HCI+NLP 2025

Fourth Workshop on Bridging Human–Computer Interaction and Natural Language Processing

Proceedings of the Workshop

November 9, 2025

The HCI+NLP organizers gratefully acknowledge the support from the following sponsors.

Sponsored by

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-353-1

Introduction

Welcome to the Forth Workshop on Bridging Human–Computer Interaction and Natural Language Processing!

The rapid advancement of natural language processing (NLP) research such as recent large language models has led to a variety of language technologies spanning a wide range of domains, such as conversational search and writing assistance. Those models are trained on vast amounts of data generated by people and rely on human feedback for continual improvement. While this widespread adoption ignites excitement, it raises pressing concerns and challenges in NLP research, such as real-world evaluation, bias and fairness, and model interpretability and explainability. Meanwhile, the field of human–computer interaction (HCI) develops rigorous methods for 1) studying and understanding human behavior to design technologies and 2) understanding how people use those technologies. Such a human-centered approach manifested in substantial efforts to understand the socio-cultural dynamics of data curation, to develop frameworks and tools to audit biases and ethical issues in intelligent systems, and to study people's interaction with language technologies and its impact on people's behavior.

This workshop aims to bridge the NLP and HCI communities to allow members of the NLP community to learn why, whether, and how methods and theories from HCI might be useful in advancing core NLP work, as well as allowing members of the HCI community to learn how advances in NLP might shape HCI research and practice centered around language technologies.

We are delighted to continue the effort of three previous editions of this HCI+NLP workshop at EACL 2021, NAACL 2022, NAACL 2024 and bring the forth edition to EMNLP 2025. In this workshop we present 38 papers, of which 25 are archival papers, and 13 are non-archival papers to be presented at the workshop but not included in the proceedings.

We would like to thank everyone who submitted their work to this workshop, as well as the program committee for their insightful review and feedback. We would also like to thank our invited speakers: Dr. Anjalie Field and Dr. Zhicong Lu.

We hope you find this workshop enjoyable! — Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dev, Siyan Li, Michael Madaio, Jack Wang, Sherry Tongshuang Wu, Ziang Xiao, and Diyi Yang

Organizing Committee

Organizing Committee

Su Lin Blodgett, Microsoft Research, Canada
Amanda Cercas Curry, CENTAI Institute, Italy
Sunipa Dev, Google Research, USA
Siyan Li, Columbia University, USA
Michael Madaio, Google Research, USA
Jack Wang, Adobe Research, USA
Sherry Tongshuang Wu, Carnegie Mellon University, USA
Ziang Xiao, Johns Hopkins University, USA
Diyi Yang, Stanford University, USA

Program Committee

Program Committee

Adeel Ahmed, Özge Alacam

Maike Behrendt, Milena Belosevic

Mahdi Dhaini, Bonaventure F. P. Dossou

Kevin Feng

Ana Valeria González, Sireesh Gururaja

Shreya Havaldar, Frederik Hvilshøj

Bahar İlgen

Khalid Al Khatib, Huda Khayrallah, Dayeon Ki, Sergey Kovalchuk

Alyssa Lees, Arun Balajiee Lekshmi Narayanan, Jenny Liang

Qianou Ma, Marianna J. Martindale, Katelyn X. Mei, Tanu Mitra, Meena Devii Muralikumar, Smitha Muthya Sudheendra

Tatiana Passali

Hope Schroeder, Michael Scriney, Nikhil Sharma, Qinlan Shen, Feifan Song, Jiao Sun

Serena Tardelli

Bram Van Dijk, Leonor Veloso

Ruyuan Wan, Zichao Wang, Philipp Wicke

Chenyang Yang

Jiawei Zhou

Invited Speakers

Anjalie Field, Johns Hopkins University, USA
Zhicong Lu, George Mason University, USA
Heloisa Candello, IBM Research, Brazil

Table of Contents

Digital Tongues: Internet Language, Collective Identity, and Implications for Human-Computer Interaction

Zi-Xiang Lin
Kang Chiao International School
{zixiang.ryan.lin@gmail.com}

Abstract

Nowadays, internet languages, including emojis, memes, hashtags, and slang, have become vital in constructing online communities' collective identities. However, all these forms of internet language can sometimes disempower people from other generations or cultures. This position paper presents an argument explaining how online forms of communication create social belonging for specific groups at the expense of users, and especially elderly people, due to interpretation hurdles. The present study aims to evaluate the relationship between the internet language and online collective identity, highlighting how patterns in internet language can inform human-computer interaction (HCI) by revealing how users' express identity, inclusion, and exclusion online.

1 Introduction

Language is a powerful tool to describe events, record dialogue, and form a collective identity. Indeed, language can be in any form that stands as a method of human communication. Theatre and Dance Assistant Professor Deborah Paredez's book "Selenidad: Selena, Latinos, and the Performance of Memory" captures how Selena utilized song, a musical performance, as a "language" to express Latino identity. Born on April 16, 1971, Selena is a Texas Tejano singer whose songs deeply reflect her Mexican-American experiences and cultural pride. As recorded by Paredez (2009), Selena not only resonates with other Latinos by featuring elements of her culture but also forms a collective identity through her actions and voice. Shifting into the 21st century, where language is utilized digitally, internet languages play a key role in building collective identity and redefining how people interact, create, and share cultural narratives, just like how Salena uses her voice. Unlike traditional communities bound by geographical borders, online communities exchange or share information on social media platforms every day at any time.

Specifically, four main types of internet languages have been created to communicate online: slang offers casual vocabulary for quick communication, emojis convey emotions and thoughts, memes capture experiences in visual formats, and hashtags create themes for posts (Barron and Bollen, 2022; Graham, 2019; Oliseyenum and Oghenetega, 2023; Petrova, 2021). Becoming more prevalent in online social platforms, internet languages' emergence raises the question: To what extent does internet language contribute to collective identity formation in online communities?

As the internet evolves as the primary global interaction, the ability to engage diverse users through the internet language is crucial for building a collective identity among online communities. When analyzing the linguistic contribution to online communities, the generational factor appears to disintegrate the online community as a whole; understanding future trends is also essential since digital transformation is rapid, cross-generational, and on the cutting edge for new lifestyles. This paper explores how internet language influences the formation of collective identity in online communities, particularly using slang, emojis, memes, and hashtags. While not proposing design solutions, the present position paper highlights how patterns in internet language can inform human-computer interaction (HCI) by revealing how users' express identity, inclusion, and exclusion online. These insights may support the future HCI and NLP research designs toward more inclusive, culturally aware interfaces that bridge generational and social gaps.

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 1–6
November 9, 2025 ©2025 Association for Computational Linguistics

2 Internet languages and Collective Identity in Online Community

Internet languages are writing language conventions and linguistic features in online communication, conveying a message (Squires, 2010). Although internet languages usually appear in "word text" such as slang and hashtags, they can be presented in graphs like emojis or memes. Specifically, slang and emojis convey ideas or emotions, while memes and hashtags connect users with shared experiences. According to English professor Graham (2019) and foreign languages associate professor Petrova (2021), the internet language bridges cultural barriers in real life as people resonate with others based primarily on feelings. Specifically, individuals can interpret the emotions of emojis and memes from their selection without knowing the context. They promote emotional resonance, allowing users to share and experience similar emotions. This ability to communicate nuanced emotions through simple graphics and texts fosters cross-cultural understanding as they promote openness, flexibility, and creativity, especially when people are lonely and can find online companionship. Thus, the widespread use of emojis and memes blends creativity with cognitive development, forming a collective identity in the communication medium.

Also, to build emotional connections, the hash character (#) implies the intention to establish a connection with others, as indicated by computational social science postdoctoral associates Barron and Bollen (2022). For instance, the hashtag in social media posts underlines the media user's identity, like "#metoo," subtly inviting others to resonate with the sender. As the number of viewers increases, more and more people will feel the same and join the discussion with similar beliefs. Ultimately, this fosters individual consciousness and expands it to the internet users' collective identity.

Yet, several studies hold the opposite argument. Social psychology professor Hogg (2016) linked online users' behavior to Social Identity Theory (SIT), enhancing in-group members' traits and reducing characteristics similar to those outside the group. This suggests that affiliation to specific groups increases, while the association between diverse groups diminishes, leading to party isolation and trait dissociation. The dynamic highlights how online interactions can undermine the online community wholly and impact social behavior, ultimately influencing users' engagement and perspectives in digital spaces.

Extending this idea to internet language, English literary study doctors Oliseyenum and Oghenetega (2023) claim that slang is "a specialized form of language variation unique to a particular social group." That is, only people in the "community," aside from additional learning, can understand the meaning of the slang. While the essences of emojis, memes, and hashtags are often easy to recognize, they also require users to grasp subtle meanings for full understanding. As a result, these "cognitive limits" restrict internet language from being used outside a specific group, supporting SIT and building barriers to online communication.

Similar observations are concluded by associate psychology professor Bäck et al. (2018). The researchers found that singular pronoun (I) usage among online platforms decreases, while plural pronoun (we/they) significantly increases, "distancing to more outgroups." This shift in pronoun use illustrates in-group bonding and highlights how digital language norms exclude newcomers, creating barriers to unity. Thus, the study reflects the decline of individual consciousness and the rise of group awareness within similar parties, supporting the SIT. Concluding the linguistic perspective, although internet languages strengthen the bonds in each group with similar traits, they aggravate the gap between different ethnic parties, instead of integrating the online community as a whole.

3 Generational Gaps and Digital Literacy

Binney (2004) indicates the interpretation of text and narratives will be "reinterpreted by another generation" as they "will continue changing and be changed." This underlines the characteristics of language, which will be interpreted differently in different eras. The study also implies that people of different generations have different views on the same language, justifying that older and younger people may have different interpretations or understandings of internet language.

Indeed, economics and informatics assistant professor Hysa et al. (2021) discovered the daily social media usage of Baby Boomers (born in 1945–1964), Generation X (born in 1965–1980), Generation Y (born in 1981–1994), and Generation

Z (born in 1995–2010) are 28.6%, 40%, 50%, and 90% respectively. Specifically, social media usage is higher among Generation Z and Y than Generation X and Baby Boomers. The findings indicate that social network utilization decreases as age increases, illustrating the generational communication differences. As younger generations seek online media as the primary transmission source, this potentially causes separation of generational groups in online communities, since youth surf on online platforms more often.

In addition, the selection of internet language by different generations also contributes to the discrepancy. According to language and literature assistant professor Azad et al. (2023), Generation Z embraces highly informal language, abbreviations, and internet language. In contrast, Baby Boomers, the older generation, prefer more formal language and traditional communication norms on social media (Puspita and Ardianto, 2024). This reflects Generation Z's adaptation to digital communication's fast-paced, visually-oriented nature, challenging the elders to comprehend the meaning. Combining Hysa et al. (2021) and Azad et al. (2023), the two studies underline the destruction of collective identity between generations due to diverse internet language preferences.

Moreover, digital generational differences also affect real-life relationships among different age groups, causing digital gaps to widen. Unlike older generations, Generation Z relies on digital communication and integrates remote work (Pichler et al., 2021). Since work behaviors differ among generations, with the younger generation employing more social platforms, the fundamental digital differences affect both internet usage and physical human interactions, creating online media barriers. As time goes on, the younger generation will embrace more digital technologies, while the older generation continues to employ the traditional working style. Thus, the elders will not collaborate with youth and therefore do not use internet language, demonstrating that internet language cannot unite people of different ages. Concluding the generational perspective, online and offline interactions will exacerbate the generational gaps, underlining that different internet language utilization dissociates generational collective identity.

4 Future Digital Platform

For future predictions, researchers believe that online communication will resolve the barriers as they integrate languages on the internet to develop a collective identity. After dozens of years, the current generational barriers will be mitigated as the current youth emerge and dominate the society, resulting in almost everyone employing internet language to communicate. In fact, world studies professor Godwin-Jones (2018) hypothesizes that internet language, the informal language choices, will gradually replace traditional formal communication as digital communication becomes more integral to everyday life. That is, the employment of slang, hashtags, emojis, and memes will all significantly increase as younger generations adopt them progressively.

Nevertheless, while the original generational barriers are mitigated, new generational barriers will emerge. For instance, Nurhayati (2025) points out Generation Alpha (born in 2011–2024) continues to grow their language repertoire, preferring conventional idioms and less complex syntax. This contrasts with Generation Z, who demonstrates a better contextualized and multi-layered comprehension of internet-based phrases. Furthermore, Melissa et al. (2024) attributes this phenomenon since Generation Alpha interacts more with short-form video platforms, including YouTube and TikTok, that makes them act differently with Generation Z, who are influenced by earlier social media culture.

Regarding prediction on blending languages, Spanish and Portuguese assistant professor Dickinson (2023) observed codeswitching, the process of changing language, in online blogs. By combining Spanish and English, Dickinson's (2023) case study reached more internet users since understanding was not limited to only English or Spanish speakers. The tendency to attract more social media users will encourage others to follow, foreshadowing higher employment of codeswitching. However, Dickinson (2023) fails to acknowledge the challenges of regional barriers or integrating languages from distinct language families. Dickinson observes residents' behaviors who live in the United States to evaluate codeswitching. Since both Spanish and English are spoken in the United States, Dickinson's study can only support language integration in places that already use those languages, failing to resolve barriers in online communication across countries and continents that speak distinct languages. As a result, the future norm of combining languages in online platforms is unlikely to happen and form a worldwide collective identity.

Proposing alternative future predictions on online communication developments, several researchers claim that the growth of emojis and memes will be faster because they do not necessarily require the viewer to know the language and understand the meaning. Specifically, computer science researchers Raj and Balachandran (2020) and Artificial Intelligence safety researcher Weng et al. (2014) predict that emojis and memes in online communication will rise, becoming more embedded in digital culture as their overall usage is expected to increase. To expand viewership, emojis and memes will experience more significant changes because they can acquire a wider range of online users, as they often transcend language barriers. Their language-independent nature allows users to easily grasp meaning, breaking communication barriers and fostering a global collective identity. Comparing Dickinson's (2023) and other researchers' predictions, future online platforms are more likely to shift toward image-based Internet languages because they demand fewer language restraints.

Elaborating on fewer text-based languages in the future, computer information researcher Penni (2016) further forecasts that video, with fewer verbal limitations, will dominate online interactions, as it is more dynamic and easier to attract attention. Penni's (2016) prediction of video evolution eliminates the possibility that viewers do not understand the emojis, memes, or text in Internet languages. As users increasingly prefer video content, communication on online social networks will shift toward video-based interactions rather than text or image-based internet language. Nevertheless, video content cannot completely avoid text usage because videos can only better engage the viewer but cannot replace verbal communication to exchange information.

To solve this, psychology professor Gernsbacher (2015) discovered that video captions benefit viewers, especially non-native speakers, who watch videos to improve comprehension by more than 60%. To spread information, both internet languages and video require the viewers to at least understand the meaning, emotion, or experience. However, a video without any language that comprehends these elements will decrease its employment as online users cannot comprehend the video, failing to engage users. Thus, although internet communication will shift toward videos, internet languages will still be required.

Combining Penni's (2016) and others' predictions, the future social media platform will employ more emojis and memes along with videos, as they require fewer language recognition abilities while engaging more participants. These new forms of online communication will allow more people to join, understand the post, and share the same ideas, promoting the collective identity of the online community.

5 Implication for Human-Computer Interaction

The online collective identity dynamic and cognitive diversity among different linguistic or social groups present problems for Human-Computer Interaction and Natural Language Processing. Specifically, internet language evolves too quickly and symbols that are not in traditional corpora. Hence, the NLP models tend to be inaccurate when performing sentiment analysis or in the task of translation (Ishita and Mamidi, 2025; Khurana et al., 2022; Raiaan et al., 2024).

Omar et al. (2022) mentions that with these dynamic, creative, and sometimes adversarial internet languages, the models will not generalize, will misunderstand meaning, or will become vulnerable to adversarial attacks (e.g. minor changes of phrasing that evade spam filters or chatbots). In fact, Ishita and Mamidi (2025) explains Generation Alpha usage of slang has not yet been sufficiently translated by AI, wherein 89% of inaccuracies came from contextual misinterpretations. Since internet language depends on context, cultural references, and pragmatic cues, such as sarcasm or irony, this fast-evolving feature urges NLP systems and programmers to bridge the gap in real-world user-generated content scenarios to improve sentiment analysis and text translation.

While bridging cognitive gaps is evident, it is also important to take note of individuality to maintain online identity. Van Der Meer (2024) highlights that NLP systems and large language models tend to capture the dominant issues or opinions raised by the majority while often neglecting the minority or dissenting voices. This issue becomes a distortion of a complete collective identity, impeding the formation of diversity upon which any robust online community is based (Burton et al., 2024). Thus, as the algorithms are trained over static or majority-centered corpora, the algorithms risk misinterpreting context-dependent expressions, neglecting minority voices, and accommodating emergent linguistic expressions. Moving forward, systems should be assisted by adaptability, inclusivity, and cultural sensitivity such that AI tools not only interpret languages but also respect and maintain diversity in online collective identities.

6 Conclusion

While internet languages will bridge groups with similar traits and develop primarily on image-based and video in the future to foster collective identity, they will disintegrate the connection between each online community, especially generations with different internet employment. As online languages evolve, the need for digital literacy will increase. Unlike traditional language learning, which often occurs in formal education settings, understanding internet language requires users to adapt to a fast-changing, informal digital environment. Effective communication in online spaces depends not only on users' ability to decode these forms but also on their skill in using them to engage and inform others meaningfully. Insights from this study may inform future HCI approaches that aim to bridge generational gaps, promote digital literacy, and support diverse forms of online expression without reinforcing exclusionary norms.

Limitations

Due to the fast-paced nature of online platforms, future internet languages may emerge rapidly, replacing existing ones. This creates linguistic gaps between regions, subcultures, and communities, making it challenging to form a cohesive collective identity online. To understand the future trends of internet languages, future research should monitor and investigate new forms of online communication continuously, accounting for the evolution of these languages as they develop.

Ethical Considerations

This position paper raises ethical concerns for future research about how internet language is interpreted in HCI design. Standardizing emojis, memes, or slang for usability risks erasing cultural and generational diversity. Future systems should avoid reinforcing dominant norms or excluding marginalized groups. Researchers should also be cautious when applying automated analysis, ensuring that meaning is not oversimplified or misused.

References

Alexander T. J. Barron and Johan Bollen. 2022. Quantifying collective identity online from self-defining hashtags. *Scientific Reports*, 12(1).

Beata Hysa, Aneta Karasek, and Iwona Zdonek. 2021. Social media usage by different generations as a tool for Sustainable tourism Marketing in Society 5.0 idea. *Sustainability*, 13(3):1018.

Deborah Paredez. 2009. *Selenidad: Selena, Latinos, and the Performance of Memory.* Duke University Press.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2022. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744.

Emma A. Bäck, Hanna Bäck, Marie Gustafsson Sendén, and Sverker Sikström. 2018. From I to We: Group formation and linguistic adaption in an online xenophobic forum. *Journal of Social and Political Psychology*, 6(1):76–91.

Helen Wilfred Raj and Santhi Balachandran. 2020. Future emoji entry prediction using neural networks. *Journal of Computer Science*, 16(2):150–157.

Iis Kurnia Nurhayati. 2025. Vernacular Evolution: A Comparative Study of Language Use in Generation Z and Generation Alpha on Social Media. *International Journal of Language and Cultural Linguistics*, 1(1):18–27.

Inamul Azad, Sugandha Chhibber, and Azra Tajhizi. 2023. How do different generations communicate on social media? A comparative analysis of language styles, emoji usage, and visual elements. *Language, Technology, and Social Media.*, 1(2):22–33.

Ishita and Radhika Mamidi. 2025. The Evolution of Gen Alpha Slang: Linguistic Patterns and AI Translation Challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 678–686, Vienna, Austria. Association for Computational Linguistics.

Janice Penni. 2016. The future of online social networks (OSN): A measurement analysis using social media tools and application. *Telematics and Informatics*, 34(5):498–517.

Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, Lucie Flek, Stefan M Herzog, Saffron Huang, Sayash Kapoor, Arvind Narayanan, Anne-Marie Nussberger, Taha Yasseri, Pietro Nickl, Abdullah Almaatouq, et al. 2024. How large language models can reshape collective intelligence. *Nature Human Behaviour*, 8(9):1643–1655.

Judith Binney. 2004. Maori oral narratives, Pakeha written texts : two forms of telling history. *New Zealand Journal of History*, 38(2):203–214.

Kendra V. Dickinson. 2023. What does it Meme? English–Spanish codeswitching and enregisterment in virtual social space. *Languages*, 8(4):231.

Lauren Squires. 2010. Enregistering internet language. *Language in Society*, 39(4):457–492.

Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2014. Predicting successful memes using network and community structure. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):535–544.

Maledo Richard Oliseyenum and Edobor Helen Oghenetega. 2023. A morpho-semantic analysis of some Nigerian internet-based slangs. *Kampala International University Interdisciplinary Journal of Humanities and Social Sciences*, 4(1):110–126.

Marwan Omar, Soohyeon Choi, Daehun Nyang, and David Mohaisen. 2022. Robust natural language processing: recent advances, challenges, and future directions. *IEEE Access*, 10:86038–86056.

Michael A. Hogg. 2016. Social Identity Theory. In *Understanding peace and conflict through social identity theory: Contemporary global perspectives*, pages 3–17. Springer International Publishing.

Michiel Van Der Meer. 2024. Facilitating Opinion Diversity through Hybrid NLP Approaches. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 272–284, Mexico City, Mexico. Association for Computational Linguistics.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Morton Ann Gernsbacher. 2015. Video captions benefit everyone. *Policy Insights From the Behavioral and Brain Sciences*, 2(1):195–202.

Paula Melissa, Maurenta Bunga Novia Siregar, Fathiha Malika Shakira, Lailan Haz, and Rahmadsyah Rangkuti. 2024. Sociolinguistic Study: Variation of Slang Words between Gen Z and Gen Alpha. *Philosophica Jurnal Bahasa Sastra Dan Budaya*, 7(2):114–125.

Robert Godwin-Jones. 2018. Chasing the Butterfly effect: Informal language learning online as a complex system. *Language Learning & Technology*, 22(2):8–27.

Sage L. Graham. 2019. A wink and a nod: The role of emojis in forming digital communities. *Multilingua*, 38(4):377–400.

Shaun Pichler, Chiranjeev Kohli, and Neil Granitz. 2021. DITTO for Gen Z: A framework for leveraging the uniqueness of the new generation. *Business Horizons*, 64(5):599–610.

Vinka Ganita Puspita and Ardik Ardianto. 2024. Code-Switching and Slang: An analysis of language dynamics in the everyday lives of Generation Z. *Linguistics Initiative*, 4(1):76–87.

Yulia Petrova. 2021. Meme language, its impact on digital culture and collective thinking. *E3S Web of Conferences*, 273:11026.

Supporting Online Discussions: Integrating AI Into the adhocracy+ Participation Platform To Enhance Deliberation

Maike Behrendt
Heinrich Heine University
maike.behrendt@hhu.de

Stefan Sylvius Wagner
Heinrich Heine University
stefan.wagner@hhu.de

Mira Warne
Heinrich Heine University
mira.warne@hhu.de

Jana Leonie Peters
Heinrich Heine University
jana.leonie.peters@hhu.de

Marc Ziegele
Heinrich Heine University
marc.ziegele@hhu.de

Stefan Harmeling
Technical University Dortmund
stefan.harmeling@tu-dortmund.de

Abstract

Online spaces provide individuals with the opportunity to engage in discussions on important topics and make collective decisions, regardless of their geographic location or time zone. However, without adequate support and careful design, such discussions often suffer from a lack of structure and civility in the exchange of opinions. Artificial intelligence (AI) offers a promising avenue for helping both participants and organizers in managing large-scale online participation processes. This paper introduces an extension of adhocracy+, a large-scale open-source participation platform. Our extension features two AI-supported debate modules designed to improve discussion quality and foster participant interaction. In a large-scale user study we examined the effects and usability of both modules. We report our findings in this paper. The extended platform is available at https://github.com/mabehrendt/discuss2.0.

1 Introduction

Online discussions and participation platforms enable people to engage in socially relevant issues. However, written exchanges in online spaces are frequently marked by a lack of structure, often leading to *information overload*, making it difficult for both participants and providers to process large volumes of contributions (Arana-Catania et al., 2021). According to Anastasiou et al. (2023), other key issues include polarization, incivility, toxic behavior, superficial content, and insufficient collaboration among participants. To address these challenges, the concept of deliberation proves particularly valuable. Deliberation is defined as the respectful and argumentative exchange of opinions aimed at making a decision. It encompasses three core dimensions: *rationality*, referring to the argumentative exchange of opinions; *civility*, which entails politeness and respect; and *reciprocity*, characterized by responsiveness and active listening (Bächtiger et al., 2009; Esau et al., 2021; Graham, 2010).

AI presents a promising opportunity to enhance deliberation, supporting both participants and organizers in creating a more structured, respectful, and engaging environment for meaningful exchange of opinions. In this work, we propose two AI-based solutions to improve online discussions, implemented for adhocracy+, an open-source participation platform.

Our contributions:

1. **Comment Recommendation Module:** To encourage user interaction and expose participants to opposing viewpoints, we developed a comment recommendation module based on a stance detection model.

2. **Deliberative Quality Module:** To enhance user engagement and improve the quality of contributed comments, we implemented a debate module that automatically detects and highlights the most deliberative comments.

3. **Application and Evaluation:** To examine the effects of the proposed modules, we conducted a large-scale panel study (N = 1,356). The results of the user study are presented in detail in the following sections.

2 Related Work

Previous efforts to integrate AI into discussion platforms have often focused on structuring and summarizing discussions. The CONSUL[1] citizen participation tool enables citizens to propose ideas to local politicians on improving their city. These proposals can be supported and discussed by other participants on the platform. To address the issue of *information overload*, Arana-Catania et al.

[1] https://consulproject.nl/en/

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 7–16
November 9, 2025 ©2025 Association for Computational Linguistics

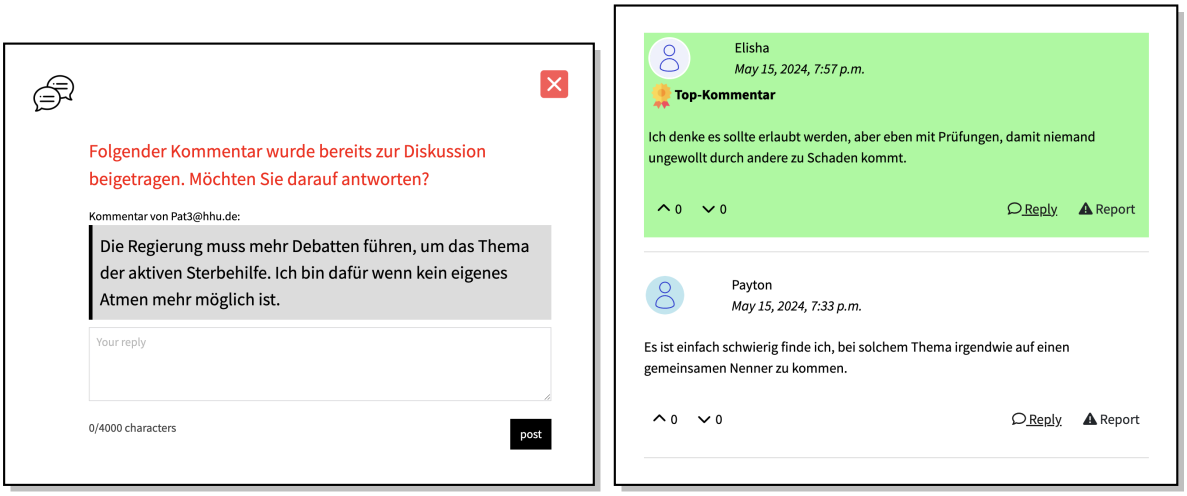


Figure 1: We propose two AI tools that we integrate into adhocracy+. **(Left) Comment Recommendation Module:** Participants are confronted with a comment that contradicts their own opinion and are asked if they want to respond. The AI tool determines the stance of the comments, which is used to propose opposing comments. Translation: The following comment has already been added to the discussion. Do you want to reply to it? **(Right) Deliberative Quality Module:** We predict a deliberative quality score (AQuA score) for each comment. Comments with a high AQuA score are sorted to the top of the discussion and highlighted in bright green and marked as "top comment".

(2021) improved the platform with several natural language processing (NLP) methods, including tools to summarize existing proposals, automatically categorize them and recommend proposals to participants according to their interests.

In the KOSMO project, an AI-supported moderation dashboard was developed for the adhocracy+ platform to assist moderators during citizen participation processes[2]. Two models were trained to identify uncivil and engaging comments (Risch et al., 2021), which are flagged for moderators, allowing them to decide on appropriate actions, such as blocking uncivil comments.

The BCause platform, created by Anastasiou et al. (2023), supports discussions with an automatic text summarization tool and an argument recommendation system. This system suggests arguments from scientific literature based on the user's stance on the discussed topic. Other examples of open-source discussion tools that incorporate AI features include Discourse[3] and Polis (Small et al., 2021) from the Computational Democracy Project.

Another notable example is Community-Pulse (Jasim et al., 2021), a platform equipped with tools for text analysis and visualization to help civic leaders to make sense of community input. It includes a sentiment analysis of contributions and topic modeling to automatically extract discussed topics.

Beyond civic tech, there is a broader body of research focused on using AI to support discussion in the context of collaborative learning (see, e.g., Kong et al. (2025)).

Similar to our approach, Yeo et al. (2024) also aim to enhance deliberative quality on online discussion platforms. They employ large language models (LLMs) to generate reflective nudges designed to promote users' self-reflection, thereby fostering more thoughtful and deliberative contributions. In our work, we focus on directly enhancing the deliberative quality of discussions by improving their reciprocity and rationality. To achieve this, we introduce two new modules for the adhocracy+ platform: (i) the *Comment Recommendation Module* that suggests comments based on whether participants are *in favor* or *against* the discussed issue, encouraging participants to engage with opposing viewpoints, and (ii) the *Deliberative Quality Module* that automatically identifies and highlights the most deliberative user comments, motivating participants to contribute further high-quality comments.

3 Features

In the following, we will discuss the features of the two implemented modules from both a technical and a user perspective.

[2] https://github.com/liqd/a4-kosmo
[3] https://meta.discourse.org/

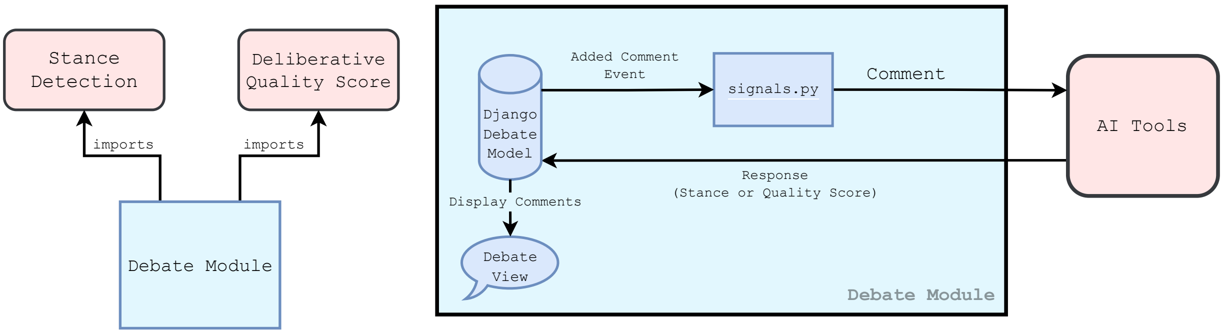


Figure 2: Overview of the architecture to extend adhocracy+ with our AI tools. **(Left)** The *debate module* imports both the *stance detection* and *deliberative quality* AI's as Python modules. **(Right)** The Django database model sends out an event when a new comment is added to the database. The event is handled in signals.py where the new comment is passed either to the stance detection or deliberative quality model. These send a response (either a stance or quality score) back to the database where the corresponding response is stored.

3.1 Enhancing Reciprocity with the Comment Recommendation Module

As previously mentioned, large-scale online discussions often involve a high volume of postings, including redundant, toxic, or uncivil content. Simultaneously, these discussions frequently lack structure, leading participants to experience information overload (Arana-Catania et al., 2021). Under this condition, participants struggle to follow the discussion and engage with others, which results in a lack of reciprocity within the conversation (Lago et al., 2019). Another consequence of information overload is dysfunctional argumentation (Klein, 2007). This, in turn, fosters the formation of small groups of participants who share similar opinions and avoid interacting with those holding opposing views (Klein, 2015).

To mitigate information overload, enhance reciprocity among participants, and improve the overall quality of discussions, we developed a *Comment Recommendation Module* integrated into the adhocracy+ debate module. This module suggests comments to the participants that reflect a point of view opposite to their own. For instance, if a participant holds an *against* stance on the debate question, the module will recommend a comment from another participant with an *in favor* stance.

Stance Detection. To detect the stance of a comment, we use an uncased German BERT Base model (Chan et al., 2020)[4] fine-tuned on the X-Stance dataset (Vamvas and Sennrich, 2020). This dataset includes 48.6k German comments on 150 political questions, answered by political office candidates in Switzerland. Since the adhocracy+ platform is specifically designed for discussion and decision making on politically relevant issues, the dataset fits our purpose very well. The fine-tuned model operates as a binary classifier, outputting either *in favor* or *against* based on a given debate question and a specific comment.

The complexity and diversity of political and social issues make it challenging to obtain sufficient labeled data for stance detection. To address this, we follow the approach of Wagner et al. (2025), leveraging synthetic data generated by LLMs. We employ Mistral-7B (Jiang et al., 2023) to generate comments reflecting an *in favor* or *against* stance. These synthetic comments are then used to further fine-tune the stance detection model. For existing comments, the synthetic data helps identify real comments that are most challenging for the model to classify. These comments can be manually labeled to further improve the model's performance. For additional details, see Wagner et al. (2025).

Comment Recommendation User Experience. The main purpose of the *Comment Recommendation Module* is to present a comment to the user that opposes their own position on the debate question. Therefore, the stance of every comment, posted in the discussion is predicted and stored into the database. When a user logs into the platform for the first time, they are prompted to indicate their stance on the debate question.

The user's position, which can be either *in favor* or *against*, is stored in the database. This information is then used to determine suitable comments for recommendation. The system retrieves com-

[4] https://huggingface.co/dbmdz/bert-base-german-uncased

ments from the database that oppose the user's stance. If multiple opposing comments are available, one is randomly selected from the list. If there is no suitable comment available, a message is displayed to the user, indicating that no comment can be suggested at that time.

The selected comment is displayed to the user in a popup window (see Figure 1, left), where they are given the option to reply. Once the user responds, the popup dialog closes, and the screen automatically scrolls to the suggested comment within the discussion. Additionally, users can reopen the suggestion popup by clicking a designated icon. When reopened, a new opposing comment (if available) is proposed for the user to reply to.

3.2 Enhancing Debate Quality and Engagement With the Deliberative Quality Module

In addition to disorganized content and dysfunctional argumentation (which diminishes reciprocity, see the previous section), online discussions face other challenges, including low-quality contributions (Klein, 2007). Addressing this issue is crucial for fostering meaningful and productive conversations. In an observational study, Wang and Diakopoulos (2022) found that manually highlighting high-quality comments in the comment section of the New York Times (referred to as the *New York Times Picks*) increases the overall discussion quality and the user engagement. The authors suggest that highlighting well-written comments is beneficial to the quality of new comments as the picked comments constitute a social feedback mechanism (Wang and Diakopoulos, 2022).

We build on these findings and develop the *Deliberative Quality Module* which aims to promote high quality comments by automatically highlighting them. It remains to be investigated whether the human component, i.e., the selection by a New York Times editor, has a significant impact on the participants' perceptions, or whether simply highlighting the comments has the same effect. To measure the deliberativeness of a user comment, we calculate the AQuA score (Behrendt et al., 2024) for each comment and define a threshold for high quality.

AQuA Score. The AQuA score, proposed by Behrendt et al. (2024), is a weighted sum of the predictions of individual BERT-based adapter models f_{θ_k} (Pfeiffer et al., 2020), fine-tuned for 20 different deliberative quality indicators. These include, i.a., *justification*, *proposing solutions*, *referencing other users* and, as an indicator for low quality, the use of incivility markers, such as *sarcasm*. Each adapter prediction is weighted with a number $w_k \in \mathbb{R}$ that is estimated from data. Some of the weights are positive, indicating a positive correlation between the respective indicator and the overall quality of the comment, and some are negative, indicating a negative correlation. The total score for a comment c is calculated as

$$s_{\text{AQuA}}(c) = \sum_{k=1}^{20} w_k f_{\theta_k}(c). \tag{1}$$

AQuA scores are normalized to the range between 0 and 5. Note that the individual predictions of AQuA are trained on expert evaluations, which are combined with weights estimated from non-expert assessments, for details see Behrendt et al. (2024). In the *Deliberative Quality Module*, AQuA scores allow us to identify high quality comments.

Deliberative Quality User Experience. The three comments with the highest predicted AQuA scores, which exceed a specified threshold, are automatically identified as top comments. They are prominently displayed above the other comments and highlighted in light green color (see Figure 1 on the right, showing only a single top comment). The other comments are displayed below the top comments in chronological order. The exact threshold depends on the discussion and can be set as a hyperparameter.

4 Implementation Details

Adhocracy+ is built on the Django framework[5] and provides a wide range of functionalities and modules to facilitate large-scale online discussions. The platform's debate module features a forum-like structure where a discussion topic is defined and displayed at the top of the page, enabling users to comment on the topic or respond to other participants' comments. Additional details about the platform's features are available on the adhocracy+ website[6]. We extend adhocracy+ by importing the AI tools into the debate module, as shown in Figure 2 (left). A more detailed view is shown in Figure 2 on the right. When a new comment is added by a user, the Django debate model fires an event,

[5]https://www.djangoproject.com/
[6]https://adhocracy.plus/info/features/

Model	Acc.	F1
BERT Base German Cased	0.7381	0.7426

Table 1: The performance on the test set of the X-Stance dataset (Vamvas and Sennrich, 2020) of the fine-tuned BERT Base German cased model we used for stance prediction.

	Deliberative Aspect	MBERT uncased
Rationality	Relevance	0.37
	Fact	0.56
	Opinion	0.57
	Justification	0.69
	Solution Proposals	0.79
	Additional Knowledge	0.78
	Question	0.87
Reciprocity	Referencing Users	0.88
	Referencing Medium	0.93
	Referencing Contents	0.81
	Referencing Personal	0.92
	Referencing Format	0.96
Civility	Polite form of Address	0.97
	Respect	0.9
	Screaming	0.81
	Vulgar	0.74
	Insults	0.87
	Sarcasm	0.48
	Discrimination	0.88
	Storytelling	0.85
	Ø Total Average (F1-Score)	0.7815

Table 2: We show the weighted average F1 score for the 20 different deliberative aspects the AQuA score adapter models are trained on.

which is handled in the `signals.py` file. Here, we import the AI tools to pass the comments to the stance detection or the deliberative quality model. The AI tools then return a response (either a stance or deliberative quality score), which is stored back to the database for the corresponding comment. This stored response is then presented by the corresponding module as shown in Figure 1. For the purposes of this study, they were implemented as distinct debate modules within the adhocracy+ platform in order to enable the separate evaluation of their respective effects. Their integration into a unified module remains a plausible direction for future development. Overall, this architecture is flexible: In our experiments, we ran the AI tools locally on a Linux server. But the AI tools could also be run as services where communication is handled via Rest API.

5 Evaluation

In the following, we analyze the effectiveness of our two proposed modules. We start by evaluating both models on existing datasets and measure how well they perform in terms of accuracy and F1 score. Furthermore, we conducted a large-scale user study to evaluate participants' satisfaction when using the modules in a real online discussion as well as to gauge the effects of the modules on other perceptions and behaviors of the participants.

5.1 Model Performance

Comment Recommendation Module Table 1 displays the performance of the German BERT Base uncased model, which was fine-tuned on the X-Stance dataset (Vamvas and Sennrich, 2020). The model reaches an accuracy of 73.81 and an F1 score of 74.26 on the test dataset.

Deliberative Quality Module A multilingual BERT base uncased model[7] serves as the basis for the trained adapter models that build the AQuA score (Behrendt et al., 2024). Table 2 lists the weighted average F1 scores on the test dataset for each of the 20 trained adapter models on deliberative aspects.

5.2 User Study

5.2.1 Methodology

To investigate the effects of both AI modules, we conducted a field experiment as part of a three-wave panel survey in July 2024. Participants were recruited from the German population through Bilendi, an online access panel provider and market research company. The final sample consisted of N = 1,356 participants with a mean age of 52 years (47% female; 58% with at least a high school diploma).

Participants joined a simulated citizens' assembly with a 10-day online discussion phase on the extended adhocracy+ platform (internally referred to as *discuss20*). They engaged in small-group discussions on two selected political topics: (1) whether active euthanasia should be legally permitted in Germany, and (2) whether the sale of alcoholic beverages should be more restricted in Germany. These topics were identified in a preliminary survey as the most engaging from a broader selection of issues.

[7] https://huggingface.co/google-bert/bert-base-multilingual-cased

#	Survey Question	Scale (1-7)
Q1	On the platform, discussion contributions were suggested to me, to which I could reply.	1 = strongly disagree, 7 = strongly agree
Q2	On the platform, contributions were marked as top comments.	1 = strongly disagree, 7 = strongly agree
Q3	To what extent did you feel that this process was supported by artificial intelligence?	1 = most certainly not, 7 = most certainly yes
Q4*	I enjoyed using discuss20.	1 = strongly disagree, 7 = strongly agree
Q5*	The functions of the discuss20 platform threatened my freedom to choose what I wanted.	1 = strongly disagree, 7 = strongly agree
Q6	All in all, I was satisfied with the discussion.	1 = strongly disagree, 7 = strongly agree
Q7*	The contributions contained arguments and justifications.	1 = strongly disagree, 7 = strongly agree
Q8*	The participants responded to the contributions of others.	1 = strongly disagree, 7 = strongly agree
Q9*	The contributions were discriminating.	1 = strongly disagree, 7 = strongly agree
Q10*	There was a wide range of opinions in the discussion.	1 = strongly disagree, 7 = strongly agree

Table 3: Excerpt from our user study survey questions. Questions that are marked with an asterisk are example questions that are part of a larger index.

The experimental design consisted of five conditions for each of the two discussed topics, aimed at testing the effects of the AI modules. These included: discussions supported by the *Comment Recommendation Module*, which either (i) recommended comments that contradicted the participant's opinion or (ii) recommended random comments. Discussions supported by the *Deliberative Quality Module*, which either (iii) highlighted three comments with the highest deliberative quality scores as "top comments" or (iv) highlighted three randomly selected comments as "top comments" and (v) discussions without AI support, serving as the control group. Participants and experimental conditions were randomly assigned, resulting in ten distinct experimental groups. Randomization checks showed no significant differences between the groups in terms of age, gender, education, or political interest.

During and after the discussions, the participants completed standardized online questionnaires to evaluate their experiences on the platform. To explore the effects of the AI modules, this user study focuses on four aspects:

1. **Manipulation effectiveness** - the extent to which participants recognized and responded to the implemented AI features.

2. **Quantitative participation** - the extent of the engagement of the participants in the discussions.

3. **Platform evaluation** - users' perceptions of the platforms usability and functions.

4. **Discussion evaluation** - participants' assessments of the discussion quality, including satisfaction and deliberative characteristics.

An excerpt of the corresponding survey questions is listed in Table 3. The effectiveness of the manipulations was measured through participants' recognition of the module-specific functions (see Q1 and Q2) and their assessment of the AI-support (see Q3). Evaluation of the platform included overall satisfaction with the platform (mean index of 5 items, see, e.g., Q4*, Cronbach's alpha = .86) as well as evaluation of the functions against the backdrop of freedom of choice (perceived autonomy, mean index of 4 items, see, e.g., Q5*, Cronbachs Alpha = .85). Lastly, evaluation of the discussions included overall satisfaction with the discussion (see Q6) and the perceived deliberative quality, evaluated across four dimensions, namely the perception of the rationality (mean index of 4 items, see, e.g., Q7*, Cronbach's alpha = .85), reciprocity (mean index of 3 items, see, e.g., Q8*, Cronbach's alpha = .89), civility (mean index of 4 items, e.g., Q9*, Cronbach's alpha .78) and diversity of the discussions (mean index of 4 items, e.g., Q10*, Cronbach's alpha .88).

As the *Comment Recommendation Module* aims to expose users to diverse viewpoints, it fosters diversity and reciprocity by encouraging interaction with opposing opinions. In contrast, the *Deliberative Quality Module* promotes civility and rationality by highlighting comments that exemplify high deliberative quality, thereby setting a constructive standard for discussion. Consequently, the analysis focuses on diversity and reciprocity for the *Com-*

	AI CR Module (n = 289)		Random CR Module (n = 276)		Control (n = 262)		F
	M	*SD*	*M*	*SD*	*M*	*SD*	
(1) Manipulation effectiveness							
Discussion contributions were suggested to me, to which I could reply	6.20[a]	1.29	5.88[a]	1.54	3.83[b]	2.22	116.18***
To what extent did you feel that the discussion was supported by artificial intelligence?	4.33[a]	1.65	4.18[a]	1.62	3.80[b]	1.71	7.47***
(2) Quantity of participation							
Average number of comments per user	12.71[a]	12.85	12.98[a]	11.85	9.16[b]	10.58	9.82***
(3) Platform evaluation							
Overall satisfaction with the platform	6.08	1.13	6.11	1.06	6.15	0.97	0.34
Experience of threats to freedom of choice	1.44	0.89	1.51	0.96	1.37	0.80	1.61
(4) Discussion evaluation							
Satisfaction with the discussion	5.98[a]	1.16	5.89[ab]	1.33	5.63[b]	1.49	4.60*
Perception of diversity	6.03[a]	0.94	5.89[ab]	1.01	5.74[b]	1.05	5.50**
Perception of reciprocity	5.64[a]	1.03	5.41[a]	1.13	4.97[b]	1.35	21.00***

n = 827, One-Way ANOVA (Post-Hoc-Test: Bonferroni/Games-Howell), *p<0.05, ** p<0.01, *** p<0.001.
Note: Groups with different code letters (a, b) differ significantly at the 5% level.

Table 4: Results of One-Way Analyses of Variance (ANOVAs) for the Comment Recommendation (CR) Module.

ment Recommendation Module and on civility and rationality for the *Deliberative Quality Module*, as these dimensions best capture the intended effects of each intervention.

5.2.2 Results

Comment Recommendation Module. We conducted One-way Analyses of Variance (ANOVAs) to investigate group-specific manipulation effectiveness, quantitative participation, platform evaluation, and discussion evaluation. A summary of the results for the *Comment Recommendation Module* is provided in Table 4. We report mean (M) and standard deviation (SD) and F-Values (F). Regarding manipulation effectiveness, participants in the *Comment Recommendation Modules* scored significantly higher on identifying this platform feature and on perceiving AI support compared to the control group. However, the participants' assessment whether the discussion was supported by AI did not significantly differ between the modules with random and AI-based comment recommendation.

Regarding participation, participants in the *Comment Recommendation Modules* wrote an average of approximately three to four more comments per user compared to the control group. Again, it was inconsequential whether the recommended comment was suggested randomly or AI-based. Regarding users' evaluation of the platform, the *Comment Recommendation Modules* did not impair users' satisfaction with the platform due to the module-specific implemented functions. Another positive finding is that the *Comment Recommendation Modules* did not restrict participants' feelings of autonomy. In contrast, regarding the effects on discussion evaluation, especially participants in the AI-supported *Comment Recommendation Module* reported a significantly higher satisfaction with the discussion and higher perception of the deliberative dimension of diversity compared to the control group. Finally, comment recommendation significantly increased participants' perception of reciprocity within the discussion compared to the control group. Regardless of an underlying AI-based recommendation, we found that recommending comments had an overall positive effect on individual participation.

Deliberative Quality Module. Table 5 provides an overview of the ANOVA results for the *Deliberative Quality Module*. Regarding manipulation effectiveness, participants in both the AI *Deliberative Quality* and Random *Deliberative Quality Modules* were significantly more likely to recognize platform contributions marked as top comments and to perceive AI support compared to the control group. However, the participants' assessment of AI support did not significantly differ between the AI *Deliberative Quality* and Random *Deliberative Quality Modules*.

In terms of participation quantity, the average number of comments per user did not differ significantly between the groups, suggesting that neither the AI *Deliberative Quality* nor the Random

	AI DQ Module (n = 289)		Random DQ Module (n = 276)		Control (n = 262)		F
	M	*SD*	*M*	*SD*	*M*	*SD*	
(1) Manipulation effectiveness							
On the platform, contributions were marked as top comments	5.81[a]	1.76	5.90[a]	1.61	3.05[b]	2.02	189.17***
To what extent did you feel that the discussion was supported by artificial intelligence?	4.20[a]	1.80	4.50[a]	1.61	3.80[b]	1.71	11.22***
(2) Quantity of participation							
Average number of comments per user	9.40	10.39	9.58	9.69	9.16	10.58	0.12
(3) Platform evaluation							
Overall satisfaction with the platform.	6.16	1.02	6.06	1.13	6.15	0.97	0.74
Experience of threats to freedom of choice	1.42	0.85	1.39	0.91	1.37	0.80	0.21
(4) Discussion evaluation							
Satisfaction with the discussion	5.71	1.51	5.63	1.45	5.63	1.49	0.27
Perception of civility	6.84	0.49	6.83	0.47	6.77	0.66	0.63
Perception of rationality	5.67	1.04	5.49	1.05	5.51	0.97	2.53

n = 791, One-Way ANOVA (Post-Hoc-Test: Bonferroni/Games-Howell), *p<0.05, ** p<0.01, *** p<0.001
Note: Groups with different code letters (a, b) differ significantly at the 5% level.

Table 5: Results of One-Way Analyses of Variance (ANOVAs) for the Deliberative Quality (DQ) Module.

Deliberative Quality Module led to an increase in users' commenting activity. Similarly, for platform evaluation, no significant differences were found in users' overall satisfaction with the platform or their perceptions of autonomy. Participants across all groups reported similarly high satisfaction and did not feel restricted in their freedom to choose actions on the platform. Finally, regarding discussion evaluation, no significant differences were observed between the groups in terms of satisfaction with the discussion, civility, or rationality. While the modules aimed to enhance discussion quality, their implementation did not result in perceptible changes in these specific evaluative dimensions.

In order to compare the actual quality of the discussions across the different groups, content analyses are currently being conducted. Preliminary results suggest that, for the topic of active euthanasia, the quality of discussions was higher in the *Deliberative Quality Module* than in the other modules. Again, however, it appears that it does not seem to make a difference whether the top comments are selected by the AI or at random.

6 Conclusion

In this work we present extensions to the adhocracy+ platform for citizen participation. We implemented two additional modules to support more deliberative online discussions. In the *Comment Recommendation Module* participants are confronted with opposing views to encourage user interaction, hence improving the reciprocity in the discussion. The *Deliberative Quality Module* aims to improve the quality of contributed comments by automatically highlighting the most deliberative ones.

In a large-scale user study, we tested the effects of both AI modules. We found that the *Comment Recommendation Module* increased participation on the platform and improved users' perception of the deliberative quality of the discussions while not diminishing their sense of autonomy. The *Deliberative Quality Module*, in contrast, did not significantly improve users' perceptions of the platform or the discussions. Still, there are indications that both modules had a positive influence on the discussions, albeit independently of whether AI was involved or not.

We see great potential in the features we presented to support human actors in conducting large online discussions. Certainly it remains an open task to improve the AI to a level, where people perceive its selection performance as far superior than random selection. The resulting platform is freely available under an open source license and can hopefully be used for political decision-making in the future.

Future Work. In the future we want to examine how both AI extensions to adhocracy+ can be further improved. This means gathering and annotating additional conversational data to fine-tune and improve both models. To further evaluate effects of both modules on the comment quality within the discussions, content analyses are currently being carried out.

7 Limitations

While our extensions to adhocracy+ introduce AI-driven enhancements, we must acknowledge several limitations.

Currently, the platform and both AI modules are only available in German. This limits accessibility for non-German speaking users and limits the potential for wider adoption.

Moreover, the effectiveness of both AI modules highly depends on the quality of their training data. They may struggle with nuanced or complex discussions, and incorrect predictions can potentially frustrate participants.

The effects we observed were predominantly very small, which may be due to the design of our study. In field experiments, numerous noise factors can influence the outcomes we measured - such as the perceived quality of discussions. At the same time, our experiments offer high external validity, as they were conducted in a realistic setting rather than under artificial laboratory conditions.

The partly non-significant differences between the AI, random, and control conditions may also be attributed to the statistical procedures employed. We used post hoc tests that apply strict corrections for multiple testing, which makes it more difficult to detect statistically significant effects.

However, when conducting planned contrast analyses, some differences between the AI-supported and Random *Comment Recommendation* and *Deliberative Quality Modules* do reach significance, suggesting that the AI-supported modules were perceived more positively by the participants than those working with random content selection.

Nonetheless, planned contrasts require more specific a priori hypotheses, which could not be formulated within the scope of this exploratory paper. Developing and testing such hypotheses remains a task for future research.

References

Lucas Anastasiou, Aldo De Moor, Barbara Brayshay, and Anna De Liddo. 2023. A tale of struggles: an evaluation framework for transitioning from individually usable to community-useful online deliberation tools. In *Proceedings of the 11th International Conference on Communities and Technologies*, C&T '23, page 144–155, New York, NY, USA. Association for Computing Machinery.

Miguel Arana-Catania, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice*, 2(3):1–22.

André Bächtiger, Susumu Shikano, Seraina Pedrini, and Mirjam Ryser. 2009. Measuring deliberation 2.0: standards, discourse types, and sequenzialization. In *ECPR General Conference*, pages 5–12. Potsdam.

Maike Behrendt, Stefan Sylvius Wagner, Marc Ziegele, Lena Wilms, Anke Stoll, Dominique Heinbach, and Stefan Harmeling. 2024. AQuA – combining experts' and non-experts' views to assess deliberation quality in online discussions using LLMs. In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE) @ LREC-COLING 2024*, pages 1–12, Torino, Italia. ELRA and ICCL.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German‘s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Katharina Esau, Dannica Fleuß, and Sarah-Michelle Nienhaus. 2021. Different arenas, different deliberative quality? using a systemic framework to evaluate online deliberation on immigration policy in germany. *Policy & Internet*, 13(1):86–112.

Todd Graham. 2010. The use of expressives in online political talk: Impeding or facilitating the normative goals of deliberation? In *Electronic Participation*, pages 26–41, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mahmood Jasim, Enamul Hoque, Ali Sarvghad, and Narges Mahyar. 2021. Communitypulse: Facilitating community input analysis by surfacing hidden insights, reflections, and priorities. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, page 846–863, New York, NY, USA. Association for Computing Machinery.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mark Klein. 2007. How to harvest collective wisdom for complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper*.

Mark Klein. 2015. A critical review of crowd-scale online deliberation technologies. *Available at SSRN 2652888*.

Xi Kong, Zhi Liu, Changsheng Chen, Sannyuya Liu, Zhenguo Xu, and Qianhui Tang. 2025. Exploratory study of an ai-supported discussion representational tool for online collaborative learning in a chinese university. *The Internet and Higher Education*, 64:100973.

Noémie Lago, Marianne Durieux, Jean-Alexandre Pouleur, Chantal Scoubeau, Catherine Elsen, and Clémentine Schelings. 2019. Citizen participation through digital platforms: the challenging question of data processing for cities. In *Proceedings of the Eighth International Conference on Smart Cities, Systems, Devices and Technologies*. DGTRE - Région wallonne. Direction générale des Technologies, de la Recherche et de l'Énergie [BE].

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. 2021. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: Revista de Pensament i Anàlisi*, 26(2).

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

Stefan Sylvius Wagner, Maike Behrendt, Marc Ziegele, and Stefan Harmeling. 2025. The power of LLM-generated synthetic data for stance detection in online political discussions. In *The Thirteenth International Conference on Learning Representations*.

Yixue Wang and Nicholas Diakopoulos. 2022. Highlighting high-quality content as a moderation strategy: The role of new york times picks in comment quality and engagement. *Trans. Soc. Comput.*, 4(4).

ShunYi Yeo, Gionnieve Lim, Jie Gao, Weiyu Zhang, and Simon Tangi Perrault. 2024. Help me reflect: Leveraging self-reflection interface nudges to enhance deliberativeness on online deliberation platforms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

User-Centric Design Paradigms for Trust and Control in Human-LLM-Interactions: A Survey

Milena Belosevic
Bielefeld University
milena.belosevic@uni-bielefeld.de

Abstract

As LLMs become widespread, trust in their behavior becomes increasingly important. For NLP research, it is crucial to ensure that not only AI designers and developers, but also end users, are enabled to control the properties of trustworthy LLMs, such as transparency, privacy, or accuracy. However, involving end users in this process remains a practical challenge. Based on a design-centered survey of methods developed in recent papers from HCI and NLP venues, this paper proposes seven design paradigms that can be integrated in NLP research to enhance end-user control over the trustworthiness of LLMs. We discuss design gaps and challenges of applying these paradigms in NLP and propose future research directions.

1 Introduction

While LLMs bring many advantages, their opacity hinders human agency and trust, as especially end users lack the necessary information and transparency to critically assess system decisions before following or acting on them (Förster et al., 2020). For this reason, there is a growing need in the field of NLP to develop methods that enhance end-user control over AI systems.

At the same time, in the domain of human-computer interaction (HCI), approximately 22 regulations comprising normative principles for mitigating AI risks and enhancing trust in AI systems had been published by 2020 (Hagendorff, 2020). While recent HCI studies explore the attitudes of different groups towards these policies (Agbese et al., 2023), their practical implementation is underexplored (Kaur et al., 2022; Perov and Golovkov, 2024), particularly regarding how to enable end users to proactively participate in controlling the trustworthiness of LLM systems.

This paper proposes seven design paradigms for enhancing end-user control over the trustworthiness of LLM systems based on a design-centered survey of novel methods from recent HCI and NLP studies. We define trustworthy LLMs using the following requirements for trustworthy AI proposed by Ethics Guidelines for Trustworthy AI (HLEG, 2019): (1) *human agency and oversight* (including fundamental rights); (2) *technical robustness and safety* (including resilience to attack and security, fallback plan and general safety, accuracy, reliability, and reproducibility); (3) *privacy and data governance* (including respect for privacy, quality and integrity of data, and access to data); (4) *transparency* (including traceability, explainability, and communication); (5) *diversity, non-discrimination* and *fairness* (including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation); (6) *environmental and societal well-being* (including sustainability and environmental friendliness, social impact, society, and democracy) and (7) *accountability* (including auditability, minimization, and reporting of negative impact, trade-offs, and redress). Although these guidelines were published before the rise of LLMs, we use them to define trust in LLMs because they were among the first to foreground a user-centered approach to trustworthy AI (Usmani et al., 2023) and remain an influential user-centered policy.

This survey contributes to human-centered approaches to LLMs by bridging regulatory perspectives on trustworthy AI from the field of HCI with their practical applications in NLP research on end users' interactions with LLMs.

2 Methodology

We surveyed original research papers (no work in progress, demo papers, posters, provocations, surveys, or extended abstracts) published in English in the ACM Digital Library and the ACL Anthology between January 1, 2022, and August 1, 2025. The start date was selected to include papers published

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 17–32
November 9, 2025 ©2025 Association for Computational Linguistics

shortly before the release of ChatGPT on November 30, 2022. While the ACM library was selected for its comprehensive coverage of HCI design research and venues (e.g., CHI) relevant to our focus, the ACL Anthology comprises work from some of the most important NLP venues, such as EMNLP and NAACL. This dual-sourced corpus provides a balanced foundation for identifying design patterns at the intersection of HCI and NLP. The following search string was used for the ACM library:

> *trust* OR “agency” OR “oversight” OR robust* OR safe* OR secur* OR accura* OR reliab* OR reproduc* OR “privacy” OR transparen* OR trace* OR explain* OR fair* OR bias* OR sustain* OR accountab* OR audit* or LLM*

The search function in ACL Anthology is limited to simple keyword queries and does not support using this search string. Therefore, we performed multiple keyword-based searches (e.g., trust LLM, transparency LLM, bias LLM) and complemented this with Google site:aclanthology.org searches to approximate Boolean logic and ensure broader coverage. A total of 1781 papers were screened from both databases. At least one of the search words had to appear in the abstract, the title, or the keywords of the paper to be included in the final dataset.

Importantly, the papers that fulfill this criterion were manually inspected to determine whether they have a clear focus on both trust (i.e., the trust aspect mentioned in the abstract) and end-user control in the full text. Accordingly, user-centric papers without a clear relationship to trust and vice versa: trust-related papers without end-user involvement in the design and/or evaluation stage (e.g., Miao and Fang 2025) or papers where the evaluation is conducted based only on datasets, performance comparison of several models, and evaluation metrics, rather than involving users and explicitly addressing how user control is achieved, were not considered. However, papers combining user studies with, for example, comparing the performance of several models, were considered (e.g., Zhou et al. 2024; Koraş et al. 2025; Dong et al. 2025).

Also, papers not explicitly dealing with language models or language model-based applications were excluded (e.g., DeVos et al. 2022). These criteria reduced the number of eligible papers to 773.

Finally, papers that considered user studies as future work (e.g., Hung et al. 2023) were excluded. In this way, 713 papers were excluded. The final list comprises 60 papers from both sources.

Papers did not need to explicitly address the AI HLEG guidelines, nor did we include studies that analyzed the guidelines themselves. Multimodal LLMs (Zhang et al., 2024; Tang et al., 2024; Chen et al., 2024) were discarded due to the broader use of text-based models. Figure 1, created with a web-based Shiny app (Haddaway et al., 2022), visualizes the PRISMA-compliant search process (Page et al., 2021).

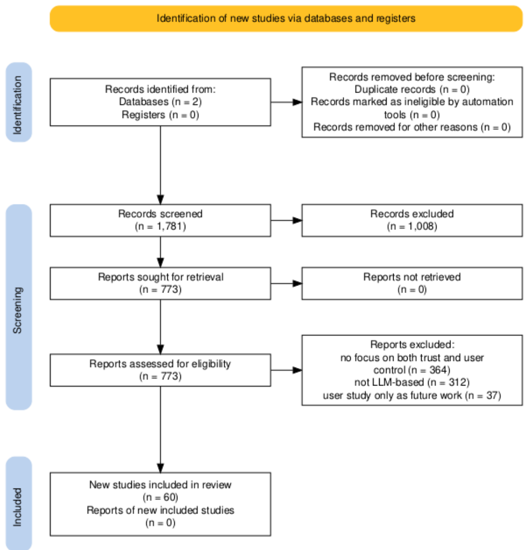


Figure 1: Overview of the literature search and screening process, following PRISMA-style structure.

Note that this is not a systematic review aiming for completeness, but a design-centered survey of recent work focusing on a synthesis of paradigms that support user control and trust in LLM systems.

Two annotators searched for the papers in the databases described above. They discussed and refined the inclusion criteria following the PRISMA paradigm (Page et al., 2021). The included papers were then annotated for their primary trust aspects (multiple assignment was allowed), and design paradigm. Annotation decisions were reached via iterative discussion. No formal inter-annotator agreement was calculated, as the focus was on interpretive synthesis. As a result, seven design paradigms are proposed (Section 3) and discussed in terms of their applications in NLP (Section 5). Note that although multiple paradigm assignments were theoretically possible, each paper was assigned to exactly one primary design paradigm based on annotator agreement.

Table 1: Design paradigms and primary user goals

Paradigm	Primary User Goal
Interface-level accuracy control	Verify factual correctness of LLM output
Workflow-aligned & domain-adapted LLM assistance	Maintain control in expert workflows
Explanation-centered approaches	Understand how and why outputs are produced
Participatory designs	Learn about LLMs; shape behavior
Interactive authoring & co-creation	Co-generate or revise outputs with AI
Style-based trust calibration	Calibrate trust based on how outputs are expressed
Privacy-aware architectures & tools	Control what personal data is exposed

3 Results

Table 1 provides an overview of design paradigms identified through inductive coding by two annotators (see Section 2). A detailed mapping of the reviewed approaches to trust aspects is provided in the Appendix. The observed skew in the distribution of papers across paradigms (interactive authoring/co-creation (13 papers) and explanation-centered approaches (12) vs. interface-level accuracy control and privacy-aware tools, four studies each) may in part reflect the methodological choices of this survey, such as the single-label annotation protocol (see Section 2). Furthermore, trust dimensions, such as accuracy, transparency, and reliability, are overrepresented probably because they are easier to operationalize through measurable interventions (e.g., confidence scores), aligning well with existing evaluation practices in NLP and HCI. In contrast, underrepresented dimensions, such as environmental and societal wellbeing, require long-term stakeholder engagement and more resource-intensive methods that are harder to implement within the scope of typical research prototypes.

Interface-level accuracy control Interface-level accuracy control refers to design approaches that equip users with interactive tools and visual cues at the interface level to help inspect, verify, and guide the factual accuracy of LLM outputs. These interfaces do not require altering the model itself, but instead focus on enhancing user control, interpretability of outputs, and trust calibration through features such as consistency checks, confidence scores, source attributions, and interactive verification workflows.

The primary goal of this paradigm is to foster accuracy and user agency by integrating transparent control mechanisms directly into the interface rather than modifying the LLM architecture.

Core strategies include tools for output verification, hallucination detection, and user-led content auditing. For example, Cheng et al. (2024) enable users to compare the factual consistency of multiple LLM outputs. Laban et al. (2024) introduce a factual editing framework that alerts users to new content, supports verification via web search, and enables tracing of model-generated edits. Formal verification has also been integrated into LLM planning tasks: Lee et al. (2025) combine model checking with user oversight. Other interfaces visualize hallucination risks or confidence scores to help users identify unreliable content (Leiser et al., 2024).

Despite promising interaction designs, this paradigm faces several challenges. First, many studies prioritize surface-level model accuracy without systematically examining how interface interventions influence other trust dimensions such as fairness, transparency, or robustness. Second, tools like confidence scores (Leiser et al., 2024) assume a high degree of AI literacy and decision-making capacity, potentially excluding non-expert users or overburdening them with the responsibility to correctly interpret, evaluate, and act on information provided by an AI system. Third, the usability and cognitive demands of these systems remain under-evaluated, as it is often unclear whether users meaningfully benefit from features like verification workflows or simply ignore them in practice.

Workflow-aligned and domain-adapted LLM assistance This design approach integrates LLMs into real-world tasks or professional practices, such as education (Kazemitabaar et al., 2024), qualitative analysis (Dai et al., 2023), legal consultation (Hu et al., 2024), banking (Gupta et al., 2025), coding (Dong et al., 2025), or clinical settings (Koraş et al., 2025), addressing domain-specific challenges of the LLM application. Users are typically given mechanisms to adapt, guide, or verify outputs in-situ, through plan-then-execute pipelines (He et al., 2025), interface-level guardrails (Liffiton et al., 2023), or feedback loops involving humans in iterative roles (Dong et al., 2025; Dai et al., 2023). Unlike generic chat interfaces, these systems align generation with domain goals and domain-specific verification routines and constraints.

The goal is to integrate LLMs into domain-specific workflows in ways that preserve user control, ensure output reliability, and align with domain-specific goals.

Examples include the restriction of LLM outputs to pseudocode in educational contexts to prevent over-reliance and support learning (Kazemitabaar et al., 2024), real-time human feedback (Gupta et al., 2025), iterative human verification Dong et al. (2025), or guardrails that prevent programmers' over-reliance (Liffiton et al., 2023), collaborative human-LLM thematic analysis and topic modeling (Dai et al., 2023; Akter et al., 2025; Choi et al., 2024).

While these paradigms offer promising forms of human-in-the-loop control, several limitations remain. First, they often assume static domain knowledge and well-formed tasks and do not adapt to rapid changes in domains like coding. Second, despite placing high cognitive demands on users (e.g., verifying assertions (Dong et al., 2025) or interpreting multi-step plans (He et al., 2025)), most designs treat users as uniformly skilled and do not assess or adjust for varying levels of domain expertise and AI literacy. This creates risks of misalignment between tool complexity and user capability and of mismatched support (either under-serving novice users or constraining experts). Finally, the integration of these designs in professional workflows raises epistemic and normative concerns since the normative assumptions integrated in designs (e.g., what counts as a "good" summary or acceptable pseudocode) are rarely made explicit or empirically evaluated. As a result, these designs may reinforce domain conventions (e.g., legal templates) without enabling critical reflection, for example in qualitative analysis (Akter et al., 2025; Choi et al., 2024).

In sum, workflow-aligned assistance offers a promising direction for domain-specific LLM use, but often relies on hidden assumptions about task stability, user capability, and normative correctness. Future work should investigate how designs could better adapt to user diversity and task ambiguity.

Participatory designs Participatory designs aim to empower users through learning and reflection, engaging them not just as passive recipients of AI output but as active collaborators, educators, assessors, or learners.

The goal is to foster AI literacy, critical awareness of LLM capabilities, and trust calibration by giving users tools to customize, question, and steer LLM behavior, particularly in educational, reflective, or interpersonal contexts.

Common strategies include user-controlled editable outputs (Chun et al., 2025), scaffolded interaction via AI literacy workshops (Theophilou et al., 2023), user-led evaluation through comparisons, subjective trust metrics (Pan et al., 2024; Zhu et al., 2025; Nguyen et al., 2024), or human-LLM evaluation of social appropriateness (Rao et al., 2025). Expert-in-the-loop approaches include collaborative prompt refinement for educational content (Reza et al., 2025) or feedback-driven role-play simulation in counseling (Louie et al., 2024).

Despite the user-centered intent, several gaps persist. First, participatory mechanisms are often introduced without sufficient onboarding or scaffolding. Users are asked to judge, configure, or collaborate with LLMs before acquiring a conceptual understanding of model behavior, which may lead to overtrust. AI literacy, while a core aim, is rarely embedded as a design prerequisite—Theophilou et al. (2023) being a notable exception. Second, customization and feedback are typically limited to surface-level tuning (e.g., tone or behavior), with little support for questioning underlying assumptions, biases, or system limitations. Third, while many systems frame participation as empowering, they may implicitly rely on user labor, placing the burden of correction, verification, and ethical reflection onto the user without adequate institutional or system-side accountability.

Overall, participatory designs signal a significant shift toward user agency and transparency, but remain underdeveloped in terms of empowering user AI literacy and critical engagement with model limitations.

Interactive authoring and co-creation This paradigm focuses on enabling users to collaborate with LLMs during complex or creative tasks (e.g., writing, prompt design, workflow creation) by enabling real-time interaction, iterative refinement, and mixed-initiative control. These systems support back-and-forth exchanges between users and LLMs, allowing users to guide, steer, edit, or evaluate intermediate outputs through customizable workflows.

The primary goal is to empower users as co-authors, prompt designers, or evaluators in creative or analytical tasks by enabling interactive, transparent, and customizable collaboration with LLMs. These systems seek to enhance human agency, reduce cognitive load, and make LLM-powered generation more interpretable and aligned with users' goals and values.

This paradigm centers on prompt chaining (Arawjo et al., 2024; Wu et al., 2022), co-auditing LLM-behavior in general (Rastogi et al., 2023), or LLM-generated biases (Prabhudesai et al., 2025) and personality traits (Zheng et al., 2025a) in particular, LLM- and human-based disinformation evaluation (Zugecova et al., 2025), co-creative authoring (Ding et al., 2023; Liu et al., 2024; Hoque et al., 2024), direct manipulation (Masson et al., 2024), and mixed-initiative control, enabling users to collaboratively shape LLM behavior via initiative-sharing interfaces (Overney et al., 2025), LLM-initiated prompt pipelines (Zhang and Arawjo, 2025) and editable preference profiles created based on user preferences (Liu et al., 2025).

Despite their promise, interactive authoring designs raise several unresolved questions. First, while many interfaces emphasize modularity, prompt chaining, or editable outputs (Arawjo et al., 2024; Wu et al., 2022; Zhang and Arawjo, 2025), it remains unclear how much initiative users actually retain in practice. Systems often alternate initiative without clearly defining the boundaries of user agency, and few studies examine whether users can override or put the model's underlying assumptions into question. Customization is usually limited to surface-level, such as prompt components, without affording deeper user control or interpretability of generation mechanisms.

Second, user literacy and feedback quality are assumed rather than supported. Designs empower users to filter outputs, flag disinformation, or assess persuasiveness (Zugecova et al., 2025; Liu et al., 2025), but offer limited scaffolding to support critical evaluation. Since there are no clear scaffolds for critical reflection, user perception of biases or auditing personality traits (Zheng et al., 2025a; Prabhudesai et al., 2025) risks being subjective and culturally dependent.

Third, while some systems highlight transparency and provenance (e.g., via interface visualizations or think-aloud protocols (Hoque et al., 2024; Rastogi et al., 2023)), it remains unclear whether such interventions are always desirable and whether more transparency always leads to better trust calibration.

Finally, there is limited evidence that these approaches generalize beyond low-stakes, exploratory domains. Many studies involve small participant samples (e.g., thirteen participants in Hoque et al. 2024), leaving open the question of how co-creation behaves under real-world constraints such as time pressure or conflicting user goals.

In sum, interactive authoring represents a promising design approach to expanding human–AI collaboration, but current work underestimates the dynamics of control and overlooks users' cognitive limitations.

Explanation-centered approaches Explanation-centered approaches aim to make LLM behavior more interpretable by providing human-understandable justifications for model predictions, such as rationales (Mishra et al., 2024), contrastive explanations (Buçinca et al., 2025; Si et al., 2024), multilevel and contextualized explanations (Monteiro Paes et al., 2025; Mei et al., 2023; Di Bonaventura et al., 2024), anchored in situ explanations (Yan et al., 2024), explanations with different confidence levels (Wang et al., 2025), saliency explanations (Pafla et al., 2024) or visualization of internal states (Spinner et al., 2024), at various stages of interaction (Kim et al., 2025; Yao et al., 2023) to help users understand how and why a model generated a particular output.

The primary goal is to empower users to interpret, question, and calibrate trust in LLM outputs by integrating user-relevant explanations into the human-LLM interaction. Rather than being a post-hoc feature, explanations are regarded as an integral part of the user experience.

However, several design limitations remain underexplored. First, explanation quality is uneven, and users are often asked to trust model-generated justifications without support for interrogating the explanation itself. For instance, saliency maps or ranked rationales assume that the model's attention aligns with human reasoning, but users are not empowered to put this alignment into question. Most designs present a single explanation type, limiting opportunities for comparison (Pafla et al., 2024).

Second, explanation interfaces often rely on static visualizations or textual input. While a few designs allow users to manipulate explanations (e.g., editable search trees or contrastive comparisons), these remain exceptions. Moreover, explanations are usually presented as final, and users can not contribute to the model's reasoning. This risks reinforcing overreliance on explanation rather than promoting interactivity and critical engagement.

Third, the cognitive demands of interpreting explanations are often overlooked. Visualizations, importance heatmaps, or rationales may be challenging to interpret for non-experts or minoritized groups, and some studies suggest that users make better decisions with external references (e.g., Wikipedia) than with model-generated explanations (Si et al., 2024). The assumption that explanations automatically enhance trust or understanding must be validated across diverse user groups and domains.

Finally, most explanation-centered designs explain one output at a time (for example, why the model gave a specific answer), but they usually don't help users understand a general model behavior, such as whether the model is biased, how it was trained, or what kinds of mistakes it tends to make overall. An exception is Yao et al. (2023) where human-annotated explanations are integrated into active learning loops for annotation support, involving users in both training and evaluation phases.

In sum, while explanation-centered interfaces enhance transparency, they risk oversimplifying the complexity of LLM behavior and limiting user agency if not designed with deeper interactivity, explanation pluralism, and user education in mind.

Style-based trust calibration Style-based trust calibration refers to design strategies that shape users' trust in LLM outputs by varying the communicative style of the output. Rather than changing the factual content, these approaches manipulate how information is conveyed, for example, by presenting the output in an assertive or a hesitant tone, or showing confidence cues and visually marking lexical indicators of uncertainty to help users form more accurate mental models of LLM reliability. The central assumption is that stylistic framing and contextual cues strongly influence user reliance, perceived transparency, and decision confidence.

The primary goal is to support better alignment between perceived and actual model capabilities. This is especially crucial in settings involving uncertainty or risk, such as healthcare, legal advice, or career guidance.

Rather than improving accuracy directly, these interventions calibrate user perception of models. Studies have tested expressions of uncertainty (e.g., first-person: "I'm not sure..." vs. impersonal: "It is not sure...") (Kim et al., 2024b), confidence disclaimers (Metzger et al., 2024), comparing hesitant versus assertive tones (Kadoma et al., 2024), trust repair techniques through apologies, denials, and promises (Pareek et al., 2024), stylistic variations across chatbot types (LLM-based vs. intent-based vs. form-based) (Zylowski et al., 2025), uncertainty markers (Zhou et al., 2024; Chen et al., 2025b), model-generated greetings (Zhou et al., 2025b), and visual disclaimers or highlights (Bo et al., 2025). These features are tested in calibrated and miscalibrated scenarios to assess their influence on user trust.

However, most work remains narrowly focused on whether a given stylistic manipulation influences trust, rather than how users can be supported in recognizing and critically engaging with such cues in everyday use. For instance, while many features are shown to affect trust in experimental setups, they are rarely integrated into interface systems with guidance or educational scaffolding. Ma et al. (2025) address this by proposing a deliberation-based interface that encourages users to reason through LLM suggestions. Yet, dealing with insufficient analytical engagement of users with AI recommendations remains an exception.

A further limitation is the tension between helping users calibrate trust and the risk of unintentionally manipulating them or introducing new ethical problems. Specifically, stylistic cues may encode cultural or gender biases, reinforce stereotypes, or mask unreliable model behavior behind persuasive style. Future work should examine how style interacts with power, and whether certain user groups are more vulnerable to over-reliance due to stylistic calibration alone.

In conclusion, while style-based approaches offer promising mechanisms for aligning user-perceived trust with actual model reliability, they raise critical open questions about fairness and the long-term effects of such calibration.

Privacy-aware architectures and tools Privacy-aware architectures and tools are systems, interfaces, or frameworks that aim to detect, minimize, or prevent privacy risks in human–LLM interaction. They enhance user awareness, control, and protection by implementing privacy safeguards either before, during, or after data exchange with LLMs. These approaches consider input redaction, output inspection, system-level manipulation detection, and user education, often grounded in user-centered design and participatory development. Unlike general security methods, this category focuses on end-user-facing privacy measures, enabling users to actively participate in managing their personal data exposure and autonomy in LLM-mediated environments.

The primary goal is to empower users to manage and protect their personal data by providing controllable tools that mitigate privacy risks at every stage of the interaction pipeline. These systems aim to increase user agency and awareness while reducing unintended data leakage, over-disclosure, or manipulation in AI-mediated communication. They address not only what LLMs can "know" or "leak", but how users can actively participate in preventing harm and making informed choices about data use, visibility, and trustworthiness.

Core strategies in this paradigm span the full privacy lifecycle, from input-level privacy control (Ngong et al., 2025) through self-disclosure detection (Dou et al., 2024), user-led data minimization via browser extensions (Zhou et al., 2025a) to post-hoc inspection (e.g., leaking personal identifiers through LLM outputs Kim et al. 2024a or detecting prompt injection attacks Lin et al. 2025) and user education (Chen et al., 2025a).

However, these designs may face adoption challenges. Many tools assume that users are both willing and able to engage in privacy management, although users may sometimes prioritize convenience or utility over caution, especially in low-stakes contexts. Moreover, privacy-aware interfaces can disrupt the user experience if they demand too much time, technical understanding, or attention. To be effective, they must be carefully adapted to the context of use and the user's mental workload, for example by being paired with automation, personalization, or persuasive design. Finally, some designs risk offloading the responsibility for privacy onto the user without addressing underlying system-level weaknesses in how LLMs handle user data. For example, asking users to identify sensitive content assumes they understand what counts as risky in the context of opaque model behavior, but this assumption may not hold. It is also unclear how such tools perform across user groups with varying levels of sensitivity to privacy issues.

In sum, more research is needed to assess how to communicate privacy risks without overwhelming users or discouraging them from critical use of LLMs. Privacy-aware tools play a crucial role in shifting privacy control closer to users, but must be designed to balance protection, usability, and psychological trust across varied real-world scenarios.

4 Theoretical perspectives

To synthesize the design strategies identified through inductive coding, we draw on three complementary frameworks from HCI and cognitive science: Activity Theory (Kuutti, 1996), Distributed Cognition (Hollan et al., 2000), and Mental Models (see an overview in Payne, 2003). These descriptive theories are suited for analyzing user-centred paradigms across NLP and HCI research.

Activity theory highlights how users engage with LLMs as tools to achieve specific goals (e.g., writing, learning). It aligns closely with interactive authoring & co-creation and workflow-aligned designs where LLMs support domain-specific tasks (e.g., Masson et al. 2024; Kazemitabaar et al. 2024), enabling users to shift from passive prompting to active participation. Participatory designs also empower users by emphasizing their agency in shaping system behavior (e.g., Theophilou et al. 2023).

Distributed cognition frames trust as emerging from the interaction between the user, the LLM system, and the interventions (e.g., visualizations, warnings), such as in interface-level accuracy control (e.g., Leiser et al. 2024) and style-based trust calibration (e.g., Zhou et al. 2024). Trust calibration is distributed across the model's suggestions, system-generated evidence, and design interventions rather than by internal understanding alone.

Referring to users' internal understandings of how LLMs work, mental models are central to explanation-centered approaches (e.g., Yan et al. 2024) that aim to scaffold reasoning about model logic, privacy-aware designs (e.g., Dou et al. 2024) that help users understand what LLMs might infer from personal data, and style-based trust calibration, which influences users' conceptual models of LLM reliability.

Additionally, our classification aligns with the more recent human-centered AI (HCAI) framework proposed by Shneiderman (2022), particularly in treating user control not only as an outcome (product) but also as a participatory design process.

The proposed design paradigms also align with principles from classical HCI, such as Norman's gulfs of execution and evaluation (see Norman, 2013, 38–40), which describe the barriers users face in acting on and interpreting system behavior. Several designs aim to reduce Norman's gulf of execution by simplifying prompt design (Zhang and Arawjo, 2025) or providing scaffolds that guide users in expressing their intentions. Others address the gulf of evaluation by offering visualizations (Spinner et al., 2024) of model decisions or contrastive explanations (Buçinca et al., 2025) to help users interpret outputs. Furthermore, activity theory helps reduce the gulf of execution by analyzing whether users can meaningfully act on interfaces to achieve their goals. Distributed cognition addresses the gulf of evaluation by highlighting how trust and understanding are mediated through interface-level cues, external visualizations, and interaction history. Finally, mental models support both gulfs by determining how users understand what actions are possible and how outputs should be interpreted. Together, these theories provide a layered perspective on user control in LLM interactions.

5 Discussion and Conclusions

This paper identified and systematized seven design paradigms that promote user control in human–LLM interaction and reflect design strategies grounded in different user goals, ranging from verifying factuality and shaping model output to managing trust and data exposure. Our design-centered perspective complements current discussions on human involvement in post-training by emphasizing user control during deployment and interaction.

While empirical studies have offered scattered examples of user-centered designs and most recent related surveys do not primarily focus on trust or have a broader scope (e.g., human-model cooperation in Huang et al. 2025), our contribution lies in synthesizing these efforts into a coherent framework that centers user goals as the organizing principle of human trust in LLMs. Across paradigms, we observe a shift from one-shot prompting toward interactive, iterative, and increasingly user-configurable LLM workflows. These designs foreground a broad spectrum of control types: perceptual (e.g., accuracy cues), procedural (e.g., workflow pausing), epistemic (e.g., explanations, varying linguistic style), and protective (e.g., privacy screening).

Yet, critical gaps remain. Although many studies mention cross-domain application (e.g. Louie et al. 2024), the variety of tested scenarios is limited. We also observe a lack of design frameworks that help practitioners balance automation and human agency. For example, many tools mediate control through additional LLMs (e.g., Pan et al. 2024), which risks reinforcing automation bias rather than supporting user autonomy. To address this, future systems could incorporate trust calibration strategies (e.g., communicative framing, interactive uncertainty visualization) that help users reflect on when and how to trust outputs. Most studies assume AI-literate end users with a high level of technical literacy. Designs rarely account for diverse user needs, e.g., those with low reading/writing literacy, limited technical expertise, or from marginalized communities. This limits the accessibility and generalizability of proposed methods. Users are often expected to interpret complex cues (e.g., factuality scores) without training. It remains unclear how to prevent over-reliance on automation while avoiding user frustration and how to balance control vs. usability, or privacy vs. personalization.

For the research at the intersection of HCI and NLP, we identify several promising directions for future work:

- Explicitly address interaction design patterns that foster meaningful user oversight (e.g., modular prompt chaining, co-creation loops).
- Expand design efforts to underexplored trust dimensions (e.g., fairness, social well-being).
- Develop participatory methods that involve diverse users in the co-design of trust-aware LLM interfaces.
- Develop systems that calibrate trust in LLMs not only by LLMs but also include human-in-the-loop review.
- To support low-literacy users, consider, for example, visual metaphors to reduce cognitive burden or interaction logging, or user interfaces with a toggle to simplify responses.
- Replace binary on/off controls with graded or layered control (e.g., co-authoring steps or adjustable initiative).
- Move beyond controlled studies to assess how trust and control evolve during prolonged, real-world interaction (in-the-wild evaluation)
- Consider long-term, real-world deployment studies to assess how interaction designs shape trust over time.

Finally, we advocate for design that enables not just enhanced control, but critical engagement with LLM behavior, especially through scaffolds that support users in questioning and modifying model output.

6 Limitations

We identify three main limitations of this study. First, as this is a design-centered survey rather than a systematic meta-analysis, two types of constraints apply: those related to paper selection and those associated with the derivation of the proposed design typology. The final scope of included papers was based on a qualitative assessment by two annotators, followed by iterative discussion to reach consensus on inclusion. Consequently, not all papers containing search terms in the abstract, title, or keywords were included. Both the paper selection and the resulting classification are thus shaped by human judgment and interpretability. In particular, some papers at the boundary between metric-driven evaluation and user-centered design were included if they contained at least partial user evaluation components, such as in Koraş et al. (2025), where the user study was exploratory and not systematic. Although many papers could plausibly be assigned to multiple paradigms, annotators were instructed to assign each paper to a single primary category. The proposed design paradigms were qualitatively derived and require further empirical validation.

Second, due to limitations in the ACL Anthology search interface (see Section 2), it was not possible to apply an identical search string across both databases. While the ACM Digital Library search allowed for complex Boolean queries, the ACL Anthology search relied on simpler keyword combinations (see Section 2). This discrepancy may have introduced a bias by potentially missing relevant ACL papers that would have matched the full ACM query. A brief comparative test or validation of coverage was not feasible, but we acknowledge that this search asymmetry could affect the completeness and balance of the corpus. Furthermore, the review does not include papers from other sources such as arXiv, which means that unpublished or in-progress work was not considered.

Third, the reviewed studies are predominantly situated in English-speaking and Western contexts, as only papers published in English were included. This limits the cultural and linguistic diversity of the findings.

7 Ethical statement

This work is a meta-analysis of published research at the intersection of HCI and NLP. We do not present or process personal data, nor do we involve human participants. All surveyed papers were selected from publicly accessible, peer-reviewed sources, excluding preprints. Where user studies are reported in the cited literature, we rely on the original authors' ethical approvals and disclosures. Care was taken to fairly represent a diverse set of approaches and to avoid overgeneralizing results.

We acknowledge that relying solely on published, English-language sources may introduce publication and cultural bias, leading to an overrepresentation of Western perspectives. This is not only a methodological limitation (see Section 6), but also an ethical concern for the generalizability and inclusivity of our findings.

References

Mamia Agbese, Rahul Mohanani, Arif Khan, and Pekka Abrahamsson. 2023. Implementing AI ethics: Making sense of the ethical requirements. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, EASE '23, page 62–71, New York, NY, USA. Association for Computing Machinery.

Syeda Sabrina Akter, Seth Hunter, David Woo, and Antonios Anastasopoulos. 2025. Costs and benefits of AI-enabled topic modeling in P-20 research: The case of school improvement plans. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 460–476, Vienna, Austria. Association for Computational Linguistics.

Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. Chainforge: A visual toolkit for prompt engineering and LLM hypothesis testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. ACM.

Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To rely or not to rely? Evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2025. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-Jun Li, and Yaxing Yao. 2025a. Clear: Towards contextual LLM-empowered privacy policy analysis and risk generation for large language model applications. In *Proceedings of the 30th International*

Conference on Intelligent User Interfaces, IUI '25, page 277–297, New York, NY, USA. Association for Computing Machinery.

Cheng Chen, Sangwook Lee, Eunchae Jang, and S. Shyam Sundar. 2024. Is your prompt detailed enough? exploring the effects of prompt coaching on users' perceptions, engagement, and trust in text-to-image generative AI tools. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, pages 1–12, Austin TX USA. ACM.

Rex Chen, Ruiyi Wang, Norman Sadeh, and Fei Fang. 2025b. Missing pieces: How do designs that expose uncertainty longitudinally impact trust in AI decision aids? An in situ study of gig drivers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 790–816, New York, NY, USA. Association for Computing Machinery.

Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. Relic: Investigating large language model responses using self-consistency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, Honolulu HI USA. ACM.

Alexander S. Choi, Syeda Sabrina Akter, JP Singh, and Antonios Anastasopoulos. 2024. The LLM effect: Are humans truly using LLMs, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054, Miami, Florida, USA. Association for Computational Linguistics.

Jiwon Chun, Yankun Zhao, Hanlin Chen, and Meng Xia. 2025. Planglow: Personalized study planning with an explainable and controllable LLM-driven system. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale*, L@S '25, page 116–127, New York, NY, USA. Association for Computing Machinery.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9993–10001, Singapore. Association for Computational Linguistics.

Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara Mcgillivray. 2024. Is explanation all you need? an expert survey on LLM-generated explanations for abusive language detection. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 280–288, Pisa, Italy. CEUR Workshop Proceedings.

Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3321–3339, Singapore. Association for Computational Linguistics.

Jinhao Dong, Jun Sun, Wenjie Zhang, Jin Song Dong, and Dan Hao. 2025. Contested: Consistency-aided tested code generation with LLM. *Proc. ACM Softw. Eng.*, 2(ISSTA).

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754, Bangkok, Thailand. Association for Computational Linguistics.

Maximilian Förster, Mathias Klier, Kilian Kluge, and Irina Sigler. 2020. *Fostering human agency: A process for the design of user-centric XAI systems*.

Abhinav Gupta, Devendra Singh, Greig A Cowan, N Kadhiresan, Siddharth Srivastava, Yagneswaran Sriraja, and Yoages Kumar Mantri. 2025. AUTOSUMM: A comprehensive framework for LLM-based conversation summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 500–509, Vienna, Austria. Association for Computational Linguistics.

Neal R Haddaway, Matthew J Page, Chris C Pritchard, and Luke A McGuinness. 2022. PRISMA2020: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst. Rev.*, 18(2):e1230.

Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.*, 30(1):99–120.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. *Plan-then-execute: An empirical study of user trust and team performance when using LLM agents as a daily assistant*. Association for Computing Machinery, New York, NY, USA.

HLEG. 2019. Ethics guidelines for trustworthy AI. Expert-group report, European Commission, Directorate-General for Communications Networks, Content and Technology, Brussels.

James Hollan, Edwin Hutchins, and David Kirsh. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.*, 7(2):174–196.

Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark effect: Supporting provenance and transparent use of large language models in writing with interactive visualization. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA. ACM.

Yutong Hu, Kangcheng Luo, and Yansong Feng. 2024. ELLA: Empowering LLMs for interpretable, accurate and informative legal advice. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 374–387, Bangkok, Thailand. Association for Computational Linguistics.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Huang. 2025. How to enable effective cooperation between humans and NLP models: A survey of principles, formalizations, and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 466–488, Vienna, Austria. Association for Computational Linguistics.

Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner, and Carolin Lawrence. 2023. Walking a tightrope – evaluating large language models in high-risk domains. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 99–111, Singapore. Association for Computational Linguistics.

Kowe Kadoma, Marianne Aubin Le Quere, Xiyu Jenny Fu, Christin Munsch, Danaë Metaxa, and Mor Naaman. 2024. The role of inclusion, control, and ownership in workplace AI-mediated communication. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. ACM.

Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: A review. *ACM Comput. Surv.*, 55(2).

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. CodeAid: Evaluating a classroom deployment of an LLM-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20, Honolulu HI USA. ACM.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024a. ProPILE: Probing privacy leakage in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Nips '23, Red Hook, NY, USA. Curran Associates Inc.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024b. "I'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA. Association for Computing Machinery.

Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Osman Alperen Koraş, Rabi Bahnan, Jens Kleesiek, and Amin Dada. 2025. Towards conditioning clinical text generation for user control. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10549–10569, Vienna, Austria. Association for Computational Linguistics.

Kari Kuutti. 1996. *Activity theory as a potential framework for human-computer interaction research*, chapter 2. MIT press.

Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. 2024. Beyond the chat: Executable and verifiable text-editing with LLMs. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–23, Pittsburgh PA USA. ACM.

Christine P. Lee, David Porfirio, Xinyu Jessica Wang, Kevin Chenkai Zhao, and Bilge Mutlu. 2025. VeriPlan: Integrating formal verification and LLMs into end-user planning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A hallucination identifier for large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA. ACM.

Mark Liffiton, Brad E Sheese, Jaromir Savelka, and Paul Denny. 2023. CodeHelp: Using large language models with guardrails for scalable support in programming classes. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, pages 1–11, Koli Finland. ACM.

Weiran Lin, Anna Gerchanovsky, Omer Akgul, Lujo Bauer, Matt Fredrikson, and Zifan Wang. 2025. LLM whisperer: An inconspicuous attack to bias LLM responses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Jiahao Liu, Yiyang Shao, Peng Zhang, Dongsheng Li, Hansu Gu, Chao Chen, Longzhi Du, Tun Lu, and Ning Gu. 2025. Filtering discomforting recommendations with large language models. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 3639–3650, New York, NY, USA. Association for Computing Machinery.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How AI processing delays foster creativity: Exploring research question co-creation with an LLM-based agent. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. ACM.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA. Association for Computational Linguistics.

Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards human-AI deliberation: Design and evaluation of LLM-empowered deliberative AI for AI-assisted decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. DirectGPT: A direct manipulation interface to interact with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. ACM.

Alex Mei, Sharon Levy, and William Yang Wang. 2023. Foveate, attribute, and rationalize: Towards physically safe and trustworthy AI. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11021–11036, Toronto, Canada. Association for Computational Linguistics.

Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering calibrated (dis-)trust in conversational agents: A user study on the persuasive power of limitation disclaimers vs. authoritative style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, Honolulu HI USA. ACM.

Qijun Miao and Zhixuan Fang. 2025. User-side model consistency monitoring for open source large language models inference services. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11610–11622, Vienna, Austria. Association for Computational Linguistics.

Aditi Mishra, Sajjadur Rahman, Kushan Mitra, Hannah Kim, and Estevam Hruschka. 2024. Characterizing large language models as rationalizers of knowledge-intensive tasks. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8117–8139, Bangkok, Thailand. Association for Computational Linguistics.

Lucas Monteiro Paes, Dennis Wei, Hyo Jin Do, Hendrik Strobelt, Ronny Luss, Amit Dhurandhar, Manish Nagireddy, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Werner Geyer, and Soumya Ghosh. 2025. Multi-level explanations for generative language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32291–32317, Vienna, Austria. Association for Computational Linguistics.

Ivoline C. Ngong, Swanand Ravindra Kadhe, Hao Wang, Keerthiram Murugesan, Justin D. Weisz, Amit Dhurandhar, and Karthikeyan Natesan Ramamurthy. 2025. Protecting users from themselves: Safeguarding contextual privacy in interactions with conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26196–26220, Vienna, Austria. Association for Computational Linguistics.

Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, Ashley Harkin, and Mahesh Prakash. 2024. My climate advisor: An application of NLP in climate adaptation for agriculture. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 27–45, Bangkok, Thailand. Association for Computational Linguistics.

Don Norman. 2013. *The design of everyday things*, 2 edition. Basic Books, London, England.

Cassandra Overney, Daniel T Kessler, Suyash Pradeep Fulay, Mahmood Jasim, and Deb Roy. 2025. Coalesce: An accessible mixed-initiative system for designing community-centric questionnaires. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 366–389, New York, NY, USA. Association for Computing Machinery.

Marvin Pafla, Kate Larson, and Mark Hancock. 2024. Unraveling the dilemma of ai errors: Exploring the effectiveness of human and machine explanations for large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, and 7 others. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Syst. Rev.*, 10(1):89.

Qian Pan, Zahra Ashktorab, Michael Desmond, Martín Santillán Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-centered design recommendations for LLM-as-a-judge. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 16–29, TBD. ACL.

Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2024. Trust development and repair in AI-assisted decision-making during complementary expertise. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 546–561, New York, NY, USA. Association for Computing Machinery.

Stephen J Payne. 2003. Users' mental models: The very ideas. In *HCI Models, Theories, and Frameworks*, pages 135–156. Elsevier.

Vadim Perov and Vladislav Golovkov. 2024. Ethics documents in the field of AI. Concepts, achievements and problems. In *2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, volume 7, pages 196–199, Yekaterinburg. Eee.

Snehal Prabhudesai, Ananya Prashant Kasi, Anmol Mansingh, Anindya Das Antar, Hua Shen, and Nikola Banovic. 2025. "Here the GPT made a choice, and every choice can be biased": How students critically engage with LLMs through end-user auditing activity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-AI collaboration in auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, Aies '23, page 913–926, New York, NY, USA. ACM.

Mohi Reza, Ioannis Anastasopoulos, Shreya Bhandari, and Zachary A. Pardos. 2025. PromptHive: Bringing subject matter experts back to the forefront with collaborative prompt engineering for educational content creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press, London, England.

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024. Large language models help humans verify truthfulness – except when they are convincingly wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1459–1474, Mexico City, Mexico. Association for Computational Linguistics.

Thilo Spinner, Rebecca Kehlbeck, Rita Sevastjanova, Tobias Stähle, Daniel A. Keim, Oliver Deussen, and Mennatallah El-Assady. 2024. -generAItor: Tree-in-the-loop text generation for language model explainability and adaptation. *ACM Trans. Interact. Intell. Syst.*, 14(2).

Yi Tang, Chia-Ming Chang, and Xi Yang. 2024. PDFchatannotator: A human-LLM collaborative multi-modal data annotation tool for PDF-format catalogs. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, Iui '24, page 419–430, New York, NY, USA. ACM.

Emily Theophilou, Cansu Koyutürk, Mona Yavari, Sathya Bursic, Gregor Donabauer, Alessia Telari, Alessia Testa, Raffaele Boiano, Davinia Hernandez-Leo, Martin Ruskov, Davide Taibi, Alessandro Gabbiadini, and Dimitri Ognibene. 2023. Learning to prompt in the classroom to understand AI limits: A pilot study. In *AIxIA 2023 – Advances in Artificial Intelligence: XXIInd International Conference of the Italian Association for Artificial Intelligence, AIxIA 2023, Rome, Italy, November 6–9, 2023, Proceedings*, page 481–496, Berlin, Heidelberg. Springer-Verlag.

Usman Ahmad Usmani, Ari Happonen, and Junzo Watada. 2023. Human-centered artificial intelligence: Designing for user empowerment and ethical considerations. In *2023 5th international congress on human-computer interaction, optimization and robotic applications (HORA)*, pages 1–7, Istanbul. Ieee.

Xinru Wang, Mengjie Yu, Hannah Nguyen, Michael Iuzzolino, Tianyi Wang, Peiqi Tang, Natasha Lynova, Co Tran, Ting Zhang, Naveen Sendhilnathan, Hrvoje Benko, Haijun Xia, and Tanya R. Jonker. 2025. Less or more: Towards glanceable explanations for LLM recommendations using ultra-small devices. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 938–951, New York, NY, USA. Association for Computing Machinery.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*, pages 1–22, New Orleans LA USA. ACM.

Litao Yan, Alyssa Hwang, Zhiyuan Wu, and Andrew Head. 2024. Ivie: Lightweight anchored explanations of just-generated code. In *Proceedings of the*

2024 CHI Conference on Human Factors in Computing Systems, Chi '24, New York, NY, USA. ACM.

Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, and Dakuo Wang. 2023. Beyond labels: Empowering human annotators with natural language explanations through a novel active-learning architecture. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11629–11643, Singapore. Association for Computational Linguistics.

Jingyue Zhang and Ian Arawjo. 2025. Chainbuddy: An AI-assisted agent system for generating LLM pipelines. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. 2024. Prompt highlighter: Interactive control for multi-modal LLMs.

Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025a. LMLPA: Language model linguistic personality assessment. *Computational Linguistics*, 51:599–640.

Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhan Luo. 2025b. Customizing emotional support: How do individuals construct and interact with LLM-powered chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025a. Rescriber: Smaller-LLM-powered user-led data minimization for LLM-based chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025b. REL-A.I.: An interaction-centered approach to measuring human-LM reliance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11148–11167, Albuquerque, New Mexico. Association for Computational Linguistics.

Tiffany Zhu, Iain Weissburg, Kexun Zhang, and William Yang Wang. 2025. Human bias in the face of AI: Examining human judgment against text labeled as AI generated. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25907–25914, Vienna, Austria. Association for Computational Linguistics.

Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopál, Katarína Marcinčinová, and Matúš Mesarčík. 2025. Evaluation of LLM vulnerabilities to being misused for personalized disinformation generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 780–797, Vienna, Austria. Association for Computational Linguistics.

Thorsten Zylowski, Nathalia Sautchuk-Patricio, Wladimir Hettmann, Katharina Anderer, Karl Fischer, Matthias Wölfel, and Peter Henning. 2025. User study on the trustworthiness, usability and explainability of intent-based and large language model-based career planning conversational agents. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers*, ICETC '24, page 46–53, New York, NY, USA. Association for Computing Machinery.

A Appendix

Interface-level accuracy control. User-led verification based on consistency of LLM responses (Cheng et al., 2024); user control of LLM edits (Laban et al., 2024); user study of LLM-based planning systems (Lee et al., 2025); user-centered development of hallucination identifier for LLMs (Leiser et al., 2024). *Primary trust aspect:* Accuracy, transparency.

Workflow-aligned and domain-adapted AI assistance. LLM-assisted topic modeling for qualitative analysis (Akter et al., 2025; Choi et al., 2024); human–LLM collaboration for thematic analysis (Dai et al., 2023); LLM code generation with user feedback (Dong et al., 2025); human-in-the-loop conversation summarization for financial advisors (Gupta et al., 2025); plan-then-execute LLM collaboration with user-in-the-loop control (He et al., 2025); LLM-based legal advice with user intervention (Hu et al., 2024); human evaluation of LLM programming assistant (Kazemitabaar et al., 2024); human evaluation of LLM-generated texts in clinical settings (Koraş et al., 2025); user evaluation of LLM-based code assistance with guardrails (Liffiton et al., 2023). *Primary trust aspect:* Accuracy, transparency, oversight; auditability.

Participatory designs. Co-designed self-directed learning planner (Chun et al., 2025); co-design of roleplay prompts with domain experts (Louie et al., 2024); climate advice via co-designed LLM interaction (Nguyen et al., 2024); user involvement in the LLM-as-a-judge concept (Pan et al., 2024); comparing human and LLM judgements of cultural adaptability (Rao et al., 2025); collaborative prompt authoring interface for homework problems (Reza et al., 2025); AI literacy education (Theophilou et al., 2023); co-creation of chatbot personas for emotional reliance (Zheng et al., 2025b); user preference of texts with different labels (LLM-generated vs. human) (Zhu et al., 2025). *Primary trust aspect:* Reliability, fairness, bias.

Interactive authoring & co-creation. Interactive prompt engineering and evaluation (Arawjo et al., 2024); human–AI co-creation of news headlines (Ding et al., 2023); provenance-driven co-writing (Hoque et al., 2024); human–LLM co-creation of research questions (Liu et al., 2024); user-aligned co-filtering of discomforting recommendations (Liu et al., 2025); direct manipulation interface (Masson et al., 2024); human–LLM co-creation of questionnaires (Overney et al., 2025); end-user auditing scaffolds for identifying LLM biases (Prabhudesai et al., 2025); LLM-based human–AI auditing (Rastogi et al., 2023); human–LLM modular prompt chaining (Wu et al., 2022); LLM-based human–AI evaluation of LLM behavior (Zhang and Arawjo, 2025); LLM-assisted user evaluation of LLM personalities (Zheng et al., 2025a); human evaluation of LLM-generated personalized disinformation (Zugecova et al., 2025). *Primary trust aspect:* Reliability, transparency.

Explanation-centered approaches. User evaluation of LLM explanations for abusive language detection tasks (Di Bonaventura et al., 2024); user evaluation of contrastive explanations (Buçinca et al., 2025); impact of LLM explanations on user reliance (Kim et al., 2025); user evaluation of safety-related LLM rationales (Mei et al., 2023); evaluation of LLM rationale quality (Mishra et al., 2024); user study with multi-level model explanations (Monteiro Paes et al., 2025); user evaluation of human vs. XAI explanations (Pafla et al., 2024); user evaluation of LLM explanations and search engines (Si et al., 2024); user evaluation of tree-of-thought visualization (Spinner et al., 2024); in-situ anchored code explanations (Yan et al., 2024); human vs. LLM rationales (Yao et al., 2023); spatially structured and temporally adaptive explanations (Wang et al., 2025). *Primary trust aspect:* Explainability, transparency, reliability.

Style-based trust calibration. Reliance interventions (Bo et al., 2025); hesitant vs. self-assured auto-complete LLM suggestions (Kadoma et al., 2024); certain vs. uncertain LLM responses (Kim et al., 2024b); interactive AI–human deliberation (Ma et al., 2025); disclaimers + high vs. low authority style in LLM responses (Metzger et al., 2024); LLM-generated trust repair strategies (Pareek et al., 2024); LLM-generated emphatic expressions of politeness (Zhou et al., 2025b); LLM-generated uncertainty markers (Zhou et al., 2024). *Primary trust aspect:* Transparency, reliability, biases.

Privacy-aware architectures and tools. User-centered self-disclosure abstraction (Dou et al., 2024); threat model for user-centered mitigation of adversarial prompts (Lin et al., 2025); user-led data minimization (Zhou et al., 2025a); privacy-safeguarding intermediary between users and LLMs (Ngong et al., 2025). *Primary trust aspect:* Privacy.

TripleCheck: Transparent Post-Hoc Verification of Biomedical Claims in AI-Generated Answers

Ana Valeria González and **Sidsel Boldsen** and **Roland Hangelbroek**
Scientific Intelligence, AI and Digital Innovation
Novo Nordisk
{avqg, isdb, rlhb}@novonordisk.com

Abstract

Retrieval Augmented Generation (RAG) has advanced Question Answering (QA) by connecting Large Language Models (LLMs) to external knowledge. However, these systems can still produce answers that are unsupported, lack clear traceability, or misattribute information – a critical issue in the biomedical domain where accuracy, trust and control are essential. We introduce **TripleCheck**, a post-hoc framework that breaks down an LLM's answer into factual *triples* and checks each against both the retrieved context and a biomedical knowledge graph. By highlighting which statements are supported, traceable, or correctly attributed, TripleCheck enables users to spot gaps, unsupported claims, and misattributions, prompting more careful follow up. We present the TripleCheck framework, evaluate it on the SciFact benchmark, analyze its limitations, and share preliminary expert feedback. Results show that TripleCheck provides nuanced insight, potentially supporting greater trust and safer AI adoption in biomedical applications.

1 Introduction

Large Language Models (LLMs) augmented with retrieval, commonly referred to as Retrieval Augmented Generation (RAG), have significantly improved question answering (QA) by grounding responses in external sources. However, despite reducing hallucinations, these systems still exhibit key failures due to inherent system design constraints (Barnett et al., 2024).

In biomedical domains, especially in real-world industry, RAG is relatively underexplored (Bunnell et al., 2025; Ng et al., 2025) but distinct challenges have been pointed out, such as the lack of standard evaluation, unique ethical risks, and recurring problems with irrelevant or misleading information that hamper adoption in a field where both accurate and traceable information is crucial[1]. Additionally, inaccurate or outdated references can compromise the quality of generated responses (Amugongo et al., 2025; Gargari and Habibi, 2025).

Human-AI collaboration research stresses the need for interaction designs that keep users engaged and aware (Song et al., 2025). Without careful explanation mechanisms, users may become overreliant on AI systems (Vasconcelos et al., 2023; Kim et al., 2024; Passi et al., 2024; Zhang et al., 2020). Paradoxically, conventional explanation techniques can increase user trust even when the AI is wrong, elevating the risk of unsubstantiated but plausible-sounding answers (Bansal et al., 2021; González et al., 2021). This underscores the need for new approaches that better surface evidence and improve claim traceability.

To address these gaps and foster appropriate trust in biomedical QA, we propose a post-hoc verification layer that provides fine-grained evidence assessment. Biomedical fact-checking presents unique challenges: knowledge is constantly updated, and contextual nuance often determines the interpretation of evidence (Sosa and Altman, 2022). Overcoming these issues requires strategies that support more nuanced, context-aware evaluations.

We introduce **TripleCheck**, a system-agnostic post-hoc verification framework that can decompose AI-generated biomedical answers into factual triples and checks each for support within both the retrieved context and a large-scale biomedical knowledge graph that aggregates literature, patents, and clinical trials among other sources. This dual approach highlights statements that are supported, traceable, correctly attributed, and flags gaps such as misattributions or conflicting evidence from various sources. This can potentially help users recognize when to be skeptical or seek further evidence to make their own conclusions. By making the support and traceability of claims explicit, TripleCheck

[1] See UN News on WHO's warnings regarding generative AI in healthcare :https://news.un.org/en/story/2023/05/1136707

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 33–47
November 9, 2025

aims to calibrate user trust and promote safe reliance on AI answers. Our main contributions are:

- We present **TripleCheck**, a verification framework for biomedical QA that cross-checks answer claims with both retrieved context and a large-scale biomedical knowledge graph.
- We evaluate TripleCheck on a scientific claim verification benchmark (SciFact (Wadden et al., 2020)), showing robust performance against supervised and zero shot alternatives and provides interpretable evidence for each decision. Our analysis shows it disentangles both supported and unsupported information in complex answers.
- We discuss real-world applications and initial expert feedback, illustrating how TripleCheck has the potential to improve trust calibration, transparency, and traceability for workflows such as literature review and clinical QA.

2 Related Work

Scientific Claim Verification Automated fact-checking has progressed from general domains such as political news to specialized areas like biomedicine. Datasets like FEVER (Thorne et al., 2018) supported claim verification against Wikipedia, while SciFact (Wadden et al., 2020) introduced the challenge of verifying scientific claims using abstracts, spurring advances in both evidence retrieval and claim classification. SciFact-Open (Wadden and et al., 2022) broadened this to open-domain settings with over 500,000 abstracts, revealing that scientific evidence is often partial or ambiguous. Other resources have stressed the importance of explainability and evidence alignment for biomedical fact-checking (Kotonya and Toni, 2020; Sarrouti et al., 2021; Saakyan et al., 2021; Kumar et al., 2025).

Beyond traditional claim verification, recent efforts leverage knowledge graphs (KGs) to reduce factual errors, especially given their ability to systematically map relationships among biomedical entities. Notably, recent benchmarks (Lin et al., 2024) challenge AI agents to cross-verify KG-derived facts against the literature, revealing that even advanced LLMs often struggle with this task. Among KG-based approaches, MedGraphRAG (Wu et al., 2024) takes a fundamentally different approach by integrating a knowledge graph directly into the retrieval and generation process, aiming to produce answers that are verified at generation time. In contrast, TripleCheck acts as a post-hoc verification layer: it operates on the output of any generative QA system, requiring no modification or re-training, but instead adding an extra verification step to independently assess claim validity. This distinction means TripleCheck can complement methods like MedGraphRAG by providing an additional safety net.

Other methods propose post-generation claim checking, such as extracting claims from model outputs for KG validation (Guan et al., 2024), or hallucination detection using structured entailment checking over generated answer triples (Sansford et al., 2024). However, these works either do not leverage an external KG for cross-checking (as in (Sansford et al., 2024)), or they lack a user-facing explanation component (as in (Guan et al., 2024)). In contrast, TripleCheck not only combines text entailment and KG validation in a dual-evidence approach, but is also designed with user-understandability and interaction in mind.

Our approach builds on these directions by proposing a zero-shot, post-hoc verification layer that can be added on top of any generative QA system. We uniquely leverage a large-scale biomedical KG to robustly cross-validate atomic answer triples, inspired by recent work in open-domain QA and fact-checking (Li et al., 2025; Kamoi et al., 2023). Importantly, TripleCheck preserves the LLM's original answer, instead surfacing supporting or contradictory evidence for each claim so users can make informed, nuanced judgment – an essential feature in the evolving and often ambiguous landscape of biomedical research.

Trust Calibration and Explainable QA Interfaces Trust calibration – the process by which user trust aligns with the true reliability of an AI system – has emerged as a critical factor in medical AI adoption (Sakamoto et al., 2024). Effective calibration can improve decision accuracy, yet achieving it remains challenging, as trust depends on perceived understandability, technical competence, and system reliability (Darvish et al., 2024). Inadequate calibration, whether overtrust or undertrust, can lead to unsafe outcomes in high-stakes biomedical environments.

There is a growing consensus that AI systems in these domains must support user understanding and oversight through explainable interfaces (Liang and Sonntag, 2025). For example, Li et al. (2024) describe an LLM-assisted QA system with ex-

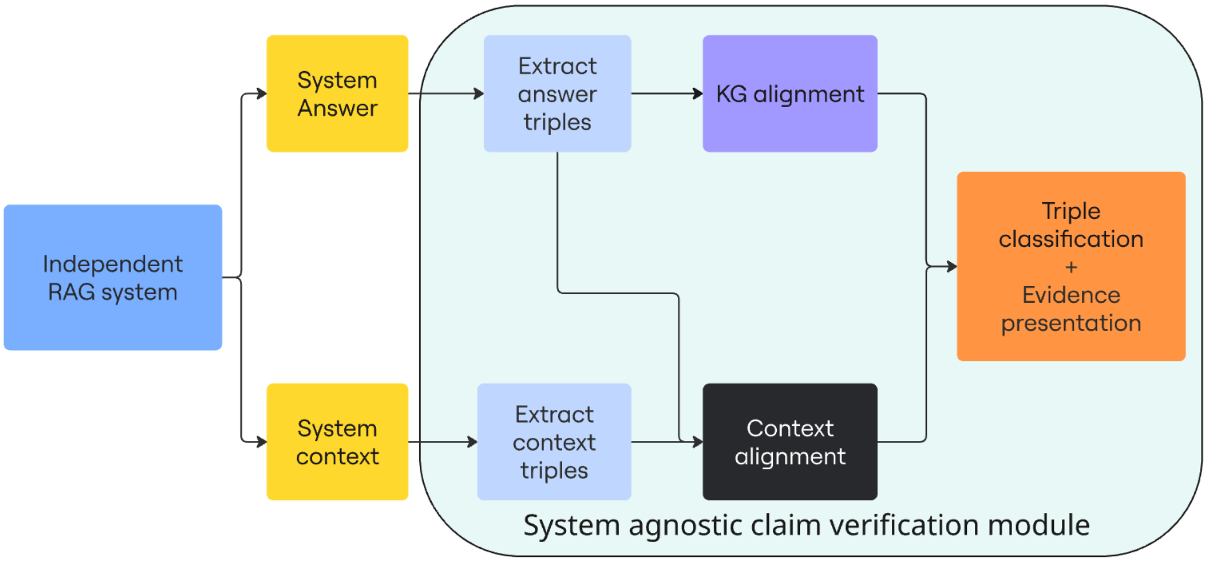


Figure 1: Overview of the TripleCheck pipeline. Given a user question and an answer from a RAG system, TripleCheck extracts atomic triples from the answer (and context) and verifies each one through two channels: (1) alignment with the retrieved context (documents or passages) and (2) cross-checking against a biomedical knowledge graph. Each triple is then labeled as supported, unsupported, or contradicted based on both evidence sources. This claim-level verification can be presented to the user as an interactive interface that highlights which parts of the answer are trustworthy and which require caution.

plicit KG integration for user control, while others caution that some explanations can inadvertently increase overtrust, even when the system is wrong (González et al., 2021; Bansal et al., 2021; Vasconcelos et al., 2023).

Effective interfaces feature interactivity, enabling users to explore not only the answer, but why and how it was produced. This approach helps foster appropriate skepticism and engagement (Rudin et al., 2022; Lai et al., 2023). In the biomedical domain, recent work by Huang et al. (2024) shows that providing multi-hop, interpretable rationales in a drug repurposing model, improved clinicians' accuracy, confidence, and decision efficiency, underscoring the value of transparent, actionable explanations. Similarly, tools such as claim verification with evidence trails (e.g., using SHAP) improve decisions, though risk overreliance without careful design (Liang and Sonntag, 2025).

While we do not fully explore the possibilities of building a sophisticated user interface in this work, TripleCheck is explicitly designed to provide users with the information needed to calibrate trust and promote informed oversight. By breaking down answers into checkable factual units, labeling each as supported, unsupported, or contradicted, and surfacing the underlying evidence from literature or knowledge graphs, TripleCheck offers fine-grained transparency. This enables users to scrutinize each claim with an appropriate level of skepticism or confidence, in line with findings within Human Computer Interaction (HCI) that emphasize user control as fundamental for trust calibration in AI (Passi et al., 2024).

3 Methodology: Post-hoc Claim Verification with TripleCheck

System Overview TripleCheck acts as a post-processor for a standard RAG pipeline. Suppose a user poses a question and the QA system produces an answer along with retrieved documents or passages as context. TripleCheck takes this answer and its supporting context as input, and performs three main steps: (1) Triple Extraction, (2) Evidence Alignment, and (3) Triple Classification. The output is a set of annotated triples derived from the answer, each marked with whether it is supported by the context and/or by the external knowledge graph along with any additional evidence surfaced. Figure 1 illustrates this workflow. As TripleCheck is system-agnostic and never alters the original answers, it can be flexibly added to any QA workflow to provide a second layer of verification.

Triple Extraction The first step breaks each answer into factual triples of the form (Subject, Predicate, Object). For example: from "A deficiency of vitamin B12 increases blood levels of homocysteine, which is a risk factor for heart disease," we extract *(vitamin B12 deficiency, increases, homocysteine levels)* and *(homocysteine, is a risk factor for, heart disease)*, each treated as an independent claim.

Our method follows recent approaches that combine large language models (LLMs) with post-hoc canonicalization of biomedical entities and relations (Zhang and Soh, 2024). It integrates two main strategies:

- **LLM-based parsing:** We prompt an LLM (GPT-4.1) with instructions (found in the appendix in Table 4, section A.2) to decompose the answer into concise factoid triples (`(Subject, Predicate, Object)`). The prompt is designed to focus on biomedical relations relevant to our KG and domain, and to avoid redundancy or overly broad statements. This captures implicit facts missed by more rigid parsers.
- **NER and RE:** In parallel, a pipeline for Named Entity Recognition (NER) and Relation Extraction (RE) identifies key biomedical entities (e.g., genes, chemicals, diseases) and the relations between them, restricted to a predefined ontology (e.g., "downregulates" , "upregulates", etc) present in our KG.

Candidate triples from both methods are merged, with further processing to expand abbreviations (e.g., "TNF" → "Tumor Necrosis Factor") and link entities to KG identifiers. Triples referencing novel or out-of-ontology entities are excluded from KG validation using relations, but retained for textual entailment-based checking. To reduce spurious alignments that could arise during the linking process, an LLM module screens for semantic consistency of the final triples to the system answer. The output is a set of cleaned, distinct factual triples asserted by the answer (see Table 1).

Contextual Evidence Alignment To measure the alignment between the answer and the context, TripleCheck evaluates whether each extracted triple is supported or refuted by the retrieved context. The triple extraction pipeline is also applied to the context documents, yielding sets of context triples for creating a similar structured comparison between claim and context as done by Sansford et al. (2024). For each answer triple, we attempt different matching strategies:

Original Claim	Extracted Triples
Albendazole is used to treat lymphatic filariasis.	(Albendazole, treats, Lymphatic filariasis)
DMRT1 is a sex-determining gene that is epigenetically regulated by the MHM region.	(DMRT1, associated with, sex determination) (MHM region, regulates , DMRT1)
Leukemia associated Rho guanine nucleotide-exchange factor represses RhoA in response to SRC activation.	(Rho guanine nucleotide exchange factor, inhibits, RhoA) (SRC activation, induces, Rho guanine nucleotide exchange factor)

Table 1: Original claims and their extracted triples. Relations and entities are additionally mapped to valid entities and relation types present in our KG.

- **Direct Support:** If the context triples set contains an identical triple to what is in the answer, the claim is marked as explicitly supported by the retrieved context.
- **No Support:** If the triple is absent in the context, it is initially treated as unsupported. However, as absence may result from novel or poorly linked entities, we leverage an LLM to assess if the context entails, contradicts, or is neutral toward the claim (instructions can be found in Table 3, section A.1). Entailment provides implicit support, contradiction triggers a warning, and otherwise the triple remains unsupported.

This strategy allows verification at the individual claim level, revealing when some aspects of an answer are substantiated while others are not.

Knowledge Graph Evidence Alignment TripleCheck simultaneously checks each triple against a biomedical KG that aggregates extracted relationships from sources like PubMed, clinical trials, and patents, among many others. We label extracted triples as:

- **KG-Supported:** If the triple or a suitable variant exists in the KG, we mark it as KG-supported. If the previous is not found, we additionally extract documents mentioning both entities in the triple and run the same textual entailment framework ran during the contextual evidence alignment step to reduce false negatives. We make the supporting evidence available.
- **KG-Contradiction:** If the KG records an opposing assertion (e.g., "A negative cause B" vs. the answer's "A positive cause B") via KG

relations or textual entailment method, we flag this as a contradiction and surface the relevant evidence.

- **KG-Unsupported:** If neither support nor contradiction is found, the claim is tagged as unsupported, suggesting either novel science, unsupported assertion or simply a gap in the KG.

Triple Classification By combining contextual evidence and KG-based validation, TripleCheck assigns each claim to one of four main verification categories:

1. **Fully Supported:** Found in both sources, indicating robust scientific consensus and proper attribution.
2. **Supported by KG Only:** Present in the KG but missing from retrieved context, flagging a retrieval or citation gap.
3. **Supported by Context Only:** Found in the context but not in the KG, pointing to possible new concepts or KG incompleteness.
4. **Unsupported:** Unsupported by either evidence channel, raising the possibility of a hallucination or unsubstantiated claim.

Additional flags are included for these cases:

1. **Contradicted in Context:** Explicitly contradicted by at least one retrieved passage, highlighting a likely error in system logic or misleading result.
2. **Contradicted in KG:** Contradicted by the knowledge graph, signaling the existence of contested information.

This fine-grained verification surfaces precisely which portions of an answer are reliable, unsupported, or contested, providing targeted feedback for both users and developers. TripleCheck never alters the original answer; users can decide how to act on verification results, while QA developers may use this information to improve retrieved citations and generation strategies.

Proprietary Components TripleCheck's implementation makes use of certain proprietary components. Specifically, our triple extraction pipeline relies on an in-house biomedical NER and RE system, trained on a broad mix of public biomedical annotations and internal corpora, to achieve

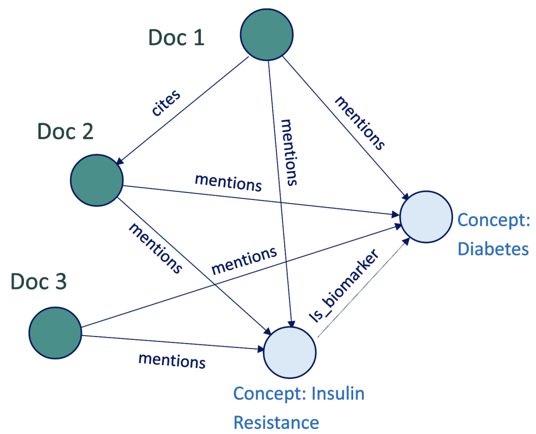


Figure 2: Simplified view of our proprietary KG: unstructured documents contain concept mentions and their relationships. We are able to trace in which documents specific relations are mentioned.

wide entity and relation coverage and high accuracy across the biomedical domain. The KG used for evidence alignment is constructed by aggregating structured relationships extracted through automated processes from scientific literature, clinical trial data, patents, and other specialized sources, some of which are not publicly available. An illustrative overview of the knowledge graph structure is shown in Figure 2. While these specific resources cannot be released due to licensing and privacy constraints, the overall TripleCheck framework is system-agnostic and designed for flexibility. Similar pipelines can be constructed using open-source biomedical NER/RE tools and knowledge graphs such as PrimeKG (Chandak et al., 2023). We encourage both academic and industry practitioners to build on or adapt our proposed framework with alternative resources, and view TripleCheck as an inspiration and blueprint for transparent, responsible biomedical QA in both open and proprietary environments.

4 Evaluation

We evaluated TripleCheck on the SciFact benchmark (Wadden et al., 2020), where claims are annotated as *Supported*, *Refuted*, or *NEI* (Not Enough Info). As access to the SciFact test set labels is no longer available [2], we perform evaluation on the development set similar to other studies (Deka et al., 2023). While we present results from several other methods on both test set and development set, our

[2]Evaluation on test set was only available via leaderboard which is now closed: `https://leaderboard.allenai.org/scifact/submissions/public`

Model	Precision	Recall	F1
Evaluated on SciFact Dev Set (Zero-Shot Setting)			
TripleCheck (ours)	0.73	0.70	0.70
PubMedBERT-mnli (Deka et al., 2023)	0.66	0.59	0.63
PubMedBERT-mnli-mednli (Deka et al., 2023)	0.84	0.75	0.79
DeBERTa-v3-base-mnli (Deka et al., 2023)	0.42	0.39	0.40
DeBERTa-v3-base-mnli-mednli (Deka et al., 2023)	0.78	0.70	0.74
Evaluated on SciFact Test Set			
Zero-NatVer (Strong et al., 2024) (zero-shot)	-	-	0.55
ClaimGen (entity-based) (Wright et al., 2022)	0.73	0.69	0.71
ClaimGen (BART) (Wright et al., 2022)	0.64	0.79	0.71
MultiVerS (Wadden et al., 2022) (weak-supervision)	0.73	0.71	0.72
VerT5erini (Pradeep et al., 2021)	0.64	0.73	0.68

Table 2: Fact verification results on SciFact. *Top:* All models evaluated on the development set in a zero-shot setting (i.e., **not** fine-tuned on SciFact train data). *Bottom:* Results on the test set, as reported in original publications; including zero-shot and weakly supervised approaches. Note: Due to test set access restrictions, only dev set results are shown for our approach.

key point is that TripleCheck delivers performance broadly in line with state-of-the-art alternatives, highlighting its practical competitiveness.

TripleCheck's output, though more fine-grained, is mapped for comparison: we label the entire data point (a claim from the scifact dataset) as *Supported* if all component triples are at least supported by the retrieved context and none are contradicted, *Refuted* if any triple is contradicted, and *NEI* otherwise. While this mapping is a simplification, it enables comparison on this benchmark.

TripleCheck achieves an F1 of 0.70 on SciFact (dev set) in a zero-shot setting without any task-specific fine-tuning, which is notable given that many comparison models, such as MultiVerS (Wadden et al., 2022) and VerT5erini (Pradeep et al., 2021), are tuned for this task. When comparing against other zero-shot approaches evaluated on the development set (Deka et al., 2023), TripleCheck achieves competitive performance, and outperforms strong baselines. This underscores TripleCheck's out-of-the-box robustness, even though our setup intentionally prioritizes transparency and explainability over strict optimization for SciFact. The results can be seen in Table 2.

Beyond aggregate scores, we also analyzed TripleCheck's outputs for cases where it provides nuanced judgments that classic fact-checkers might miss. We found that about 10% of SciFact's unsupported or contradicted claims were in TripleCheck's *Supported by KG Only* category. Upon inspection, it was evident that while the claim was not supported by the context, the claim was not a non factual claim, and we were able to collect evidence from the biomedical knowledge graph supporting this as an established fact. This reflects a traceability gap, highlighting where a claim may be true even if not cited. Our proposed step for such claims is to improve traceability by fetching additional data, rather than labeling the full claim as non factual. For example, in Scifact, the claim "A deficiency of vitamin B12 increases blood levels of homocysteine." is labeled as unsupported against the context, however, this is a known fact that is well supported in the KG.

4.1 Preliminary Feedback from a Domain Expert

While a full user study was beyond the scope of this research, we solicited preliminary feedback from a biomedical researcher to assess TripleCheck's practical value. The expert reviewed 30 claim-context pairs from the SciFact dataset, along with our system's triple-level evidence (see Figure 3 in the appendix). Feedback was collected via a structured questionnaire and follow up interview. Several key themes emerged:

Granular Verification and Trust Calibration. The domain expert confirmed that decomposing answers into factual triples substantially increased clarity and enabled a more nuanced, calibrated approach to trusting system outputs. Rather than treating each answer as a single unit, the triple-based breakdown highlighted exactly which subclaims were well-supported, which were missing evidence, and where there was explicit contradiction, echoing prior findings on the value of graph-based and evidence-traceable explanations in medical AI (Johnson et al., 2024). This allowed for more *selective skepticism* according to the expert: reliable portions of an answer could be accepted at face value, while unsupported or contested subsections triggered further review.

Role and Value of Knowledge Graph Support. Feedback emphasized that KG evidence often served as a crucial complement to retrieved context, especially for well-established biomedical facts that may not appear in the narrow selection of retrieved literature. The expert noted that, in practice, when an answer was supported *only* by the KG, they took it as a signal of a gap in retrieval coverage rather than a problem with the claim's validity, pointing to the fact that the user sees the

KG as a more objective and trustworthy source of truth. This aspect of traceability was highly valued for both – confirming canonical domain knowledge and helping efficiently flag true retrieval errors – demonstrating the importance of multi-channel verification over text-only methods. The distinction between KG-backed, context-backed, and unsupported can enable an action-oriented workflow: claims could be triaged for acceptance, additional investigation, or citation gap-filling.

UI Suggestions, and Information Overload. While the expert found the surfacing of supporting evidence to be confidence-boosting, to further reduce cognitive load and speed up review, UI suggestions were made such as: entity highlighting, displaying synonyms, and visually denoting the location of each triple within the evidence. Explanations that grew too detailed or technical could overwhelm non-specialists, consistent with recent findings on explanation overload (Hoffman et al., 2023). The expert also mentioned that as the goals change, the user might be interested in going deep into a topic, while at other times they want to get a high-level overview, therefore, controlling the level of depth and being able to explore and expand based on evidence could be useful. Finally, layered or toggleable presentation and simplified language were highlighted as desirable features.

Gold Standard Inconsistencies and Multiple Verification Channels. The expert occasionally detected that some claims labeled as *Supported* in SciFact were *not substantiated* by the provided abstracts, illustrating limitations of relying on single-source, gold-standard labels. This further supported the premise that multi-evidence verification is necessary to uncover gaps, avoid propagation of citation errors, and empower users to make cautious, context-sensitive decisions.

Taken together, this preliminary expert feedback strongly supports TripleCheck's approach to transparent, claim-level verification across multiple evidence channels. The integration of both KG and literature-derived support increases trust calibration, traceability, and user agency. The decomposition of answers not only aligns with real-world expert workflows but also makes the process of validation more actionable, helping users efficiently accept, investigate, or contest subclaims as needed. Comprehensive, interactive user studies remain a target for future work, but these results demonstrate significant potential for TripleCheck to promote safer and more reliable biomedical AI adoption.

5 Use Cases and Discussion

TripleCheck is broadly applicable to scenarios where users need to trust but verify AI-generated answers. We discuss a few use cases and their potential impact:

Literature Review Assistant: Researchers often use QA systems to quickly summarize findings across papers (e.g., "What causes condition X?"). TripleCheck would allow them to see which claimed causes are well-established (supported by multiple sources or KG) versus which are tentative more contested. It can also reveal if the system's answer includes claims not actually found in any cited papers, prompting the researcher to do a deeper dive for those claims.

Regulatory Document Drafting: In writing reports for drug approval or clinical guidelines, every statement needs a reference. An AI assistant could assist in drafting a section (e.g. drug efficacy) and TripleCheck would immediately flag any statement that lacks backing from the retrieved studies or known medical facts. This helps authors more quickly pinpoint those evidence gaps, saving time and preventing unsubstantiated claims from slipping through.

Clinical Decision Support: A clinician asking an AI assistant about treatment recommendations could benefit from TripleCheck's breakdown. For example, if the answer says "Drug A improves outcome Y and is not associated with side effect Z," TripleCheck might show the first claim is supported by a trial but the second claim is unsupported because the system didn't actually retrieve evidence about side effect Z. The clinician thus knows to be cautious or look up that specific point.

Improving QA System Development: TripleCheck can be used offline by developers of biomedical QA systems to analyze where the system tends to hallucinate or omit citations. If many answers have support only coming from the KG, it may mean the system is relying on prior knowledge not present in the retrieved text — maybe the retrieval component needs improvement. If many answers have "Unsupported" triples, the LLM might be overgeneralizing, suggesting a need for better grounding or post-editing.

Hypothesis Generation: Beyond verification, TripleCheck can assist in hypothesis generation by identifying claims that are plausible yet unsupported by the current evidence base. By inverting the verification output, users can systematically surface statements that are not confirmed in retrieved context or the knowledge graph. These unsupported claims can be then further investigated to see if they highlight potential gaps in scientific knowledge and serve as starting points for novel research questions.

By design, TripleCheck encourages a habit of verification. Rather than replacing human judgment, it guides users to the relevant evidence (or absence thereof). This aligns with the goal of safer deployment of AI in biomedicine: the human expert remains in the loop, making final decisions with a clearer view of the AI's reliability on each sub-point.

6 Limitations and Future Work

While promising, TripleCheck has several limitations:

- **Evaluation is still preliminary**: To date, we lack large-scale studies or professional user testing to validate the usability and benefits of TripleCheck. A crucial and active next step will be conducting a user study to quantitatively evaluate TripleCheck's impact on verification accuracy, confidence, and efficiency, similar to the approach of Huang et al. (2024), who assessed how interpretable explanations improved clinicians' decision-making. The next step is to compare users with and without access to TripleCheck as they assess AI-generated answers, thereby testing whether our framework enhances trust calibration and decision quality. This study will focus on three key outcomes: users' accuracy in claim verification, the time taken for assessment, and their confidence in their decisions.

- **User Experience Considerations**: Highlighting every claim in an answer can lead to information overload and overwhelm users. Careful interface design (e.g., toggleable detail levels) and user training are needed to ensure clarity. Tooltips or onboarding materials could assist users in interpreting verification results. Further exploration on how to build an efficient user interface is an area of future work.

- **Incomplete Knowledge Graph Coverage**: TripleCheck relies on a KG that, while extensive, is not exhaustive. It may lack very recent findings, rare conditions, or new technologies, leading to true claims being labeled as unsupported in KG. Expanding coverage and dynamically updating ontologies could have a positive impact.

- **Triple Extraction Quality**: The accuracy of information extraction directly affects downstream processing. Errors can occur with complex or explanatory sentences, leading to split, merged, or inaccurate triples. While an LLM verification step mitigates some issues, extraction errors can still cause correct claims to be labeled as unsupported and vice versa.

- **Added Latency and Complexity**: The pipeline introduces extra processing (LLM extraction, KG lookup, textual entailment verification) that increases latency. Processing each answer is slower compared to simpler QA systems, and optimizations may be needed for real-time applications.

- **Proprietary Resources**: As previously discussed, various components of TripleCheck are proprietary. While we provide our main TripleCheck system description to support reproducibility, this limitation may hinder exact replication by the research community. As an area of future work, we aim to benchmark public alternatives on fully open resources and encourage efforts to develop analogous public alternatives.

7 Conclusion

We presented **TripleCheck**, a post-hoc verification framework for biomedical QA that decomposes LLM-generated answers into factual triples and verifies each against both retrieved context and a large-scale biomedical knowledge graph. Our SciFact evaluation demonstrated that TripleCheck achieves competitive zero-shot performance while providing fine-grained, interpretable evidence for each claim. Initial expert feedback also suggested that this approach can support more calibrated trust, improve detection of unsupported or contested claims, and aid decision-making in biomedical settings. This initial feedback aligns well with anticipated real-world use, supporting the practical value of TripleCheck in biomedical workflows.

While promising, TripleCheck faces challenges such as refining user interfaces to manage information load, and expanding coverage of supporting knowledge. Most notably, future user studies are necessary to measure TripleCheck's real-world impact on verification accuracy and user confidence.

TripleCheck represents a step toward more transparent, accountable biomedical AI by offering actionable, triple-level evidence to end users and developers. We hope this work encourages further development of evidence-aware QA frameworks, advancing safe and trustworthy use of AI in biomedicine.

Acknowledgments

We are grateful to Zhang Zhong for generously sharing domain expertise and insightful feedback on the practical value of TripleCheck. We also thank colleagues from the Scientific Intelligence team at Novo Nordisk for helpful discussions throughout the development of this work. Their comments and suggestions contributed to refining our approach. Finally, we acknowledge the anonymous reviewers whose thoughtful critiques will shape future iterations of this research.

References

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, CAIN '24, page 194–199, New York, NY, USA. Association for Computing Machinery.

David Bunnell, Mary Bondy, Lucy Fromtling, Emilie Ludeman, and Krishnaj Gourab. 2025. Bridging ai and healthcare: A scoping review of retrieval-augmented generation - ethics, bias, transparency, improvements, and applications.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Mahdieh Darvish, Jan-Hendrik Holst, and Markus Bick. 2024. Explainable ai in healthcare: Factors influencing medical practitioners' trust calibration in collaborative tasks.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2023. Multiple evidence combination for fact-checking of health-related information. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 237–247, Toronto, Canada. Association for Computational Linguistics.

Omid Kohandel Gargari and Gholamreza Habibi. 2025. Enhancing medical ai with retrieval-augmented generation: A mini narrative review. *Digital health*, 11:20552076251337177.

Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do explanations help users detect errors in open-domain qa? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1103–1116.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18126–18134.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257.

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N Nadkarni, Benjamin S Glicksberg, Nils Gehlenborg, and Marinka Zitnik. 2024. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613.

Ruth Johnson, Michelle M Li, Ayush Noori, Owen Queen, and Marinka Zitnik. 2024. Graph artificial intelligence in medicine. *Annual review of biomedical data science*, 7(2024):345–368.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. 2025. Sciclaimhunt: A large dataset for evidence-based scientific claim verification. *arXiv preprint arXiv:2502.10003*.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1369–1385.

Harry Li, Gabriel Appleby, and Ashley Suh. 2024. Linkq: An llm-assisted visual interface for knowledge graph question-answering. In *2024 IEEE Visualization and Visual Analytics (VIS)*, pages 116–120. IEEE.

Xiangci Li, Sihao Chen, Rajvi Kapadia, Jessica Ouyang, and Fan Zhang. 2025. Minimal evidence group identification for claim verification. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 103–111, Albuquerque, New Mexico. Association for Computational Linguistics.

Siting Liang and Daniel Sonntag. 2025. Explainable biomedical claim verification with large language models. *arXiv preprint arXiv:2502.21014*.

Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z Li, and Kaicheng Yu. 2024. Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science. *arXiv preprint arXiv:2407.00466*.

Karen Ka Yan Ng, Izuki Matsuba, and Peter Chengming Zhang. 2025. Rag in health care: A novel framework for improving communication and decision-making by addressing llm limitations. *NEJM AI*, 2(1):AIra2400380.

Samir Passi, Shipi Dhanorkar, and Mihaela Vorvoreanu. 2024. Appropriate reliance on generative ai: Research synthesis. Technical report, Technical Report MSR-TR-2024-7. Microsoft. https://www. microsoft. com/en-us ....

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific claim verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Tetsu Sakamoto, Yukinori Harada, Taro Shimizu, and 1 others. 2024. Facilitating trust calibration in artificial intelligence–driven diagnostic decision support systems for determining physicians' diagnostic accuracy: Quasi-experimental study. *JMIR Formative Research*, 8(1):e58666.

Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. 2024. Grapheval: A knowledge-graph based llm hallucination evaluation framework. *arXiv preprint arXiv:2407.10793*.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jaeyoon Song, Zahra Ashktorab, Qian Pan, Casey Dugan, Werner Geyer, and Thomas W Malone. 2025. Interaction configurations and prompt guidance in conversational ai for question answering in human-ai teams. *arXiv preprint arXiv:2505.01648*.

Daniel N Sosa and Russ B Altman. 2022. Contexts and contradictions: a roadmap for computational drug repurposing with knowledge inference. *Briefings in bioinformatics*, 23(4):bbac268.

Marek Strong, Rami Aly, and Andreas Vlachos. 2024. Zero-shot fact verification via natural logic and large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17021–17035, Miami, Florida, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38.

David Wadden and et al. 2022. Scifact-open: Towards open-domain scientific claim verification. In *Findings of EMNLP 2022*, pages 347–359.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2596–2611.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An LLM-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836, Miami, Florida, USA. Association for Computational Linguistics.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305.

A Example prompts

A.1 Entailment/contradiction prompt

In Table 3, we show the prompt used for assessing textual entailment at different stages. We used the same prompt to verify final triples are aligned with system answer, to verify the triple is aligned with the context and to verify that the triple is aligned to any external evidence we have found via the KG.

A.2 Triple extraction

In Table 4, we show the instructions used for extracting initial triples. The initial triples were additionally linked to ontology terms using our proprietary entity linking system, and were afterwards verified against the actual claim to ensure consistency in the final triples set.

B Expert feedback questionnaire

Figure 3 shows how evidence was initially presented to the expert for initial feedback on what could make the result more useful. In Table 5, we have additionally compiled some of the key comments coming both from the written feedback and interview categorized into themes.

System Prompt for LLM-based textual entailment

```
You are a claim-verification system.
      Your task is to determine
     whether the given statement is
     supported (directly, indirectly
     , or can be reasonably inferred
     , even if this requires
     combining context and general
     biological knowledge) by the
     provided context.

CONTEXT:
{context}

STATEMENT TO VERIFY:
"{statement}"

VERIFICATION RULES:
1. Answer "YES" if the statement is
     supported by the context, can
     be logically inferred from the
     context, **or if it is
     biologically plausible and
     consistent with accepted
     scientific background knowledge
     .** You can accept reasonable
     combinations of entities as
     long as the overall logic is
     supported, even if not every
     link is explicitly present in
     the context. Do not be overly
     strict about requiring explicit
      verbatim phrasing or full
     mechanistic details **favor a
      positive answer if the overall
      claim is well-supported or
     reasonably implied.**
2. Answer "CONTRADICTION" if the
     statement clearly contradicts
     the context.
3. Answer "NO" only if there is
     insufficient information, the
     claim is irrelevant, or
     biological plausibility is
     seriously lacking or unclear.

RESPONSE FORMAT:
Begin with "YES", "CONTRADICTION",
     or "NO" on its own line. Do not
      start in any other way.
Then provide a brief, evidence-based
      explanation that quotes or
     paraphrases relevant portions
     of the context and/or uses well
     -accepted biological background
      if relevant.

YOUR VERIFICATION:
```

Table 3: System prompt for textual entailment as used in this work.

System Prompt for LLM-based Triple Extraction

```
You are an expert extracting entities and relations from scientific text.

    Given an answer to a scientific question, extract the claims in triples format.

    Your output must be a valid JSON array containing exactly one object per triple in this format:
    [["subject1", "relation1", "object1"], ["subject2", "relation2", "object2"], ...]
    **CRUCIAL RULES     READ CAREFULLY:**
    1. Do NOT use intervention phrases, experimental treatments, or contextual language as entities:
        - Disallow: "PARN targeting", "PARN inhibition", "knockout of PARN", "overexpression of X", "
            activation of Y"
        - Allow only: the core biological entity/process itself (e.g., "PARN", "TP53", "insulin
            maturation")

    2. Use concise, ontology-friendly names (2-4 words max), established biomedical terms, no
         abbreviations unless standard.

    3. DO NOT encode intervention or experiment type in subject/object. NEVER use experimental
         manipulation phrases as entities.

    - Do not use long descriptive phrases or qualifiers as entities.
    - Use 2-4 words maximum for each entity, and keep them concise and ontology-friendly.
    - Use established gene names, protein names, disease terms, and biological processes if possible.

    **Direct Entity-Relation-Entity Guidance:**
    - PREFERRED: ['Gene X', 'Directed Link', 'Process Y']
    - AVOID:    ['Knockdown of Gene X', 'Directed Link', 'Upregulated Process Y']
    - Do not build effects (like "loss", "increase", or "compromised state") into the entity. Use the
          proper relation instead.

    **Decompose Complex Entities:**
    - Break up complex cause-effect phrases into multiple, simpler, functionally meaningful triples
         only using entities present in standard biomedical ontologies.

    **Relation Types (use only these) but keep in mind the mentioned above:**
        -Focus mostly on these
        Directed Link: Direct interaction between entities. Can include correlations or associations.
              IMPORTANT When in doubt of direction, use this.
        Negative Cause: Causes a decrease or inhibition in the target entity.
        Not Directed Link: Interaction without specified direction.
        Not Negative Cause: Does not lead to a negative effect.
        Not Positive Cause: Does not lead to a positive effect.
        Positive Cause: Causes an increase or stimulation in the target entity.

        PPI (Protein-Protein Interaction): Interaction affecting protein function.
        DDI (Drug-Drug Interaction): Interaction affecting drug effectiveness.

        - These can be used to but limit them
        ACTIVATOR: Increases activity of a process or molecule.
        AGONIST: Initiates response by combining with a receptor.
        AGONIST-ACTIVATOR: Initiates and enhances activity.
        AGONIST-INHIBITOR: Acts as agonist and inhibitor.
        ANTAGONIST: Inhibits physiological action of another.
        DIRECT-REGULATOR: Directly modulates target activity.
        INDIRECT-DOWNREGULATOR: Indirectly decreases target activity.
        INDIRECT-UPREGULATOR: Indirectly increases target activity.
        INHIBITOR: Slows or prevents chemical reactions.
        PART-OF: Entity is a component of a larger structure.
        PRODUCT-OF: Entity is a result of a process.
        SUBSTRATE: Molecule acted upon by an enzyme.
        SUBSTRATE_PRODUCT-OF: Substrate converted into a product.
        undefined: Relationships not yet characterized or classified in this list but are still valid
            .

    - For abbreviations, prefer the full name if confidently available from context.
    - Both subject and object must be concise entities/concepts, not specific statements, modifiers,
         or experimental constructs.
    - Do NOT repeat triples (even if synonyms are used in the text).
    - If none of the relations are present, use "undefined". Do NOT invent new relations.
    - If there are no triples present, return [].
    - Your output must be valid JSON directly parsable by `json.loads()` as a list of triple lists (
         not nested or with extra structure)  e.g. [["subject1", "Directed Link", "object1"], ["
         subject2", "Part-Of", "object2"]].
    - Do NOT include any explanations or text outside the JSON array.
```

Table 4: System prompt for triples extraction as used in this work.

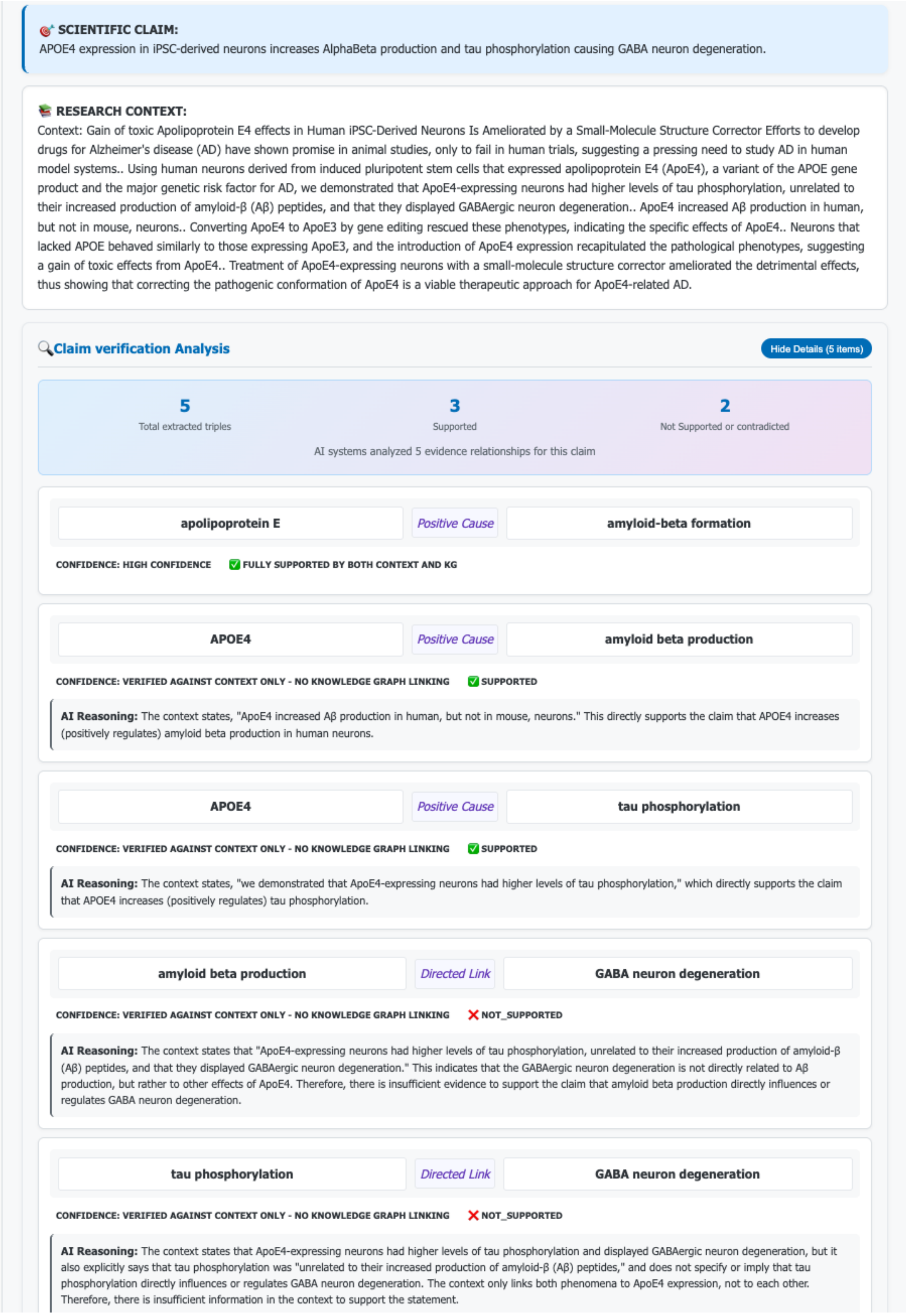


Figure 3: User feedback interface. While the intended use is in an interactive QA setting, this preliminary study presented the interface in a static, questionnaire format to collect initial expert feedback. Future work will focus on exploring different presentation formats and interactive modes.

Theme: Evidence & Explanation Quality
"The explanation directly referenced the supporting evidence, which was helpful."
"The additional reasoning summarized relevant points well, presented additional evidence and matched my interpretation."
"Sometimes the explanation focused too much on specific concepts, making it less broadly useful."
"Having both 'supported' and 'contradicted' reasoning was logical; it's important to consider context, while ultimately I would say this is a supported claim, as an expert I have similar concerns as the contradictions surfaced."
"If I feel an answer is incomplete or uncertain, I'll ask for more detail or reasoning before accepting it."
"If the context doesn't really support the claim, I become wary and might not trust that part of the answer. So having the additional evidence is key"
Theme: User Interface & Usability
"Highlighting the triple location or the keywords like gene names or important biological terms would help me quickly locate evidence in the text."
"Some explanations were overly technical or as difficult to follow as the original literature. It could be nice to choose how deep to go yourself"
"It would help to see synonyms of entities or have key parts of the triple highlighted directly in the evidence."
"Claims can be hard to understand if you are not an expert in the topic, simplified breakdowns or highlights would make it easier."
Theme: Knowledge Graph (KG) Value
"KG support was very useful, especially when the retrieved context didn't cover established facts."
"Recognizing when information is canonical, even if not in the provided context, adds confidence."
"I often trust facts from the KG more, especially when the answer is missing context evidence, it gives reassurance about general scientific truth."
"Sometimes, the KG picked up on a missing fact from the literature, and that signaled an issue with the context rather than a problem with the claim itself."
Theme: Exploration & Workflow
"If most triples are supported, I move on, if any aren't, I dig deeper or ask for more sources. Seeing the breakdown helps me focus."
"Having access to more detailed evidence when I want it, without being overwhelmed would make deciding whether claims are true or not easier."
Theme: Areas for Improvement
"Going through the retrieved context can already be complex, so have a simplified language in the breakdown would be helpful."
"Balance between pointing out specifics and giving a general overview in reasoning. User should be able to choose how deep to go into the details"

Table 5: Sample categorized expert feedback (paraphrased) from the TripleCheck evaluation.

Rethinking Search: A Study of University Students' Perspectives on Using LLMs and Traditional Search Engines in Academic Problem Solving

Md. Faiyaz Abdullah Sayeedi[1], Md. Sadman Haque[1], Zobaer Ibn Razzaque[1], Robiul Awoul Robin[1], Sabila Nawshin[2]

[1]United International University, Bangladesh [2]Indiana University Bloomington, USA

Emails: {msayeedi212049,mhaque221592,zrazzaque221135,rrobin221564}@bscse.uiu.ac.bd; snawshin@iu.edu

Abstract

With the increasing integration of Artificial Intelligence (AI) in academic problem solving, university students frequently alternate between traditional search engines like Google and large language models (LLMs) for information retrieval. This study explores students' perceptions of both tools, emphasizing usability, efficiency, and their integration into academic workflows. Employing a mixed-methods approach, we surveyed 109 students from diverse disciplines and conducted in-depth interviews with 12 participants. Quantitative analyses, including ANOVA and chi-square tests, were used to assess differences in efficiency, satisfaction, and tool preference. Qualitative insights revealed that students commonly switch between GPT and Google: using Google for credible, multi-source information and GPT for summarization, explanation, and drafting. While neither tool proved sufficient on its own, there was a strong demand for a hybrid solution. In response, we developed a prototype, a chatbot embedded within the search interface, that combines GPT's conversational capabilities with Google's reliability to enhance academic research and reduce cognitive load.

1 Introduction

The rapid advancement of artificial intelligence has significantly reshaped the ways in which university students seek academic information and engage in research activities (Pirzado et al., 2024). Traditionally, search engines like Google have served as the dominant tool for retrieving scholarly content due to their accessibility, breadth of indexed materials, and access to verified sources. However, the emergence of LLMs, such as OpenAI's ChatGPT, has introduced a new paradigm—offering students direct, conversational responses and contextualized summaries that can streamline information consumption (Alberth, 2023).

This evolution in digital research tools raises important questions about how students perceive and utilize these systems, particularly in academic settings where accuracy, credibility, and efficiency are critical. Prior research suggests that while LLMs facilitate rapid content summarization and task-specific assistance, their reliability varies depending on context and task complexity (Divekar et al., 2024; Xu et al., 2023). Conversely, search engines provide access to a wide range of authoritative sources but often require users to sift through multiple links and evaluate conflicting information independently. Several studies have documented this complementary behavior—students tend to use LLMs for explanation and drafting while relying on search engines for fact-checking and source validation (Caramancion, 2024; Spatharioti et al., 2023).

Despite their respective strengths, both tools also have well-documented limitations. LLMs can generate confident yet incorrect outputs, potentially misleading users (Xu et al., 2023), while traditional search engines can lead to information overload and inefficiency in time-sensitive academic contexts. As a result, students are increasingly adopting a hybrid approach—strategically switching between LLMs and search engines to balance speed with credibility (Sakirin and Said, 2023; Kapoor et al., 2024). However, this constant toggling between tools introduces cognitive overhead and fragmented workflows, especially when performing complex academic tasks.

To investigate these dynamics systematically, this study addresses the following research questions:

- **RQ1:** How do university students perceive the usability, efficiency, and satisfaction of LLMs compared to traditional search engines in academic problem-solving?
- **RQ2:** What patterns of tool usage emerge when students perform academic tasks with

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 48–59
November 9, 2025

either or both tools?

- **RQ3:** What are students' preferences and expectations for an integrated solution that combines the strengths of both systems?

2 Literature Review

LLMs have significantly reshaped how individuals learn, make decisions, and retrieve information. While traditional search engines like Google have long been the primary tool for academic information seeking, recent research increasingly explores how LLMs compare in terms of usability, task performance, and user trust. Divekar et al. (2024) examined how university students use LLMs like ChatGPT alongside traditional search engines for learning new topics. Their findings indicate that while LLMs support rapid summarization and ease of understanding, their effectiveness varies depending on the complexity and nature of the task. In a similar vein, Kumar et al. (2024) analyzed how students use LLMs to generate SQL queries. They observed that LLM assistance improved query formulation and contributed positively to the learning experience.

Several studies have also investigated task completion performance. Spatharioti et al. (2023) conducted a randomized experiment and found that LLM users completed decision-making tasks more quickly and with fewer queries. However, the authors warned of a major drawback: users often overtrust LLM outputs, especially when incorrect answers are presented confidently. They suggested the inclusion of confidence indicators to mitigate this issue. Xu et al. (2023) echoed this concern, emphasizing the need for rigorous fact-checking when relying on LLM responses.

In terms of task preference, Caramancion (2024) evaluated 20 types of information-seeking scenarios and concluded that users favored traditional search engines for fact-based queries, while preferring LLMs for creative or complex tasks. Supporting this, Sakirin and Said (2023) found that nearly 70% of participants preferred ChatGPT-style conversational interfaces due to their personalization, perceived efficiency, and convenience. Extending these findings, Wazzan et al. (2024) studied image geolocation tasks and observed that tool selection often influenced user strategy: LLMs were used more intuitively, while traditional tools required structured navigation.

The issue of credibility remains central. Kapoor et al. (2024) argued that despite the convenience and rapidity of AI tools, traditional search methods remain more reliable for academic research. In contrast, LLMs often lack source transparency, which can be problematic in scholarly settings. To address this trade-off, researchers have proposed hybrid models. Bal and Nath (2009) explored metasearch engines that aggregate content from various sources to improve accuracy, and Caramancion (2024) advocated for systems that combine the contextual depth of LLMs with the source validation strengths of search engines.

However, existing studies have primarily evaluated LLMs and traditional search engines in isolation or through task-specific comparisons, without fully exploring how students naturally combine both tools in academic workflows (Xu et al., 2023). There is a lack of empirical research that integrates both performance metrics and user perspectives to understand this hybrid usage behavior (Bansal, 2023). While tools like Perplexity AI (Perplexity AI, 2024) attempt to bridge this gap by combining AI-generated responses with source links, and Google has introduced AI summaries through its Search Generative Experience (SGE), these systems remain largely static, lacking personalization, real-time adaptation, and task-specific reasoning. This study addresses these limitations through a mixed-methods approach and the design of a user-informed, context-aware prototype.

3 Methodology

To explore university students' preferences and usage behaviors regarding LLMs and traditional search engines for academic tasks, we employed a mixed-methods approach that combined quantitative and qualitative data collection and analysis, shown in Figure 1. This design allowed us to examine both broad patterns and deeper user experiences in a complementary manner.

We first conducted an online survey that collected responses from 109 university students across a range of academic disciplines. While the survey primarily targeted students in technology-related fields such as Computer Science and Engineering (CSE), Electrical and Electronics Engineering (EEE), and Data Science, it was also distributed to students from other areas, including Business Administration and Medicine, to ensure diversity. The questionnaire included both closed- and open-

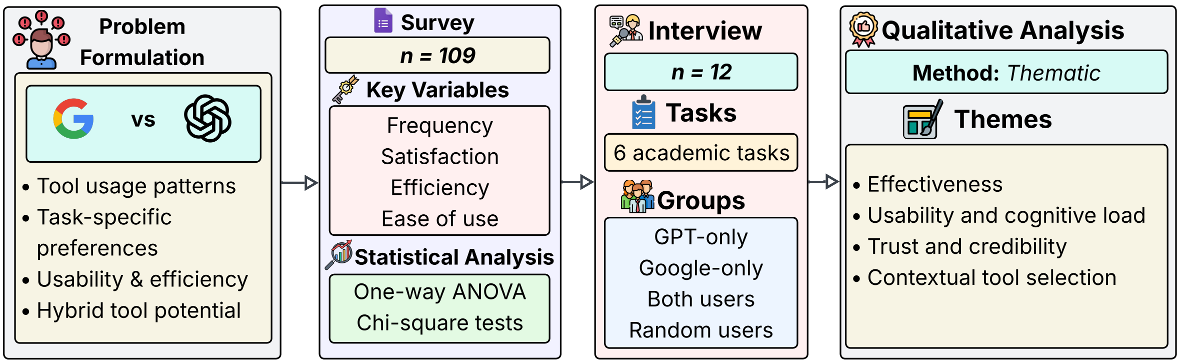


Figure 1: Overview of the study methodology. A mixed-methods approach was employed in this study. (1) The survey phase ($n = 109$) captured quantitative data and analyzed using different statistical tests. (2) The qualitative phase included in-person interviews ($n = 12$), where participants completed six academic tasks and were grouped based on tool usage. Thematic analysis of open-ended responses and interview transcripts led to four themes.

ended questions designed to assess tool usage frequency, satisfaction, efficiency, and perceived ease of use when using GPT-based LLMs and traditional search engines like Google. Descriptive statistics were used to summarize the data, and inferential statistical tests, one-way ANOVA and chi-square tests, were employed to evaluate differences in user perceptions and the influence of demographic variables on tool preference.

To enrich and validate the survey findings, we conducted in-depth in-person interviews with 12 students from CSE, EEE, and BBA backgrounds, primarily recruited from United International University (UIU), Dhaka, Bangladesh. Each participant was asked to complete six academic tasks: summarizing a research article, solving a coding problem (only for CSE students), addressing a circuit-related issue (only for EEE students), analyzing business data (only for BBA students), drafting a formal academic email, and comparing two popular academic concepts. Participants were divided into four groups based on their tool usage behavior: GPT-only users, Google-only users, balanced users who alternated between both tools, and random-choice users who freely switched between GPT and Google depending on preference.

The qualitative analysis synthesized insights from both the open-ended survey responses and interview transcripts. Thematic coding focused on perceived effectiveness, usability, trustworthiness, and the contextual factors that influenced tool selection. This analysis provided a comprehensive understanding of how students navigate the strengths and limitations of both tools and how their choices are shaped by the nature of the academic task, familiarity with the subject, and perceived cognitive effort.

4 Demographics

4.1 Survey

The survey included 109 participants from a range of academic disciplines and demographic backgrounds. While the majority of participants came from CSE, the sample also included students from EEE, BBA, Data Science, Mathematics, Biochemistry and Biotechnology, and Medicine. Table 1 summarizes the distribution by department, CGPA, gender, and age group.

The survey instrument included Likert-scale questions assessing perceptions of traditional search engines (e.g., Google) and LLM-based tools (e.g., ChatGPT). Participants responded on a 5-point scale: *Never*, *Rarely*, *Occasionally*, *Frequently*, and *Always*. Each set of questions was repeated for both tool categories, covering four core dimensions:

- *How often do you use the following tools for academic tasks?*
- *How satisfied are you with the accuracy of information provided by the following tools?*
- *How efficient are these tools in helping you complete academic tasks?*
- *How easy are these tools to use for academic purposes?*

Participants rated these items separately for both traditional search engines and LLM-based tools. At

Department	
Computer Science and Engineering (CSE)	72
Electrical and Electronics Engineering (EEE)	13
Bachelor of Business Administration (BBA)	10
Data Science	8
Mathematics	4
Biochemistry and Biotechnology	1
Bachelor of Medicine, Bachelor of Surgery (MBBS)	1
CGPA Range	
3.81 – 4.00	23
3.51 – 3.80	30
3.01 – 3.50	36
2.50 – 3.00	16
Below 2.50	4
Gender	
Male	81
Female	28
Age Range	
18 – 20 years	5
21 – 25 years	95
26 – 30 years	9

Table 1: Distribution of participants by department, CGPA range, gender, and age group.

the end of the survey, participants were also asked to indicate their overall preference.

This combination of parallel metrics and comparative judgment allowed for consistent statistical comparisons across tools, while the final preference item offered insight into holistic user inclinations.

4.2 In-person Interview

To complement the survey findings and provide deeper insights into tool-related behaviors, we conducted in-person interviews with 12 students from varied academic backgrounds, primarily from CSE, EEE, and BBA programs. The interview protocol included a structured sequence of six academic tasks, designed to simulate common university-level activities: (1) summarizing a research article, (2) solving a coding problem (for CSE students), (3) answering a circuit-related question (for EEE students), (4) analyzing business data (for BBA students), (5) drafting a formal email, and (6) comparing two popular academic concepts. These tasks were selected based on consultations with domain instructors and a review of typical coursework assignments, ensuring contextual relevance and varying cognitive demands. The goal was to observe how tool choice affected task strategy, accuracy, and efficiency across both discipline-specific and general academic activities.

To assess performance, we developed a task-specific rubric in consultation with faculty members in relevant fields. For example, the coding task was evaluated based on correctness and code clarity; the summarization task was scored on coherence, coverage, and conciseness; and the comparison task was assessed for clarity of distinctions and logical reasoning. Each task was scored independently by two evaluators to ensure inter-rater reliability.

Participants were divided into four groups based on their tool usage patterns during task completion: (1) GPT-only users, (2) Google-only users, (3) tool-balancing users (who used both tools sequentially), and (4) random-choice users (who selected tools freely for each task). This grouping was used to compare differences in accuracy and completion time across task types and to explore how tool-switching behavior aligned with user preferences and task complexity. The interviews also included open-ended reflections on tool usability, trust, and perceived strengths or limitations. These qualitative responses were thematically analyzed to supplement quantitative trends and inform design recommendations.

5 Quantitative Analysis

5.1 Survey Results

Closed-ended survey responses were converted into numerical values for analysis, using a 5-point Likert scale coded as follows: *Never (0)*, *Rarely (1)*, *Occasionally (2)*, *Frequently (3)*, and *Always (4)*. This enabled calculation of means, medians, modes, and standard deviations across four core dimensions: usage frequency, satisfaction, efficiency, and ease of use, for both traditional search engines and LLM-based tools.

The analysis revealed that LLM-based tools were used more frequently and received more favorable ratings across all metrics. The mean usage frequency for traditional search engines was 2.33, with a median of 2.0 and a mode of 2, suggesting occasional use among participants. In contrast, LLM-based tools had a higher mean frequency of 2.79, with a median of 3.0 and a mode of 3, indicating more frequent use. Satisfaction with traditional search engines yielded a mean of 1.99, a median of 2.0, and a mode of 2, reflecting a generally neutral to slightly unsatisfied experience. LLM-based tools, on the other hand, had a higher satisfaction mean of 2.47, with a median and mode of 3.0, indicating moderate satisfaction. Standard deviations for both tools were around 0.9, suggesting consistency in responses.

Efficiency ratings followed a similar trend. Traditional search engines received a mean score of 2.06 (median = 2.0, mode = 2), whereas LLM-based tools were perceived as more efficient, with a mean of 2.65, median of 3.0, and mode of 2. The variability in responses was moderate for both tools, with standard deviations of 0.94 and 1.00, respectively. In terms of ease of use, traditional search engines had a mean score of 2.10, a median of 2.0, and a mode of 2. LLM-based tools again outperformed, with a mean of 2.74, a median of 3.0, and a mode of 3. The standard deviation for LLM ease of use (1.12) was slightly higher, reflecting greater variability in responses. These comparisons are visually presented in Figure 2.

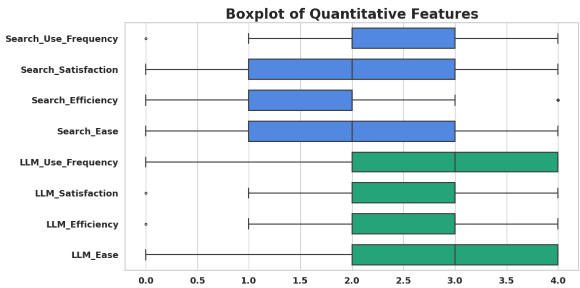


Figure 2: Boxplot of Quantitative Features: This figure presents a comparative analysis of key usability factors between traditional search engines and LLM-based tools. The top four features—*Search_Use_Frequency*, *Search_Satisfaction*, *Search_Efficiency*, and *Search_Ease*—represent user responses related to traditional search engines. The bottom four—*LLM_Use_Frequency*, *LLM_Satisfaction*, *LLM_Efficiency*, and *LLM_Ease*—correspond to user experiences with large language models.

To assess whether the observed differences in user perceptions between traditional search engines and LLM-based tools were statistically significant, we conducted a series of one-way repeated measures ANOVA tests across four dimensions: usage frequency, satisfaction, efficiency, and ease of use. This within-subjects design was appropriate as each participant rated both tools, allowing direct comparison of matched responses. The results revealed significant differences in all cases: usage frequency, $F(1, 108) = 14.82$, $p < 0.001$; satisfaction, $F(1, 108) = 18.95$, $p < 0.001$; efficiency, $F(1, 108) = 21.37$, $p < 0.001$; and ease of use, $F(1, 108) = 17.04$, $p < 0.001$. These findings indicate that participants consistently rated LLM-based tools higher than traditional search engines across all usability dimensions. All F-values were positive, as expected in ANOVA, and each test was independently conducted per variable. Assumptions of normality and sphericity were evaluated and satisfied, supporting the reliability of the results. Overall, the statistical evidence confirms that the differences in ratings are not due to chance but reflect a significant and consistent user preference for LLM-based tools in academic contexts.

To explore whether tool preference was influenced by participant background, a chi-square test was performed to examine associations between tool preference (LLM, search engine, or both) and demographic variables such as age group, gender, and academic department. The chi-squared statistic was χ^2(6, N=109) = 2.012, with a p-value of 0.570. Since the p-value exceeds the 0.05 threshold, we fail to reject the null hypothesis. This indicates that tool preference is not significantly associated with any of the demographic factors analyzed.

In summary, the survey results demonstrate that participants generally prefer LLM-based tools over traditional search engines across all major dimensions of usability. While individual backgrounds, such as department or gender, did not significantly influence this preference, the performance gap between the two tools was consistently supported by both descriptive and inferential statistical analysis.

5.2 In-person Interview

To analyze the data collected from the in-person interviews, we examined both quantitative and qualitative aspects of participant performance while completing a series of structured academic tasks. A total of 12 students participated in this phase of the study. They were assigned six academic tasks representative of common university-level activities. These tasks were selected to reflect a range of cognitive demands, from analytical reasoning to written communication. Participants were grouped based on their tool usage strategy: GPT-only users, Google-only users, tool-balancing users (who used both tools sequentially), and random-choice users (who selected tools freely at each step). Table 2 summarizes the key quantitative findings from your in-person interview.

5.2.1 Accuracy Analysis

Each participant's response was manually evaluated using a predefined scoring rubric tailored to each task type. For instance, the coding task was assessed based on functional correctness and code

Tool Usage Group	Accuracy (%)	Time (min)
GPT-only	83	19
Google-only	78	24
Tool-balancing	90	30
Random-choice	82–88	22–29

Table 2: Summary of task performance by tool usage group. Accuracy (%) refers to the average task score based on a predefined rubric. Time (min) indicates the average completion time across six academic tasks.

readability, while the summarization task was rated on coverage, conciseness, and coherence. The rubric ensured consistency and objectivity across evaluations. Participants who relied exclusively on GPT achieved an average accuracy of 83%, suggesting that LLMs were effective in generating structured responses, particularly for summarization and drafting. In contrast, Google-only users attained an average accuracy of 78%, likely due to the additional effort required to navigate, synthesize, and rephrase content from multiple sources. Participants who employed both tools in a balanced, complementary fashion demonstrated the highest performance, averaging 90% accuracy. Their use of GPT for synthesis and Google for verification allowed for improved reliability and content quality. Among the eight random-choice users, who selected tools freely based on task needs, accuracy ranged from 82% to 88%, depending on the complexity of the task and the appropriateness of tool selection.

5.2.2 Completion Time Analysis

We also recorded the time taken by each participant to complete the assigned tasks. The completion times ranged from a minimum of 13 minutes to a maximum of 42 minutes across all participants. On average, GPT-only users completed tasks the fastest, requiring approximately 19 minutes. This efficiency can be attributed to the conversational nature of LLMs, which reduces the need to browse multiple webpages. Google-only users required more time, around 24 minutes on average, due to the iterative process of selecting, reading, and extracting relevant content from diverse sources. Participants who used both tools took the longest, with an average completion time of 30 minutes. However, this group also achieved the highest accuracy, suggesting a trade-off between speed and performance. The random-choice group showed the most variability in completion time, ranging from 22 to 29 minutes. Their timing appeared to be influenced by both task complexity and personal familiarity with the chosen tools. In general, the results indicate that while GPT-based tools provide speed and ease of access, combining them with traditional search engines can lead to improved accuracy, albeit at the cost of increased task duration.

6 Qualitative Analysis

The qualitative analysis draws on open-ended survey responses and in-person interview transcripts to explore participants' perceptions, experiences, and decision-making strategies when using GPT and Google for academic tasks. We employed a thematic analysis approach to identify recurring patterns and categories in the qualitative data. Initial coding was conducted independently by two researchers who reviewed all textual responses line-by-line. Codes were then grouped into broader themes through iterative comparison and refinement until consensus was reached.

Four major themes emerged from the data: (1) task suitability and tool preference, (2) perceptions of reliability and accuracy, (3) workflow efficiency and cognitive load, and (4) usability and interaction experience.

Task Suitability and Tool Preference. Participants frequently distinguished between tools based on the academic task. GPT was consistently described as effective for quick answers, summarization, and writing support. One participant noted, *"I use ChatGPT whenever I need to summarize something quickly or generate a draft; it saves a lot of time."* In contrast, Google was preferred for tasks requiring deeper exploration and source triangulation. For example, a BBA student shared *"Google helps me see what different sources are saying, especially when I need to analyze business trends from multiple angles."*

Perceptions of Reliability and Accuracy. Trust emerged as a key factor in tool selection. While GPT was appreciated for its fluency and coherence, several participants expressed concerns about outdated or generalized responses. One remarked, *"Sometimes GPT gives an answer that sounds right but isn't actually correct, so I double-check with Google."* Google was consistently rated as more trustworthy for fact-checking and citing sources, though some respondents reported difficulty in assessing source quality or encountering contradictory information.

Workflow Efficiency and Cognitive Load. Many participants described GPT as a way to streamline academic tasks, particularly under time pressure. For instance, a CSE student commented, *"Instead of going through five different websites, I just ask GPT and get a concise answer."* However, this benefit was counterbalanced by reports of multitool use. Students who used both GPT and Google acknowledged that switching between them increased task duration but ultimately improved their understanding and output quality. This dual strategy was especially common for tasks involving coding, data analysis, or structured writing.

Usability and Interaction Experience. GPT was often framed as a conversational assistant or "personal tutor" that guided the student through a problem interactively. In contrast, Google was seen as more traditional but stable. As one student described, *"ChatGPT feels like someone is explaining things to me, but with Google I have to do all the work to find and compare stuff."* Interface familiarity and preferred mode of information delivery influenced tool preference, particularly for students less comfortable with long-form search or unfamiliar domains.

Overall, students perceived GPT and Google not as competing tools but as complementary components of their academic workflow. GPT was favored for its speed, language generation, and summarization abilities, while Google remained essential for verifying facts and consulting credible sources. The choice of tool depended largely on the type of task, the user's prior knowledge, and their need for either convenience or verification. These findings highlight the nuanced, context-dependent strategies students adopt when navigating digital information tools.

7 Discussion

The findings from both the survey and in-person interviews reveal a nuanced interplay between LLMs and traditional search engines in academic information-seeking behavior. GPT-based systems were consistently valued for their ability to provide structured, coherent, and contextually relevant responses. Their strengths were particularly evident in tasks requiring rapid summarization, coding support, or written content generation, where participants appreciated the speed and reduced cognitive effort offered by conversational interfaces. However, while LLMs excelled in usability and perceived efficiency, their limitations, such as occasionally outdated or overly generalized content, prompted students to cross-reference with more authoritative sources.

In contrast, traditional search engines like Google remained the preferred tool for in-depth research, source validation, and academic rigor. Students highlighted Google's extensive access to peer-reviewed literature, academic websites, and multiple viewpoints as vital for tasks requiring critical evaluation or citation. Nonetheless, participants also reported experiencing information overload and inefficiencies due to the need to manually sift through links, assess credibility, and synthesize fragmented content. These trade-offs suggest that tool preference is not static but shaped by the academic task's complexity, time constraints, and the student's familiarity with the subject matter.

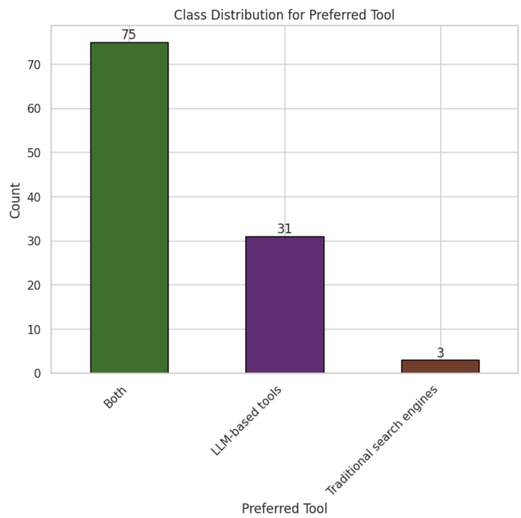


Figure 3: Class Distribution of the Preferred Tool among the Students

A recurring theme across both quantitative and qualitative data was the strong interest in a hybrid model that seamlessly integrates the complementary strengths of GPT and Google. As illustrated in Figure 3, a significant portion of participants expressed a desire for an academic support tool that combines GPT's conversational and summarization capabilities with Google's multi-source, real-time information retrieval. Such a system would enable users to receive concise responses with embedded citations and links to original sources, streamlining the verification process without sacrificing depth or credibility. Participants viewed this hybrid approach as a way to reduce cognitive load, eliminate repetitive tool-switching, and enhance learning outcomes through more fluid academic workflows.

While LLMs and traditional search engines serve distinct purposes, students view them as complementary rather than competing tools. The integration of their respective advantages, LLMs for generation and synthesis, and search engines for depth and verification, represents a promising direction for the future of academic information retrieval. These findings underscore the importance of designing intelligent, context-aware tools that adapt to students' diverse needs while upholding standards of reliability and academic integrity.

These findings also raise a key question: ***What happens when students use GPT and Google together: Does it help or hurt?*** The evidence from our study suggests a compelling answer. Participants who used both tools, using GPT to quickly summarize and clarify complex topics, and Google to verify facts and consult authoritative sources, consistently outperformed those who relied on either tool alone. These hybrid users achieved the highest task accuracy (90%), demonstrating that the strategic integration of LLMs and traditional search engines not only complements their respective strengths but also minimizes their individual weaknesses. Although this dual-tool approach required more time, students perceived it as a worthwhile trade-off for greater confidence, deeper understanding, and higher-quality outcomes. This finding underscores the potential of thoughtfully designed hybrid systems to support academic workflows, reducing cognitive load while maintaining rigor and trustworthiness in the learning process.

8 Proposed Prototype

Drawing on user insights from both survey responses and interviews, we propose a conceptual prototype that integrates GPT-based assistance directly into the traditional search engine interface.[1] The goal is to address the cognitive and logistical burden of switching between tools by creating a unified platform that combines the conversational utility of LLMs with the source-rich infrastructure of search engines. The prototype is designed as an embedded chatbot, positioned unobtrusively in the corner of the search interface, allowing users to engage in interactive, context-aware dialogue without disrupting their familiar browsing workflow.

Unlike standalone LLM interfaces, the proposed assistant does not replace standard search results. Instead, it complements them by offering real-time summaries, follow-up clarifications, and cross-source syntheses derived from retrieved documents. For example, when a user performs a Google search, the assistant can instantly summarize key points from the top-ranked results, offer bullet-point comparisons across sources, or simplify complex academic texts. Users can also ask follow-up questions to refine or extend the information, eliminating the need to manually revisit and interpret multiple pages.

The prototype's novelty lies in its hybrid architecture that allows toggling between raw search content and AI-enhanced interpretation. Crucially, each AI-generated insight is accompanied by links to the original source, promoting transparency and reducing the risk of hallucinated or unverifiable responses. This feature directly addresses the concern, voiced by multiple participants, regarding LLM trustworthiness in academic work.

Although conceptual in its current form, the prototype was informed by empirical findings from this study and inspired by real-world user preferences. Its contribution lies in reimagining academic information retrieval as an interactive and adaptive process. Over time, the assistant could learn user preferences, discipline-specific language, and search habits to deliver more relevant and personalized guidance. By embedding this intelligent layer into the search experience, the prototype aims to reduce cognitive load, increase search efficiency, and promote evidence-based academic practices, ultimately bridging the current gap between generation and verification in digital research tools.

9 Conclusion

Our study demonstrates that university students adopt a complementary approach to academic information retrieval, using LLMs for quick explanations and drafting, and traditional search engines for verification and accessing credible sources. Our mixed-methods findings underscore the task-dependent nature of tool preference and reveal strong interest in a hybrid model. While the proposed prototype remains conceptual and the interview sample was limited, the results offer practical insights for designing AI-assisted academic tools. Future work will focus on expanding participant diversity, validating qualitative themes, and implementing a functional prototype to assess real-world usability and impact.

[1] https://shorturl.at/t6Tf8

Acknowledgments

We sincerely thank everyone who participated in the survey and interviews for their time and experience sharing.

Limitations

This study provides important insights into students' use of LLMs and traditional search engines for academic tasks; however, several limitations should be acknowledged. The in-person interview sample was relatively small ($n = 12$) and primarily drawn from technology-related disciplines, which may limit the generalizability of the findings. There is also potential sampling bias, as the majority of participants were Computer Science and Engineering (CSE) majors from a single university context, which may not reflect the experiences or preferences of students from other academic backgrounds or institutions. While the qualitative analysis surfaced meaningful themes, it lacked intercoder reliability checks, and the survey relied on self-reported data that may be affected by recall or social desirability bias. Moreover, the proposed prototype remains at a conceptual stage and has not yet been implemented or tested in real academic environments, leaving its practical impact unverified. Lastly, the study did not account for factors such as digital literacy, prior experience with AI tools, or task complexity, all of which could influence tool preferences and performance.

Ethical Considerations

All procedures involving human participants in this study were conducted in accordance with ethical research standards. Participation in both the survey and in-person interviews was voluntary, and informed consent was obtained from all participants prior to data collection. Respondents were assured of anonymity and confidentiality, and no personally identifiable information was collected or stored. The data were used solely for research purposes and analyzed in aggregate to protect individual identities. As the study did not involve vulnerable populations, clinical interventions, or sensitive topics, risk to participants was minimal. The conceptual prototype proposed in this study does not process real user data and poses no immediate privacy concerns. Future implementation of the prototype will incorporate robust data protection, user consent mechanisms, and institutional ethical review as necessary.

References

U Alberth. 2023. The use of chatgpt in academic writing: A blessing or a curse in disguise? *TEFLIN Journal-a Publication on the Teaching and Learning of English*, 34:337–52.

Satinder Bal and Rajender Nath. 2009. A comparative study of traditional search engines with the metasearch engines. *Ultra scientist*, 21(2M):597–610.

A Bansal. 2023. Optimizing rag with hybrid search and contextual chunking. *Journal of Emerging Applications in Science and Technology*, 5.

Kevin Matthe Caramancion. 2024. Large language models vs. search engines: Evaluating user preferences across varied information retrieval scenarios. *arXiv preprint arXiv:2401.05761*.

Rahul R Divekar, Sophia Guerra, Lisette Gonzalez, and Natasha Boos. 2024. Choosing between an llm versus search for learning: A highered student perspective. *arXiv preprint arXiv:2409.13051*.

Prakriti Kapoor, Shweta Mahida, Sharon John, and 1 others. 2024. Ai-driven adaptive systems for personalized library research assistance. *Library of Progress-Library Science, Information Technology & Computer*, 44(3).

Harsh Kumar, Mohi Reza, Jeb Mitchell, Ilya Musabirov, Lisa Zhang, and Michael Liut. 2024. Understanding help-seeking behavior of students using llms vs. web search for writing sql queries. *arXiv preprint arXiv:2408.08401*.

Perplexity AI. 2024. Perplexity ai: Conversational search engine. Accessed: August 11, 2025.

Farman Ali Pirzado, Awais Ahmed, Román A Mendoza-Urdiales, and Hugo Terashima-Marin. 2024. Navigating the pitfalls: Analyzing the behavior of llms as a coding assistant for computer science students-a systematic review of the literature. *IEEE Access*.

Tam Sakirin and Rachid Ben Said. 2023. User preferences for chatgpt-powered conversational interfaces versus traditional methods. *Mesopotamian Journal of Computer Science*, 2023:22–28.

Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744*.

Albatool Wazzan, Stephen MacNeil, and Richard Souvenir. 2024. Comparing traditional and llm-based search for image geolocation. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pages 291–302.

Ruiyun Xu, Yue Feng, and Hailiang Chen. 2023. Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135*.

A Survey Instrument

Survey Questionnaire (Selected Items)

Tool Usage Frequency (Likert scale: Never to Always)

- How often do you use Google for academic tasks?
- How often do you use GPT-based tools for academic tasks?

Perceived Satisfaction, Efficiency, Ease of Use (Likert scale)

- How satisfied are you with the accuracy of results from each tool?
- How efficient are these tools in completing academic tasks?
- How easy are these tools to use?

Tool Preference

- Which tool do you prefer overall: Google, GPT, or Both?

Open-Ended

- In what scenarios do you prefer GPT over Google or vice versa?
- What limitations have you faced when using these tools?

B Interview Tasks

Assigned Academic Tasks

Participants were given six structured academic tasks designed to simulate realistic coursework challenges across different disciplines:

1. **Summarize a Research Abstract (All Participants)**
   Read a 250-word abstract from a peer-reviewed article and produce a concise 3–5 sentence summary capturing the main objective, methods, and findings.

2. **Solve a Coding Problem (CSE Only)**
   Write a Python function to compute the factorial of a number, ensuring proper input validation and code documentation.

3. **Analyze a Circuit Diagram (EEE Only)**
   Interpret a simple resistive circuit with three resistors and a voltage source. Calculate total resistance and current using Ohm's Law.

4. **Interpret a Business Chart (BBA Only)**
   Given a bar chart showing quarterly revenue for three products, provide a 5–6 sentence interpretation of trends, anomalies, and business implications.

5. **Draft a Formal Email (All Participants)**
   Write a professional email to your course instructor requesting an extension on an assignment. The email should be polite, concise, and persuasive.

6. **Compare Two Academic Concepts (All Participants)**
   Write a short paragraph comparing "quantitative" vs. "qualitative" research methods, highlighting key differences and use cases.

Participants were grouped by tool usage pattern: GPT-only, Google-only, tool-balancing (both sequentially), and random-choice (free selection per task).

C Task Evaluation Rubric

Rubric for Evaluating Task Accuracy (0–10 Scale)

Each academic task was scored on a scale from 0 (poor) to 10 (excellent), based on specific content and skill-based criteria. Rubrics were standardized across evaluators to ensure consistency.

1. Summarization Task

- **Coverage of Key Ideas (0–4):** Accurately identifies main purpose, methods, and findings.
- **Conciseness and Clarity (0–3):** Avoids redundancy; sentences are readable and logically ordered.
- **Language Accuracy (0–3):** Grammar, punctuation, and vocabulary are appropriate for academic tone.

2. Coding Task (CSE Only)

- **Correctness (0–4):** Produces correct output for sample inputs.
- **Code Quality (0–3):** Structured, readable, and modular.
- **Input Handling and Comments (0–3):** Includes input validation and descriptive inline comments.

3. Circuit Analysis Task (EEE Only)

- **Correct Calculation (0–5):** Accurate application of formulas (e.g., Ohm's Law).
- **Interpretation and Units (0–3):** Correct labeling and use of units.
- **Clarity of Steps (0–2):** Clear logical progression of calculations.

4. Business Chart Interpretation (BBA Only)

- **Insightfulness (0–4):** Accurately identifies trends, anomalies, and patterns.
- **Relevance (0–3):** Comments relate meaningfully to business implications.
- **Clarity (0–3):** Well-structured explanation with clear language.

5. Formal Email Draft

- **Professional Tone and Structure (0–4):** Proper salutation, closing, and paragraphing.
- **Persuasiveness (0–3):** Presents a clear and reasonable justification.
- **Grammar and Clarity (0–3):** Language is appropriate, polite, and error-free.

6. Concept Comparison

- **Content Accuracy (0–4):** Identifies valid, discipline-appropriate distinctions.
- **Comparative Logic (0–3):** Clearly outlines similarities/differences.
- **Language and Coherence (0–3):** Academic tone and logical flow.

D Thematic Codebook

Thematic analysis of open-ended survey responses and interview transcripts resulted in four overarching themes. Each theme is described in Table 3, along with its associated codes and representative participant quotes.

Parent Theme	Child Code	Definition / Quote Example
Effectiveness	Task Fit	Tool suitability for academic tasks, which refers to how well a tool matches the academic task at hand. *"GPT is great for summaries, but not so much for detailed citations."*
Usability & Cognitive Load	Ease of Use	Simplicity of interaction with the tool, which means how intuitive and straightforward users find the tool. *"ChatGPT saves me time by avoiding extra clicks."*
	Information Overload	Frustration with excessive irrelevant results, which describes frustration due to excessive, often irrelevant, search results. *"Google gives too many links and I get lost trying to pick one."*
Trust and Credibility	Source Verification	Need for citable sources, which refers to the extent to which students cross-check the tool output with credible sources. *"I trust Google more when I need something fact-checked."*
Contextual Tool Selection	Task Type Influence	Decision to use a tool depends on the academic context, which describes tool choice based on the academic context or subject matter. *"For programming help, I use GPT; for research papers, I go with Google."*

Table 3: Thematic mapping of child codes derived from open-ended survey responses and interviews. Quotes show typical user sentiment for each theme.

First Impressions from Comparing Form-Based and Conversational Interfaces for Public Service Access in India

Chaitra C R, Pranathi Voora, Bhaskar Ruthvik Bikkina*, Bharghavaram Boddapati, Srivant Vishnuvajjala, Vivan Jain, Prajna Upadhyay, Dipanjan Chakraborty

BITS Pilani, Hyderabad Campus

{p20210024,f20222012}@hyderabad.bits-pilani.ac.in, bhaskar-ruthvik@tamu.edu*

{f20230696,f20220113,f20220140,prajna.u,dipanjan}@hyderabad.bits-pilani.ac.in

Abstract

Accessing government welfare schemes in India remains difficult for emergent users: individuals with limited literacy, digital familiarity, or language support. This paper compares two mobile platforms that deliver the same scheme-related information but differ in interaction modality: myScheme, a government-built, form-based Android application, and Prabodhini, a voice-based conversational prototype powered by generative AI and Retrieval-Augmented Generation (RAG).

Through a task-based comparative study with 15 low-income participants, we examine usability, task completion time, and user preference. Drawing on theories such as the Gulf of Execution and Zipf's Law of Least Effort, we show that Prabodhini's conversational design and support for natural language input better align with emergent users' mental models and practices. Our findings highlight the value of multimodal, voice-first NLP systems for improving trust, access, and inclusion in public digital services. We discuss implications for designing accessible language technologies for marginalised populations.

1 Introduction

India's central and state governments have long adopted a welfare-oriented approach to governance, offering numerous social protection schemes to support the elderly, low-income, and marginalised populations. These schemes also target workers in the unorganised sector, which comprises approximately 92% of the country's workforce (Sakthivel and Joddar, 2006). While well-intentioned and potentially transformative, the actual uptake and utilisation of these services remains low (Rahman and Pingali, 2024).

Several barriers hinder effective access to welfare schemes. Although e-governance platforms have made these services digitally accessible, emergent users, individuals with limited digital experience, often facing low literacy, low income, and poor infrastructure access (Thies et al., 2015) struggle to engage meaningfully with them. Key obstacles include a lack of awareness about available schemes, difficulties in navigating complex form-based interfaces, and associated costs of access, such as relying on cyber cafés to fill out forms for nominally "free" services (Chakraborty et al., 2017).

These barriers reflect broader mismatches between the expectations embedded in digital interfaces and the lived realities of emergent users. Norman's concept of the *Gulf of Execution* (Norman, 1986) offers a useful lens here: systems like myScheme require users to translate their needs into the language and structure of the interface, rather than allowing users to express their goals in familiar terms. Furthermore, these systems often violate *Zipf's Law of Least Effort* (Zipf, 2016), which suggests that users prefer interaction paths that demand the least cognitive and physical effort. By relying heavily on hierarchical forms, structured data fields, and pre-defined filters, current platforms place the burden of adaptation on the user, thus worsening exclusion.

This paper explores whether interaction modality—specifically, traditional form-based interfaces versus conversational, voice-based ones—affects the usability and accessibility of mobile information systems for emergent users. We address the research question: *How do different mobile interaction modalities—namely, form-based graphical interfaces versus voice-based conversational systems—affect emergent users' ability to seek and access information about government welfare schemes?*

To investigate this, we conducted a comparative user study with 15 participants drawn from low-income, blue-collar workers employed at a

*Work done as an undergraduate student at BITS Pilani, Hyderabad Campus.

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 60–71
November 9, 2025

university campus in South India. The two platforms we evaluated deliver identical government welfare information but differ in design philosophy and interaction modality. The first is **myScheme**, a government-built Android application that relies on hierarchical menus and form-filling interfaces. The second is **Prabodhini** (Sanskrit for *awakening*) (Figure 1), a prototype conversational app developed in our lab. Prabodhini uses a backend powered by GPT-4 and Retrieval-Augmented Generation (RAG) applied to data sourced from the official myScheme website. It is designed through a user-centred process and includes voice-input capabilities in regional Indian languages using off-the-shelf text-to-speech (TTS) and speech-to-text (STT) engines. The technical details of Prabodhini are available in (Jain et al., 2025). We compared Prabodhini with the **myScheme** application, as it is the only government-released platform of its kind, and the information it provides is considered authoritative and valid.

Unlike myScheme, which presents users with dense static text and long application forms, Prabodhini breaks down information into small, actionable conversational nuggets, enabling users to query the system using natural language—either spoken or typed. This design not only reduces the *Gulf of Execution* but also aligns with the mental models and digital practices of mobile-first emergent users. Many participants are already accustomed to voice interactions through tools like Google Search, and conversational systems like Prabodhini leverage these affordances to improve accessibility and confidence.

Our findings show that users preferred Prabodhini over the form-based alternative. Conversational, voice-first systems improved access, inclusion, and trust for users often excluded from digital services. This study bridges HCI and NLP by applying a large language model (GPT-4) to reduce usability barriers for low-literate, mobile-first users in India. Prabodhini uses a Retrieval-Augmented Generation (RAG) pipeline to deliver scheme information through natural language queries. By combining speech input and scenario-based design, it makes government services more accessible in low-resource settings. This work brings NLP research closer to real-world, socially relevant HCI challenges.

2 Related Work

2.1 Tools and Interfaces for Emergent Users

Emergent users—those with limited literacy or digital exposure—require contextually adapted, inclusive interfaces. Human Computer Interaction for Development (HCI4D) and Information and Communication Technology for Development (ICTD) research has emphasised designing for this population to prevent trickle-down marginalisation (Jones et al., 2017). Prior work spans multiple domains: banking (Melo et al., 2023; Mohammed et al., 2023), education (Ngoon et al., 2024), shopping (Mohammed et al., 2023), health (Reen et al., 2024), and government services (Mehtälä and Nieminen, 2019). These studies stress usability for low-literate users, recommending culturally grounded design (Medhi et al., 2006).

In the e-governance context, Mehtala et al. (Mehtälä and Nieminen, 2019) and Srivastava et al. (Srivastava et al., 2021) highlight the importance of participatory and user-centred approaches. Our work builds directly on these insights by evaluating a government welfare app and introducing a voice-first conversational alternative designed for emergent users.

2.2 Information Seeking by Emergent Users

Theories like Zipf's Law of Least Effort (Zipf, 2016) and Dervin's Sense-Making Theory (Dervin and Naumer, 2009) stress that users prefer minimal effort and context-sensitive systems. For marginalised groups, Chatman (1991), Dhaygude and Chakraborty (2020) and Aribandi et al. (2022) show that trust, familiarity, and sociocultural norms shape engagement. Emergent users tend to favour human sources or simplified interfaces (Robinson, 2010).

Technologies like Interactive Voice Response System (IVRS) (Joshi et al., 2014; Kazakos et al., 2016; Patel et al., 2009; Srinivasan et al., 2013; Chakraborty and Seth, 2015; Chakraborty et al., 2017), icon-based UIs (Medhi et al., 2011), and Android apps (Cuendet et al., 2013; Chandel and Doke, 2013; Shah and Sengupta, 2018) have been developed to address these needs. Conversational agents (CAs) offer another promising modality (Prasad et al., 2019; Vaccaro et al., 2018; Jain et al., 2018; Purington et al., 2017).

Kodagoda et al. (Kodagoda et al., 2009) and Malthouse et al. (Malthouse, 2023) observe that emergent users often accept the first satisfactory

(a) Screen 1 (b) Screen 2 (c) Screen 3

Figure 1: Prabodhini's interface. Screen 1 prompts users to select their preferred language. In Screen 2, users are required to choose their state since several schemes are state-specific. Screen 3 allows users to either type or orally input their queries.

result or abandon searches if unsuccessful. Our system, Prabodhini, addresses these behavioural tendencies by enabling open-ended voice queries and delivering concise, localised responses.

2.3 Conversational vs. Static Information Systems

While traditional GUIs rely on structured navigation and text input, they assume a level of literacy that emergent users may not possess (Følstad and Brandtzæg, 2017; Budiu, 2018). Conversational systems mitigate this by supporting multimodal input and dynamic dialogue (Zhang et al., 2018). Studies comparing conversational and static systems show improved usability, satisfaction, and efficiency with dialogue-based interfaces (Balloccu and Reiter, 2022; Kaushik and Jones, 2023; Roy, 2024).

In particular, chatbots have shown benefits in domains requiring explanation and guidance, such as diet tracking (Balloccu and Reiter, 2022) or search tasks (Kaushik and Jones, 2023). Wagner et al. (Wagner, 2004) advocate for conversational knowledge management to enhance accessibility.

Our study contributes to this literature by comparing a government-built form-based app (MyScheme) with a generative AI-powered, voice-first system (Prabodhini). We show how conversational design, regional language support, and scenario-driven interaction reduce the barriers that static interfaces impose on emergent users.

2.4 Recent Advances in HCI+NLP

Recently, the intersection of Human-Computer Interaction (HCI) and Natural Language Processing (NLP) has received growing attention, particularly in domains such as civic participation, healthcare, education, and accessibility. Heuer and Buschek (2021) presents five methodological proposals that bridge HCI and NLP, positioning them within the context of machine learning–based NLP systems and their implications for user experience design. Complementing this, Sultana et al. (2022) examines challenges associated with popular NLP dataset types, framing their critique through the lens of narrative-based methods commonly used in HCI. Their work highlights opportunities for NLP techniques to enrich qualitative narrative analysis and inform the development of more inclusive, user-centred datasets.

Building on this foundation, Guridi et al. (2025) emphasises that the adoption of NLP tools within government settings is not merely a matter of technical performance but is heavily influenced by internal stakeholder incentives and the need to demonstrate political legitimacy. In response to these insights, we present a human-centered voice-first prototype designed to improve access to legal and policy information for the emergent users.

3 Methodology

We conducted a comparative user study to examine how interaction modality influences the usability of mobile welfare apps for emergent users. We evaluated two Android applications: the government-developed *myScheme*, which uses form-based, text-heavy interfaces, and *Prabodhini*, a voice-based conversational app developed in our lab. Both apps used the same backend data, allowing a controlled comparison of interaction design.

Participants engaged in structured tasks on both apps. We measured task completion time, recorded observational notes, and administered a usability questionnaire. A doctoral researcher and two undergraduate students facilitated the sessions, which lasted approximately 30 minutes each. All interactions were conducted in Telugu or Hindi.

A pilot with two participants revealed difficulties with the standard System Usability Scale (SUS) due to low literacy and unfamiliarity with Likert scales. We therefore created a simplified binary-response questionnaire, inspired by SUS constructs but adapted to suit the participant group (see Table 3).

3.1 Participants

We recruited 15 participants (6 women, 9 men) using convenience sampling. All were low-income, low-literate workers employed on a university campus through service-outsourcing agencies, with basic familiarity with smartphones. Their educational backgrounds ranged from Class 5 to Class 12. Twelve participants spoke Telugu, and three spoke Hindi. Monthly incomes ranged from INR 13,000 to INR 15,000 (approx. USD 157–181). Oral consent was obtained from all participants. Table 1 summarises the demographic details of the participants.

3.2 Prabodhini

Prabodhini is a light-weight, mobile-friendly platform that employs chain-of-thought prompting over GPT-4o, layered on top of RAG, to generate context-aware and personalised responses. The features of the applications are derived from the findings of our prior work (Chaitra et al., 2025). It also introduces Actionable Information Retrieval (AIR), where user queries are categorised into procedural, yes/no, or informative types, enabling step-by-step voice-guided interactions instead of dense text. A lightweight design, supported by a hybrid retrieval pipeline and demographic personalisation, ensures accessibility for low-text-literate users (Chaitra et al., 2025). This design emphasises voice-first interaction, progressively leading users to precise answers while reducing reliance on text literacy.

3.3 Procedure

Each participant used both applications on the same Android phone connected to the same mobile data network. After a brief tutorial, participants completed the tasks independently. App order was randomised, though slightly imbalanced (nine used myScheme first, six used Prabodhini first). However, six participants each who used either Prabodhini or myScheme first, completed Task 1. We did not disclose which app was developed by the researchers.

After completing the tasks, we logged the participants' responses to the usability questionnaire, and the participants engaged in a brief semi-structured interview. We manually recorded observations of user behaviour and interface challenges. Interviews were audio-recorded with consent.

3.4 Task Design

Tasks reflected common actions for accessing government schemes:
1) Find a relevant scheme.
2) Check eligibility criteria.
3) Understand the application process.

All participants completed Task 1. Only two proceeded to Tasks 2 or 3, citing language barriers in myScheme or being confident about being able to use Prabodhini later. Given this, our analysis focuses on Task 1 as a representative entry-point task for evaluating usability. Interaction challenges and support needs were recorded throughout. Network-induced delays were excluded from task timing. Findings are presented in Section 4.2.

3.5 Ethical Considerations And Positionality

In the absence of a formal ethics board in our university, we followed ethical self-regulation guidelines from Dearden et al. (Dearden and Kleine, 2018). Participation was voluntary and anonymous. The researchers are trained in Human-Computer Interaction and computer science, with prior experience designing technologies for underserved communities in India. We approached the study with a commitment to participatory, respectful engagement. Local languages were used throughout the study to minimise power imbalances and foster trust.

4 Findings

We conducted both qualitative and quantitative analyses of the data obtained through the study.

Participants	Gender	Occupation	Age	Qualification	First Language
P1	F	House Keeping Staff	32	Class 5	Telugu
P2	F	House Keeping Staff	32	no formal education	Telugu
P3	F	House Keeping Staff	35	Class 10	Telugu
P4	M	Student Hostel Attendant	26	Class 10	Telugu
P5	F	House Keeping Staff	36	no formal education	Telugu
P6	F	House Keeping Staff	28	Class 10	Telugu
P7	F	House Keeping Staff	38	no formal education	Telugu
P8	M	Office boy	25	Diploma	Telugu
P9	M	Hostel Attendant	42-46 (not sure)	Classs 9	Hindi
P10	M	Security Guard	49	Class 9	Telugu
P11	M	Security Guard	42	Class 12	Telugu
P12	M	Security Guard	53	Class 8	Telugu
P13	M	Security Guard	35	Class 10	Telugu
P14	M	Security Guard	36	Class 10	Hindi
P15	M	Security Guard	25	Class 12	Hindi

Table 1: Demographic details of the participants

Qualitative data were logged through observations while the participants interacted with the apps and any feedback the participants provided after the tasks. The quantitative data collected is the task completion time for the tasks defined in Section 3.4 and response to the usability questionnaire (Table 3).

4.1 Qualitative Findings

In this section, we report findings from the qualitative data collected during the study. We undertook a thematic analysis (Clarke and Braun, 2017) of the observation logs and participant feedback, and the identified themes were arranged into the following subsections.

4.1.1 Language Barriers and Localisation

myScheme is available only in English and Hindi, which posed a significant barrier for participants who were more comfortable in other regional Indian languages. Several users struggled to navigate the app due to unfamiliar terminology and the absence of language options tailored to their needs. This challenge was particularly acute for those with limited literacy or no formal exposure to English. Participants expressed frustration when faced with an interface that they could not comprehend. As one participant remarked when weighing the pros and cons of myScheme: **“We do not want an application that is in English” (P10)**.

In contrast, Prabodhini let users choose their preferred language during setup (Figure 1a). At the time of this research, it supported English, Hindi, Tamil, and Telugu. This enabled all participants—Hindi or Telugu speakers—to use the app in a familiar language, reducing cognitive and linguistic barriers.

4.1.2 Lack of Discoverability and Mental Models in the myScheme App

The design of myScheme overlooks the mental models and information-seeking habits of its intended users. Its features mirror web interfaces for educated, digitally literate audiences, influenced by Western usability norms. For example, the app’s search function assumes users know scheme names, but none of our participants used it. Lacking prior knowledge of schemes or eligibility, they couldn’t initiate keyword searches, making the feature effectively unusable for this group.

In contrast, Prabodhini allows users to express their needs in natural language via voice or text input. The system returns relevant schemes based on the scenario. This approach aligns more closely with the mental models of emergent users, who typically frame their queries in terms of personal circumstances rather than formal scheme names. We drew upon findings from a prior study (Chaitra et al., 2025), where researchers had documented this preference for scenario-driven interaction, and incorporated those insights into the design of Prabodhini.

4.1.3 Mismatch Between User Capabilities and App Requirements

Another key challenge participants faced when using myScheme stemmed from the mismatch between their capabilities and the design expectations embedded in the app. To receive personalised scheme recommendations, users were required to fill in a form that captured personal and demographic details. This process introduced several barriers:

1. **Time-Consuming and Tedious:** Participants found the form-filling process laborious and often needed assistance to proceed, especially when selecting from dropdown menus or entering structured information.

2. **Unfamiliar Terminology:** Several form fields used jargon or abstract categories that did not resonate with participants' lived experience. For example:

   - *BPL Status:* Users were asked to indicate whether they belonged to the Below Poverty Line (BPL) category. Most participants were either unaware of their status or confused by the question, as definitions of BPL vary across states and are rarely part of everyday discourse.
   - *Occupation Classification:* Users had to select from predefined categories, many of which used technical language such as "organised" or "unorganised" sector. These terms lacked salience for participants, who struggled to map their own work (e.g., housekeeping or security work) onto the listed options.
   - *Urban/Rural Classification:* The form asked whether users lived in an "urban" or "rural" area. Participants found this terminology abstract and suggested simpler alternatives like "city" or "village", which aligned better with their vocabulary and everyday references.

Ultimately, only two participants managed to complete the form independently. Even then, the resulting scheme suggestions were often irrelevant or inapplicable to their state of residence. Most users required repeated assistance and expressed frustration with the form's complexity. These findings echo prior research on information accessibility barriers in public digital systems (Ahmed et al., 2013).

Prabodhini addressed this gap by allowing users to pose open-ended queries in natural language. This interaction style eliminated the need for categorical precision and reduced the cognitive burden on users. For example, one participant asked: **"I am from Ponnala village. I want to open a stationery shop. Tell me which schemes can I avail?" (P13)**. The system responded with a curated list of relevant schemes, including eligibility and application details, based on the described scenario, without requiring the user to translate their needs into formal classifications.

4.1.4 Reliability and Trust in Conversational Interfaces

The myScheme application includes a chatbot intended to assist users in locating relevant information through natural language queries. While this feature holds potential for simplifying access, participants reported frequent issues with its responsiveness. In multiple instances, the chatbot failed to return results or became unresponsive mid-query, leading users to abandon the attempt or try again later. Such inconsistencies not only disrupted the flow of interaction but also diminished users' trust in the system's reliability.

For emergent users—who may already be cautious or uncertain when engaging with digital services—technical failures can reinforce negative perceptions and discourage future use. Prior studies have highlighted how unreliable interfaces reduce user confidence and erode trust in public digital platforms (Asogwa, 2013; Verdegem and Verleye, 2009).

In contrast, Prabodhini handled user queries without noticeable lag or disruption during our study sessions. Its backend processed requests reliably, whether entered via speech or text, allowing participants to explore information without the frustration of broken interactions. This consistency emerged as a key factor contributing to participants' preference for Prabodhini over myScheme.

4.1.5 Perceived Value and Challenges of Voice Input in Prabodhini

Participants widely appreciated the voice input functionality in Prabodhini, which allowed them to articulate queries orally in their native language. Many users found this mode of interaction intuitive and aligned with their prior experience using voice features in mainstream apps. One participant described the interface as familiar: **"It is like in Google" (P5)**, referring to their familiarity with using voice input in native languages on the Google search interface. This perceived similarity enhanced their confidence and willingness to explore the app, especially among users who found typing in local languages difficult or unfamiliar.

The availability of voice input in regional languages, specifically Telugu and Hindi, further contributed to the system's accessibility. Several participants noted that they often use voice features when searching on YouTube or sending voice notes on messaging platforms. Prabodhini's interface leveraged this familiarity to reduce friction during task completion.

By contrast, myScheme did not offer a voice input option, which many users identified as a limitation. The absence of multimodal input made it more difficult to navigate the app, particularly for those who were hesitant to type or read lengthy text in non-native languages.

That said, some participants encountered usability issues with the voice feature in Prabodhini. Specifically, users were occasionally unsure whether the app was actively listening, due to the lack of clear feedback cues in the interface. These issues were attributed to minor bugs and inconsistencies in how the listening state was communicated. While they did not prevent task completion, these glitches highlight the need for improvements in real-time feedback design and system responsiveness.

4.1.6 Challenges with Speech Output in Native Languages

Some participants noted issues with the quality of Prabodhini's text-to-speech (TTS) responses when interacting in their native language. These problems became more pronounced in low-connectivity environments, where the app defaulted to an offline TTS engine lacking Indian accents or natural prosody. As a result, users found certain responses difficult to understand. This is noted by the work conducted by Jiao et al. (2024) as well.

Despite these limitations, participants appreciated the app's provision of a text transcript alongside the spoken output. This feature allowed users to read the response if they had trouble understanding the audio, thereby preserving a degree of independence and continuity in the interaction. While the clarity of voice responses remains an area for improvement, the availability of multi-modal feedback helped mitigate the impact of occasional poor audio rendering.

4.2 Quantitative Findings

We measured task completion time for each application based on participants' performance in Task 1, the only task completed by all 15 users. Timing was recorded from the moment participants began interacting with the app until a list of welfare schemes was returned. We excluded delays caused by data fetching, as these depended on mobile network conditions rather than interface design.

Table 2 shows the task completion time for each participant. On average, participants completed Task 1 in 49 seconds (95% CI 25.74 – 72.26) using Prabodhini. All 15 participants successfully finished the task. In contrast, the average completion time on myScheme was 118 seconds (95% CI 81.1 – 154.9), based on data from nine participants. The remaining six could not use myScheme due to its lack of Telugu language support.

To control for ordering effects, we compared task times based on which platform was used first. When Prabodhini was used first, the average time was 50 seconds (95% CI 5.92 – 94.08); for myScheme, it rose to 127 seconds (95% CI 72.43 – 181.57). Of nine myScheme users, only two completed the task unaided—one via the Hindi interface—while others struggled with complex terms, poor navigation, and unfamiliar forms.

For Prabodhini, occasional delays were linked to issues with the voice input feature (see Section 4.1.5), particularly when the app failed to clearly indicate whether it was listening. Despite this, users were generally able to complete tasks without assistance.

In addition to task timing, we administered a six-item usability questionnaire adapted from the SUS framework. Participants selected their preferred app for each item. Table 3 presents the distribution of responses. The results reveal a clear preference for Prabodhini. Participants rated it more positively across all dimensions, including ease of use, confidence, and perceived complexity. By contrast, myScheme was often described as cumbersome and difficult to use independently. These findings suggest that Prabodhini's voice-first, conversational design better supports the needs and expectations of emergent users.

5 Discussion and Conclusion

Our findings highlight the considerable challenges faced by emergent users when engaging with digital services that rely on form-based interfaces, technical jargon, or limited language options. Participants in our study struggled with myScheme's rigid form structure, abstract categories (e.g., "ur-

Participants	First Platform Used	mySchemeTask Completion Time	Prabodhini Task Completion Time	Comments
P1	myScheme	-	0:00:03	-
P2	Prabodhini	-	0:01:32	User faced issue with microphone usage
P3	myScheme	0:02:05	0:00:05	-
P4	Prabodhini	0:02:08	0:00:34	Received help to fill the form in myScheme app
P5	myScheme	-	0:01:36	User faced issue with microphone usage
P6	myScheme	0:01:01	0:00:04	User needed extra time to locate the scheme in the myScheme app
P7	Prabodhini	-	0:00:04	-
P8	myScheme	0:01:28	0:00:28	User needed extra time to locate the scheme in the myScheme app
P9	myScheme	0:03:17	0:01:02	User interacted with the myScheme app in Hindi language
P10	Prabodhini	0:02:01	0:01:33	User faced issue with microphone usage
P11	myScheme	0:02:56	0:02:11	User faced issue with microphoneusage
P12	Prabodhini	-	0:01:15	Received help in phrasing the question
P13	myScheme	0:01:53	0:00:37	-
P14	Prabodhini	0:00:51	0:00:03	-
P15	myScheme	-	0:01:01	User faced issue with microphone usage

Table 2: Task completion time across platforms, along with issues reported by participants. Time is denoted in the format hh:mm:ss. '-' in the Time columns indicate the participant was not able to perform the task on the platform. The Comments column has additional observations made by the researchers.

Questions	Prabodhini	myScheme
Q1. Which of the two applications would you like to use frequently?	13	2
Q2. Which of the two applications is more complex?	1	10
Q3. Which of the two applications is easier to use?	13	2
Q4. Which of the two applications do you think most people would learn to use very quickly?	15	0
Q5. Which of the two applications is more cumbersome to use?	5	10
Q6. Which of the two applications can you use more confidently?	14	1

Table 3: Participant response on the usability questionnaire, evaluating application usability in terms of frequency of use, complexity, ease of use, and user confidence. Columns 2 and 3 represent the number of users preferring the respective app. For Q2, four users responded that neither platform was complex.

ban/rural", "BPL", "organised sector"), and absence of Telugu language support. These issues reflect a deeper misalignment between the design assumptions of such platforms and the mental models of their intended users. This mismatch can be understood through Norman's concept of the *Gulf of Execution* (Norman, 1986), which describes the gap between a user's goals and the actions a system requires to accomplish them. For many participants, myScheme demanded an understanding of administrative terms, hierarchical filters, and input formats that did not map onto their everyday knowledge or vocabulary. In contrast, Prabodhini's design—anchored in natural language input, voice interaction, and scenario-driven queries—narrowed this gulf by allowing users to express goals in their own terms and receive structured information in response. The system's conversational structure and its allowance for open-ended inputs also align with the principle underpinning *Zipf's Law of Least Effort* (Zipf, 2016), which suggests that users prefer interaction paths that require minimal cognitive and physical effort. By enabling users to speak queries naturally—rather than navigate nested menus or input structured forms—Prabodhini reduced friction and encouraged continued engagement. Participants' ability to complete tasks more quickly and independently is indicative of an interface that leverages interaction affordances suited to its target users.

The preference for voice input also underscores the growing familiarity of emergent users with conversational modalities. Participants likened Prabodhini to Google's voice search, referencing their existing use of voice-based interaction in apps like YouTube or messaging platforms. This familiarity and the sense of control it fostered contributed to the success of Prabodhini's mobile interface and demonstrate how leveraging well-understood input methods can enhance usability.

Our study also revealed that the quality and reliability of voice interaction matter greatly. Some participants struggled to discern the output when Prabodhini defaulted to an offline TTS engine lacking natural Indian accents. Others were confused when the app failed to clearly indicate whether it was actively listening. These issues highlight the importance of responsive feedback and robust system design—particularly in mobile contexts where connectivity may be intermittent. Future iterations must incorporate more effective visual and auditory cues to support multimodal interaction feedback.

Our findings reinforce the importance of localising interface language and terminology. Terms like "city" and "village" resonated more with participants than "urban" or "rural", illustrating how familiar vocabulary can reduce cognitive load. Prior work has shown that culturally resonant interfaces enhance user trust and improve task success among underserved groups (Medhi et al., 2010; Soares, 2015). Designers of mobile services must move beyond mere translation and towards localisation strategies that reflect users' linguistic, social, and cognitive contexts. Designing mobile governance platforms for emergent users demands resilient, mobile-native, voice-first interfaces. Scaling such systems requires addressing language diversity, interface robustness, and continuous participatory evaluation.

This study contributes to the intersection of HCI and NLP by showing how large language models and speech interfaces can support information access in low-resource settings. Prabodhini combines GPT-4 with a Retrieval-Augmented Generation (RAG) pipeline, speech-to-text input, and text-to-speech output to support natural language queries in regional languages. These components helped reduce the cognitive load of form-based systems and enabled mobile-first, low-literate users to find relevant government scheme information. Our findings offer design implications for building inclusive conversational agents that work reliably in multilingual, low-connectivity environments. Future NLP systems must prioritise transparency, localisation, and robustness to serve marginalised users more effectively.

6 Limitation

While our study design included three tasks, participant interaction with the myScheme application was largely limited to Task 1. Only two participants completed subsequent tasks using myScheme, and six were unable to use it at all. While this restricted direct comparison across all tasks, it also underscores the practical usability barriers present in myScheme. Thus, our analysis focuses on Task 1, where comparable engagement was feasible across both systems.

7 Acknowledgements

We would like to sincerely thank the Maintenance Division, BITS Pilani Hyderabad Campus, for connecting us with the participants for this study. We also extend our gratitude to the participants for their voluntary participation, and to the reviewers for their valuable guidance in refining this paper. Their support and feedback were instrumental in enabling us to engage with the community and improve the quality of our work.

References

Syed Ishtiaque Ahmed, Steven J Jackson, Maruf Zaber, Mehrab Bin Morshed, Md Habibullah Bin Ismail, and Sharmin Afrose. 2013. Ecologies of use and design: individual and social practices of mobile phone use within low-literate rickshawpuller communities in urban bangladesh. In *Proceedings of the 4th Annual Symposium on Computing for Development*, pages 1–10.

Anurag Aribandi, Divyanshu Agrawal, and Dipanjan Chakraborty. 2022. Note: Evaluating trust in the context of conversational information systems for new users of the internet. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '22, page 574–578, New York, NY, USA. Association for Computing Machinery.

Brendan E Asogwa. 2013. Electronic government as a paradigm shift for efficient public services: Opportunities and challenges for nigerian government. *Library Hi Tech*, 31(1):141–159.

Simone Balloccu and Ehud Reiter. 2022. Comparing informativeness of an nlg chatbot vs graphical app in diet-information domain. *arXiv preprint arXiv:2206.13435*.

Raluca Budiu. 2018. The user experience of chatbots. https://www.nngroup.com/articles/chatbots/.

C. R. Chaitra, Prajna Upadhyay, and Dipanjan Chakraborty. 2025. Is chatgpt ready for indian-language speakers? findings from a preliminary mixed methods study. In *Human-Computer Interaction. Design and Research*, pages 193–214, Cham. Springer Nature Switzerland.

Dipanjan Chakraborty, Mohd Sultan Ahmad, and Aaditeshwar Seth. 2017. Findings from a Civil Society Mediated and Technology Assisted Grievance Redressal Model in Rural India. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, ICTD '17, New York, NY, USA. ACM.

Dipanjan Chakraborty and Aaditeshwar Seth. 2015. Building Citizen Engagement into the Implementation of Welfare Schemes in Rural India. In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*, ICTD '15, pages 22:1–22:10, New York, NY, USA. ACM.

Priyanka Chandel and Pankaj Doke. 2013. A comparative study of voice and graphical user interfaces with respect to literacy levels. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, pages 1–2.

Elfreda A Chatman. 1991. Life in a small world: Applicability of gratification theory to information-seeking behavior. *Journal of the American Society for information science*, 42(6):438–449.

Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology*, 12(3):297–298.

Sebastien Cuendet, Indrani Medhi, Kalika Bali, and Edward Cutrell. 2013. Videokheti: Making video content accessible to low-literate and novice users. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2833–2842.

Andy Dearden and Dorothea Kleine. 2018. Minimum ethical standards for ictd/ict4d research.

Brenda Dervin and Charles M Naumer. 2009. Sense-making. *Encyclopedia of communication theory*, 2:876–880.

Mrunal Dhaygude and Dipanjan Chakraborty. 2020. Rethinking design of digital platforms for emergent users: Findings from a study with rural indian farmers. In *Proceedings of the 11th Indian Conference on Human-Computer Interaction*, pages 62–69.

Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of hci. *interactions*, 24(4):38–42.

Jose A Guridi, Cristobal Cheyre, and Qian Yang. 2025. Thoughtful adoption of nlp for civic participation: Understanding differences among policymakers. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–27.

Hendrik Heuer and Daniel Buschek. 2021. Methods for the design and evaluation of HCI+NLP systems. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 28–33, Online. Association for Computational Linguistics.

Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. 2018. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22.

Vivan Jain, Srivant Vishnuvajjala, Pranathi Voora, Bhaskar Ruthvik Bikkina, Bharghavaram Boddapati, C. R. Chaitra, Dipanjan Chakraborty, and Prajna Upadhyay. 2025. Prabodhini: Making large language models inclusive for low-text literate users. In *Advances in Information Retrieval*, pages 438–444, Cham. Springer Nature Switzerland.

Cathy Jiao, Aaron Steinfeld, and Maxine Eskenazi. 2024. Examining prosody in spoken navigation instructions for people with disabilities. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 1–12, Mexico City, Mexico. Association for Computational Linguistics.

Matt Jones, Simon Robinson, Jennifer Pearson, Manjiri Joshi, Dani Raju, Charity Chao Mbogo, Sharon Wangari, Anirudha Joshi, Edward Cutrell, and Richard Harper. 2017. Beyond "yesterday's tomorrow": future-focused mobile interaction design by and for emergent users. *Personal and Ubiquitous Computing*, 21:157–171.

Anirudha Joshi, Mandar Rane, Debjani Roy, Nagraj Emmadi, Padma Srinivasan, N Kumarasamy, Sanjay Pujari, Davidson Solomon, Rashmi Rodrigues, DG Saple, and 1 others. 2014. Supporting treatment of people living with hiv/aids in resource limited settings with ivrs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1595–1604.

Abhishek Kaushik and Gareth JF Jones. 2023. Comparing conventional and conversational search interaction using implicit evaluation methods. *arXiv preprint arXiv:2303.09258*.

Konstantinos Kazakos, Siddhartha Asthana, Madeline Balaam, Mona Duggal, Amey Holden, Limalemla Jamir, Nanda Kishore Kannuri, Saurabh Kumar, Amarendar Reddy Manindla, Subhashini Arcot Manikam, and 1 others. 2016. A real-time ivr platform for community radio. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 343–354.

Neesha Kodagoda, Wong Wong, and Nawaz Kahan. 2009. Identifying information seeking behaviours of low and high literacy users: combined cognitive task analysis. In *9th Bi-annual International Conference on Naturalistic Decision Making (NDM9)*. BCS Learning & Development.

Eugene Malthouse. 2023. Confirmation bias and vaccine-related beliefs in the time of covid-19. *Journal of Public Health*, 45(2):523–528.

Indrani Medhi, Ed Cutrell, and Kentaro Toyama. 2010. It's not just illiteracy. In *India HCI 2010/Interaction Design & International Development 2010*. BCS Learning & Development.

Indrani Medhi, Somani Patnaik, Emma Brunskill, SN Nagasena Gautama, William Thies, and Kentaro Toyama. 2011. Designing mobile interfaces for novice and low-literacy users. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 18(1):1–28.

Indrani Medhi, Aman Sagar, and Kentaro Toyama. 2006. Text-free user interfaces for illiterate and semiliterate users. In *2006 international conference on information and communication technologies and development*, pages 72–82. IEEE.

Joanna Mehtälä and Marko Nieminen. 2019. Combining design science and user-centred methods in m-government service design in namibia. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, pages 244–254.

Giselle Lorrane Nobre Melo, Nicoly Da Silva Menezes, Ingrid Moreira Miranda Da Silva, Luciano Arruda Teran, and Marcelle Pereira Mota. 2023. Inspecting the accessibility of instant payment systems from the perspective of low literacy people. In *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems*, pages 1–11.

Khadijah D Mohammed, Victoria Uren, Sian Joel-Edgar, and Priscilla Omonedo. 2023. Usability and user experience of mobile applications: A case of functional illiterates in nigeria. In *Proceedings of the 4th African Human Computer Interaction Conference*, pages 98–105.

Tricia J Ngoon, Vikram Kamath Cannanure, Kaja Jasinska, Sharon Wolf, and Amy Ogan. 2024. " i believe i did not preach into the desert": Opportunities & challenges in scaling teacher mentorship through mobile technology in rural côte d'ivoire. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 232–242.

Donald A Norman. 1986. Cognitive engineering. *User centered system design*, 31(61):2.

Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S Parikh. 2009. A comparative study of speech and dialed input voice interfaces in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 51–54.

Archana Prasad, Sean Blagsvedt, Tej Pochiraju, and Indrani Medhi Thies. 2019. Dara: A chatbot to help indian artists and designers discover international opportunities. In *Proceedings of the 2019 on Creativity and Cognition*, pages 626–632. Association for Computing Machinery.

Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. 2017. " alexa is my new bff" social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pages 2853–2859.

Andaleeb Rahman and Prabhu Pingali. 2024. Social welfare 'schemes' to an economic security 'system'. In *The Future of India's Social Safety Nets: Focus, Form, and Scope*, pages 357–425. Springer.

Jaisheen Kour Reen, Gerry Chan, and Rita Orji. 2024. icare: Insights from the evaluation of an app for managing stress among working-class indian women. *International Journal of Human–Computer Interaction*, pages 1–20.

Mark A Robinson. 2010. An empirical analysis of engineers' information behaviors. *Journal of the American Society for information Science and technology*, 61(4):640–658.

Rajdeep Roy. 2024. Conversational ai chatbots vs traditional customer support: Which is better for your business? `https://shorturl.at/XgwKu`.

S Sakthivel and Pinaki Joddar. 2006. Unorganised sector workforce in india: trends, patterns and social security coverage. *Economic and Political Weekly*, pages 2107–2114.

Hirav Shah and Amit Sengupta. 2018. Designing mobile based computational support for low-literate community health workers. *International Journal of Human-Computer Studies*, 115:1–8.

Marcos André Barroso Soares. 2015. Designing culturally sensitive icons for user interfaces: An approach for the interaction design of smartphones in developing countries. Master's thesis, Universidade do Porto (Portugal).

Vivek Srinivasan, Vibhore Vardhan, Snigdha Kar, Siddhartha Asthana, Rajendran Narayanan, Pushpendra Singh, Dipanjan Chakraborty, Amarjeet Singh, and Aaditeshwar Seth. 2013. Airavat: An Automated System to Increase Transparency and Accountability in Social Welfare Schemes in India. In *Proceedings of the Sixth International Conference on Information and Communications Technologies and Development: Notes - Volume 2*, ICTD '13, pages 151–154, New York, NY, USA. ACM.

Ayushi Srivastava, Shivani Kapania, Anupriya Tuli, and Pushpendra Singh. 2021. Actionable ui design guidelines for smartphone applications inclusive of low-literate users. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Sharifa Sultana, Renwen Zhang, Hajin Lim, and Maria Antoniak. 2022. Narrative datasets through the lenses of NLP and HCI. In *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–54, Seattle, Washington. Association for Computational Linguistics.

Indrani Medhi Thies and 1 others. 2015. User interface design for low-literate and novice users: Past, present and future. *Foundations and Trends® in Human–Computer Interaction*, 8(1):1–72.

Kristen Vaccaro, Tanvi Agarwalla, Sunaya Shivakumar, and Ranjitha Kumar. 2018. Designing the future of personal fashion. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Pieter Verdegem and Gino Verleye. 2009. User-centered e-government in practice: A comprehensive model for measuring user satisfaction. *Government information quarterly*, 26(3):487–497.

Christian Wagner. 2004. Wiki: A technology for conversational knowledge management and group collaboration. *Communications of the association for information systems*, 13(1):19.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186.

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books.

Out of the Box, into the Clinic? Evaluating State-of-the-Art ASR for Clinical Applications for Older Adults

Bram van Dijk[1], **Tiberon Kuiper**[3], **Sirin Aoulad si Ahmed**[3], **Armel Lefebvre**[1], **Jake Johnson**[3], **Jan Duin**[3], **Simon Mooijaart**[3,4], **Marco Spruit**[1,2]
[1]Department of Public Health and Primary Care, Leiden University Medical Center
[2]Leiden Institute of Advanced Computer Science, Leiden University
[3]Center for Medicine for Older People, Leiden University Medical Center
[4]Department of Internal Medicine, Leiden University Medical Center
Correspondence: b.m.a.van_dijk@lumc.nl

Abstract

Voice-controlled interfaces can support older adults in clinical contexts – with chatbots being a prime example – but reliable Automatic Speech Recognition (ASR) for underrepresented groups remains a bottleneck. This study evaluates state-of-the-art ASR models on language use of older Dutch adults, who interacted with the `Welzijn.AI` chatbot designed for geriatric contexts. We benchmark generic multilingual ASR models, and models fine-tuned for Dutch spoken by older adults, while also considering processing speed. Our results show that generic multilingual models outperform fine-tuned models, which suggests recent ASR models can generalise well out of the box to real-world datasets. Moreover, our results indicate that truncating generic models is helpful in balancing the accuracy-speed trade-off. Nonetheless, we also find inputs which cause a high word error rate and place them in context.

1 Introduction

Although there is a surge of interest in AI-driven applications like chatbots in the health domain (Guo et al., 2024; Huo et al., 2025), tailoring them to groups underrepresented in AI research remains a challenge. Older populations are one example: because they often have different needs and preferences when interacting with AI (van Dijk et al., 2025; Klaassen et al., 2025), their involvement in system development is key for building clinically relevant systems in geriatrics, the field in healthcare concerned with the health of older adults. This population is increasing in size globally (World Health Organisation, 2023), while personnel shortages in healthcare become pressing (Eurofound, 2023); yet addressing these challenges with AI warrants systems that align well with older adults.

Voice control is a key element in chatbots in geriatrics, as older individuals may struggle with small fonts, icons, and typing text in standard interfaces (Khamaj, 2025). Yet, implementing voice control is not obvious, as the performance of Automatic Speech Recognition (ASR) systems depends on the representation of older adults in training data, and on their articulation, speech volume and technological literacy (Klaassen et al., 2025). Moreover, evaluation of state-of-the-art ASR systems on realistic data of older adults is lagging behind.

In this short paper, we evaluate recent ASR models on older adults' language use in interaction with `Welzijn.AI`. This is a new digital platform for older users, that currently features a prototype chatbot to converse about clinically relevant topics like quality of life and frailty. Audio data of 10 older Dutch adults interacting with `Welzijn.AI` were transcribed using generic multilingual ASR models (Whisper and Voxtral), and models fine-tuned specifically for Dutch or older Dutch populations (Whisper and Wav2Vec2). We find that i) generic multilingual models outperform fine-tuned models, and also that ii) truncating larger generic models helps striking a good balance between accuracy and speed. These findings also hold for a similar subset of the Mozilla Common Voice dataset (Ardila et al., 2020) we use as benchmark.

2 Background

Though work on evaluating ASR models for older adults exists, extrapolating findings to realistic contexts is hard. Performance of ASR models sometimes remains implicit in downstream use, for example through evaluation of overlap of linguistic features extracted from model vs. human-generated transcripts (Naffah et al., 2025), or by using only part of ASR models (like the audio encoder) in predicting cognitive impairment in older persons (Agbavor and Liang, 2024). Other work tailors ASR systems to older adults by drawing on additional databases of individual users (Xu et al., 2025), which is not always feasible in a clinical

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 72–78

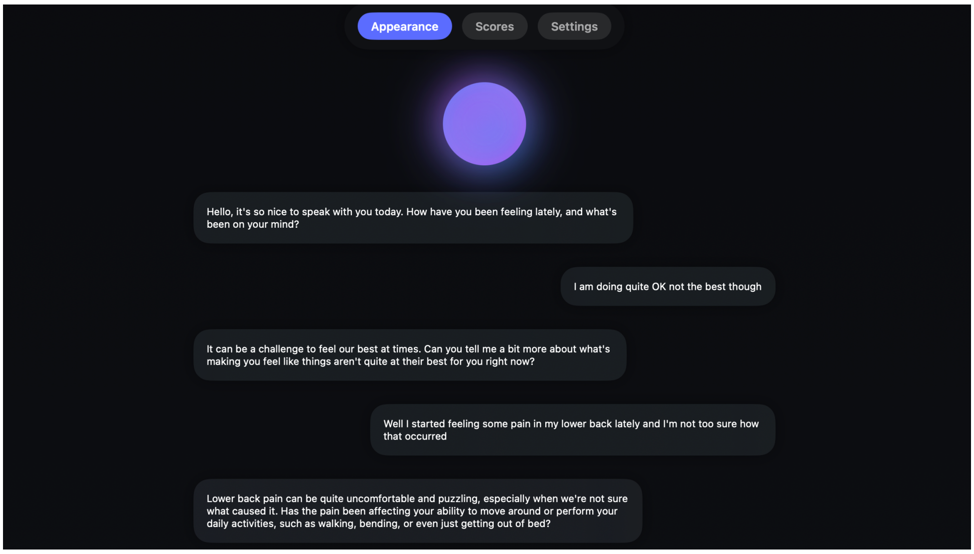


Figure 1: Interface of `Welzijn.AI` with an example conversation. Users press the purple button to activate the ASR functionality and start responding, after which their speech is transcribed and rendered on the screen. Chatbot responses are read out with a text-to-speech model. The 'Scores' button shows information extracted on quality of life and frailty, 'Settings' allows choosing different ASR models, and 'Appearance' returns the user to the conversation on display. We focus in this paper on the conversation resulting from interaction with this prototype.

context due to privacy concerns. However, Xu et al. (2025) also show that fine-tuning generic multilingual models with speech from older adults increases ASR performance. The work by Shekoufandeh et al. (2025) explores this further, by fine-tuning Whisper as recent ASR model (Radford et al., 2023) on the Dutch JASMIN-CGN dataset. This dataset includes language use of older adults in human-machine interaction settings (Cucchiarini and Van hamme, 2012), which is potentially relevant to `Welzijn.AI`.

3 Materials and methods

Ten older Dutch adults (≥65 years) from a volunteer panel of the outpatient clinic for geriatrics at the Leiden University Medical Center (LUMC) were included. The Institutional Review Board of the LUMC approved the study and all participants provided informed consent. Participants were instructed not to share personal health information with `Welzijn.AI`, but rather to impersonate a peer. Participants interacted individually with the chatbot, with one experimenter standing by.

An impression of `Welzijn.AI` is given in Figure 1. Here we focus on the system's ability to support conversations; we refer for architectural details to van Dijk et al. (2025). The chatbot was driven by the `meta-llama/Llama-3.3-70B-Instruct`[1] model (Grattafiori et al., 2024), prompted to structure the conversation around the EQ-5D and Clinical Frailty Scale (Brooks, 1996; Rockwood et al., 2005). These are validated geriatric instruments to assess quality of life and frailty, by retrieving information about mobility, mental wellbeing, physical independence, and so on. These instruments can be presented via surveys, in conversations with clinicians, or in our case, by a chatbot (Figure 1).

The chatbot was used on a laptop. Interactions took 5-10 minutes and were recorded with a handheld device. The default ASR model in `Welzijn.AI` was `openai/whisper-large-v3`, which in early testing we found to work best.

Type	Example
Orthographic	Hi, I am uh... feeling great today.
Orthographic_clean	Hi I am feeling great today
Normalised	hi i am feeling great today

Table 1: Examples of gold transcript types.

Hugging Face ID	Params
`mistralai/Voxtral-Mini-3B-2507`	4.68B
`openai/whisper-large-v2`	1.55B
`openai/whisper-large-v3`	1.55B
`golesheed/whisper-native-elderly-9-dutch`	1.54B
`golesheed/wav2vec2-xls-r-1b-dutch-3`	963M
`openai/whisper-large-v3-turbo`	809M
`openai/whisper-medium`	769M
`openai/whisper-small`	244M

Table 2: Models used for our ASR experiments.

[1]We will use Hugging Face IDs to denote AI models.

Hugging Face ID	WER Welzijn.AI				WER Common Voice			
	Orth.	Orth_c.	Norm.	*Time*	Orth.	Orth_c.	Norm.	*Time*
mistralai/Voxtral-Mini-3B-2507	.17	.11	.09	3.75	.05	.04	.04	3.72
openai/whisper-large-v2	.19	.12	.10	4.69	.05	.04	.04	4.24
openai/whisper-large-v3	**.12**	**.07**	**.06**	3.41	**.04**	**.03**	**.03**	3.71
golesheed/whisper-native-elderly-9-dutch	.40	.22	.14	4.59	.29	.18	.07	4.07
golesheed/wav2vec2-xls-r-1b-dutch-3	.49	.37	.36	**.99**	.30	.19	.19	**.89**
openai/whisper-large-v3-turbo	.16	.10	.08	1.43	.06	.04	.04	1.47
openai/whisper-medium	.19	.13	.11	2.35	.07	.06	.06	2.48
openai/whisper-small	.26	.18	.17	1.11	.12	.10	.10	1.19

Table 3: Word Error Rate (WER) is the edit distance between prediction and reference (sum of substitutions, deletions, and insertions), divided by the length of the reference, so here denotes the average number of errors per reference word, for Orthographic (*Orth.*), Orthographic_clean (*Orth_c.*), and Normalised (*Norm.*) gold transcripts. Processing time (*Time*) in average seconds per input. Best results in bold.

Recorded user speech was after the interactions separated from chatbot responses and segmented using `PyDub`[2] and `pyannote` (Bredin, 2023). We obtained 199 segments with an average length of 3.4 seconds, due to the turn-taking nature of the conversation. Our sample totalled 11 min. and 15 sec., so is small, but still valuable as data from clinical contexts is challenging to obtain. Besides chatbot data, we also drew 200 random samples from the Common Voice dataset (Ardila et al., 2020) of Dutch older individuals ($\geq$ 60 years), which totals 17 min., and 37 sec. As this data concerns written text read out loud, it should intuitively be an easier benchmark for ASR models.

For obtaining gold standard (i.e. human) reference transcriptions, segments were transcribed with the default `openai/whisper-large-v3` model, and subsequently corrected by the first author. Since the Word Error Rate (WER) as standard metric in the ASR field is sensitive to fillers, capitalisation, and punctuation, and since choosing the 'right' reference depends on the use case, we created three types of human and model transcriptions (via rule-based postprocessing) as visible in Table 1: *orthographic* (including fillers, capitalisation, and punctuation) *orthographic_clean* (only capitalisation, as also used in Figure 1), and *normalised* (no fillers, capitalisation or punctuation). Orthographic transcription is useful in that it provides additional structure, though speech content is often arguably sufficiently preserved in normalised transcriptions, with orthographic_clean transcription striking a balance between the strictest and most flexible evaluation scenarios.

Table 2 shows the ASR models included in our experiments. As can be seen, we focus on models from the `Whisper` model family as the current standard in the field, and include `mistralai/Voxtral-Mini-3B-2507` (Liu et al., 2025) as new potential competitor. We included `golesheed/whisper-native-elderly-9-dutch` as the Whisper model fine-tuned for older Dutch adults by Shekoufandeh et al. (2025), and we also included `golesheed/wav2vec2-xls-r-1b-dutch-3`, a work-in-progress Wav2Vec2 model (Baevski et al., 2020) fine-tuned for general Dutch, as older but potentially fast contender. We note that these fine-tuned models do not output fillers, capitals or punctuation by default, hence evaluate them with normalised transcripts.

Our motivation for this set of ASR models is that they are all small (compared to Voxtral's 24B variant for example) and open weights, hence suitable for local/private downstream applications. We take processing time into account since a good balance between accuracy and speed is key in many applications, so include models of different sizes.

Our code for the ASR pipeline is available.[3] All experiments were carried out on a Macbook Pro M1 16GB using the `Hugging Face` ecosystem for ASR and `PyTorch`'s MPS GPU acceleration backend. Due to privacy restrictions we cannot share recordings nor transcripts of the interactions.

4 Results

WER and processing time per model are given in Table 3. Regarding WER, on both the `Welzijn.AI` and Common Voice datasets, `openai/whisper-large-v3` as generic multilingual model outperforms all other models, also regarding normalised transcripts as

[2] https://github.com/jiaaro/pydub

[3] https://github.com/bma-vandijk/asr_pipelines

#	Model	Prediction	Reference	WER
1	openai/whisper-large-v2	hartelijk bedankt voor het kijken en tot de volgende keer *thank you cordially for watching and until next time*	gaat wel *it's okay*	5
2	golesheed/whisper-native-elderly-9-dutch	poet hem in oranje *loot him in orange*	goedemiddag *good afternoon*	4
3	openai/whisper-large-v2	ik denk dat het weer zo is als het altijd is *I think that it is as it always is*	ongeveer zoals altijd *roughly as usual*	3.33
4	openai/whisper-medium	ik ben benieuwd *I am curious*	goedemiddag *good afternoon*	3
5	openai/whisper-small	voor de wereld *for the world*	goedemiddag *good afternoon*	3
6	mistralai/Voxtral-Mini-3B-2507	groen de medaillon *green the medallion*	goedemiddag *good afternoon*	3
7	golesheed/wav2vec2-xls-r-1b-dutch-3	goede midda *good midda*	goedemiddag *good midday*	2
8	openai/whisper-large-v3-turbo	ja het is heel erg goed *yes it is very well*	weer te ingewikkeld *again too complicated*	2
9	openai/whisper-medium	bedankt voor het kijken *thank you for watching*	gaat wel *it's okay*	2
10	mistralai/Voxtral-Mini-3B-2507	het zelf gaat het *it self goes it*	hetzelfde uiteraard *the same of course*	2

Table 4: Example inputs with high WER at the sample level. English (literal) translations in italics.

most flexible evaluation scenario. In terms of processing time, on both datasets, the Wav2Vec2 model fine-tuned on Dutch language use (not specifically older language users) (golesheed/wav2vec2-xls-r-1b-dutch-3) is the fastest, but also the least accurate.

For chatbot systems like Welzijn.AI, understanding the trade-off between WER and processing time is crucial, given that in chatbots other components also impose processing time, and seconds may greatly impact the perceived quality of the experience. We visualise performance w.r.t. normalised transcripts in Figure 2. Here we see that openai/whisper-large-v3-turbo, which is essentially openai/whisper-large-v3 with a truncated decoder 1/8 its size, strikes the best balance. The 'nearest' improvement in WER concerns openai/whisper-large-v3, which is about three times slower, while the 'nearest' improvement in processing time comes from openai/whisper-small, at the expense of more than twice its WER. Figure 2 also shows that for all models, the Common Voice dataset is easier regarding WER, though not all models process these data faster. For Common Voice data the same observations hold regarding models that are the 'nearest' improvements in WER and processing times compared to openai/whisper-large-v3-turbo: they lead to considerable drops in speed or accuracy respectively. Our findings align with earlier work that shows that overall, larger ASR models perform better on commonly used datasets compared to smaller models (Atwany et al., 2025), though they also

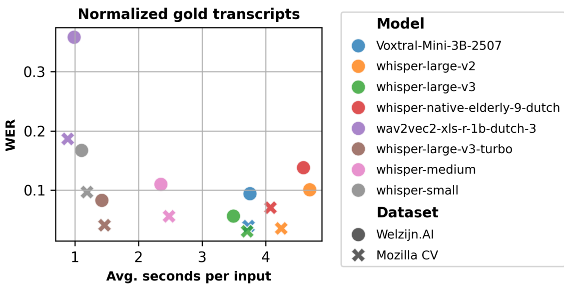


Figure 2: Overview of accuracy vs. processing time.

note that larger numbers of parameters eventually yield diminishing returns. This finding helps understand why openai/whisper-large-v3-turbo shows a relatively small performance drop while still being about 50% smaller than openai/whisper-large-v3.

4.1 Error analysis

To disclose common pitfalls in using ASR models on Dutch, we provide ten examples of the 50 predictions with highest WER (≥ 2) in Table 4, which as expected come only from Welzijn.AI data. Errors in predictions in deep neural ASR models like Whisper can be categorised in various ways. *Mishearings* could be induced by ambiguous or unclear phonemes in the input, or confusion of phonemes by the model (e.g. 'this guy' vs. 'the sky'). *Hallucinations* are errors where the prediction has no semantic or phonetic relation to the reference. *Looping* is when a model keeps repeating previously recognized speech ('Welcome to Amsterdam to Amsterdam to Amsterdam'). Furthermore, it is also known that deep neural ASR models are sensitive to non-speech audio signals in the background caused by e.g. objects or animals (Barański et al., 2025), which is not obvious to trace in the predictions, but in live interaction settings like `Welzijn.AI` something to take into account.

Examples 2, 3, 6, 7 and 10 seem cases of mishearings, where 2 and 6 have at least some phonetic (but not semantic) alignment with the reference; examples 3, 7 and 10 also have some semantic overlap. Examples 1, 4, 5, 8, 9 have no clear phonetic or semantic link with the reference thus qualify as hallucinations; 1 and 9 are probably frequency effects from Whisper's training data (video transcriptions) (Barański et al., 2025), meaning the models prioritize patterns in the training distribution over the actual audio input. All in all, our examples suggest high WER is not limited to just a few types of models, which aligns with earlier documented unpredictability in state-of-the-art ASR models across the board (Koenecke et al., 2024; Atwany et al., 2025) , and this finding should inform their development and deployment in high-stakes contexts.

5 Discussion and conclusion

We evaluated state-of-the-art ASR models on transcribing language use of older adults interacting with the `Welzijn.AI` chatbot, which was designed for geriatrics. We included various generic multilingual models as well as models fine-tuned on language use of older Dutch adults and general Dutch. We found that on both `Welzijn.AI` and Common Voice data, generic multilingual models perform better than fine-tuned models, with `openai/whisper-large-v3` as best model achieving WERs of .06 and .12 for normalised and orthographic transcriptions of realistic `Welzijn.AI` conversations. Interestingly, its truncated variant `openai/whisper-large-v3-turbo` struck the best balance between accuracy and processing speed, the latter being crucial in chatbot systems used in real-time. This is useful from a systems development perspective, since truncated models may perform well out of the box, without the need for training smaller architectures from scratch, or for additional data for group- or task-specific fine-tuning. Future work should further support this claim by using larger samples across settings.

To put our results in perspective, other evaluations of state-of-the-art ASR models on the entire Dutch subset of the Common Voice dataset reported WERs of .06 by `mistralai/Voxtral-Mini-3B-2507` (Liu et al., 2025), and .04 by `openai/whisper-v3-large`, assuming the strictest evaluation scenario (orthographic).[4] We obtained similar results for our subset of Common Voice spoken by Dutch Adults: .04 for the same Whisper and .05 for the same Voxtral model. Hence, it seems that for generic multilingual models, changing the target population does not imply large performance degradation if the task is straightforward (reading text out loud).

Still, our `Welzijn.AI` data is conversational, hence results are harder to put in perspective. Though the WER on orthographic transcripts for `Welzijn.AI` data triples for the same Voxtral and Whisper models (.17 and .12 respectively) , given the different nature of read speech and conversational language, this is not a dramatic loss of performance. Recent work has reported WERs for a variety of datasets transcribed with our best performing model `openai/whisper-v3-large`, as large as .32 for English speech recordings in home environments (BERSt), and .23 for English meeting recordings (AMI) (Atwany et al., 2025).

We also attempted to categorize WER error types. We saw that mishearings were as frequent as hallucinations. Hallucinations, however, are potentially more problematic for systems like `Welzijn.AI`, as for a user who is unable to make sense of the resulting transcription, trust will erode faster compared to a mishearing, which still has some semantic or phonetic similarity.

Strategies to improve WER and mitigate hallucinations, could include more independent language modelling components that take specific contexts into account. When the audio input for the decoder is noisy, its language prior generates a transcript based on its own distribution instead of the input, which may well be out-of-context. Mitigation could involve generating candidate predictions and evaluating their likelihoods from the perspective of a domain-specific language model, or combining the decoder's predicted token probabilities directly with the prediction of a context-specific language model (see also Zhou and Li, 2025). Hence, some exciting work remains for making real-world impact with recent ASR models.

[4]Results of latest model reported on `https://github.com/openai/whisper`, August 6 2025.

6 Limitations

Though dedicated GPUs in high performing computing clusters have higher bandwidths and are faster, hence the default choice in experiments like ours, unified architectures (such as provided by Apple's Silicon M series) are receiving more attention nowadays due to their benefits in terms of latency, energy consumption, and system footprint, and are recognized as efficient and competitive alternatives to dedicated GPUs (Hübner et al., 2025; Kenyon and Capano, 2022). So faster processing times could probably be attained by using dedicated GPUs or by using a more recent M-chip. Still, we anticipate a scenario where one device hosts multiple smaller AI models to do different tasks, for which our setup can provide a good lower bound.

Also, though we tried to make comparison fair for different ASR models by evaluating with different kinds of transcripts, developing further transcript normalisations to take 'acceptable errors' into account, e.g. writing numbers in digits or letters ('8', 'eight'), were beyond the scope of the current work. This means that there can be some noise in our performance estimates.

References

Felix Agbavor and Hualou Liang. 2024. Multilingual prediction of cognitive impairment with large language models and speech analysis. *Brain sciences*, 14(12):1292.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Hanin Atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, and Bhiksha Raj. 2025. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. *arXiv preprint arXiv:2502.12414*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. Investigation of whisper asr hallucinations induced by non-speech audio. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *24th INTERSPEECH Conference (INTERSPEECH 2023)*, pages 1983–1987. ISCA.

Richard Brooks. 1996. EuroQol: the current state of play. *Health policy*, 37(1):53–72.

Catia Cucchiarini and Hugo Van hamme. 2012. The JASMIN speech corpus: recordings of children, non-natives and elderly people. In *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme*, pages 43–59. Springer.

Eurofound. 2023. Measures to tackle labour shortages: Lessons for future policy. https://www.eurofound.europa.eu/en/publications/2023/measures-tackle-labour-shortages-lessons-future-policy. (accessed 26 March 2025).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, and 1 others. 2024. Large Language Models for Mental Health Applications: Systematic Review. *JMIR mental health*, 11(1):e57400.

Paul Hübner, Andong Hu, Ivy Peng, and Stefano Markidis. 2025. Apple vs. Oranges: Evaluating the Apple Silicon M-Series SoCs for HPC Performance and Efficiency. *arXiv preprint arXiv:2502.05317*.

Bright Huo, Amy Boyle, Nana Marfo, Wimonchat Tangamornsuksan, Jeremy P Steen, Tyler McKechnie, Yung Lee, Julio Mayol, Stavros A Antoniou, Arun James Thirunavukarasu, and 1 others. 2025. Large language models for chatbot health advice studies: A systematic review. *JAMA Network Open*, 8(2):e2457879–e2457879.

Connor Kenyon and Collin Capano. 2022. Apple Silicon Performance in Scientific Computing. In *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–10. IEEE.

Abdulrahman Khamaj. 2025. AI-enhanced chatbot for improving healthcare usability and accessibility for older adults. *Alexandria Engineering Journal*, 116:202–213.

Willemijn Klaassen, Bram van Dijk, and Marco Spruit. 2025. A Review of Challenges in Speech-Based Conversational AI for Elderly Care. *Studies in health technology and informatics*, 327:858–862.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.

Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, and 1 others. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.

Ava Naffah, Valeria A Pfeifer, and Matthias R Mehl. 2025. Spoken Language Analysis in Aging Research: The Validity of AI-Generated Speech to Text Using OpenAI's Whisper. *Gerontology*, 71(5):417–423.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Kenneth Rockwood, Xiaowei Song, Chris MacKnight, Howard Bergman, David B Hogan, Ian McDowell, and Arnold Mitnitski. 2005. A global clinical measure of fitness and frailty in elderly people. *CMAJ*, 173(5):489–495.

Golshid Shekoufandeh, Paul Boersma, and Antal van den Bosch. 2025. Improving the inclusivity of dutch speech recognition by fine-tuning Whisper on the JASMIN-CGN corpus. *arXiv preprint arXiv:2502.17284*.

Bram MA van Dijk, Armel EJL Lefebvre, and Marco R Spruit. 2025. Welzijn. AI: Developing responsible conversational AI for the care of older people through stakeholder involvement. *Maturitas*, 199(108616).

World Health Organisation. 2023. Mental health of older adults. https://www.who.int/news-room/fact-sheets/detail/mental-health-of-older-adults. (accessed 29 April 2025).

Haiying Xu, Haoze Liu, Mingshi Li, Siyu Cai, Guangxuan Zheng, Yuhuang Jia, Jinghua Zhao, and Yong Qin. 2025. EchoVoices: Preserving Generational Voices and Memories for Seniors and Children. *arXiv preprint arXiv:2507.15221*.

Shilin Zhou and Zhenghua Li. 2025. Improving Contextual ASR via Multi-grained Fusion with Large Language Models. *arXiv preprint arXiv:2507.12252*.

MobileA3gent: Training Mobile GUI Agents Using Decentralized Self-Sourced Data from Diverse Users

Wenhao Wang[1,3,4] **Mengying Yuan**[5] **Zijie Yu**[2,3,4] **Guangyi Liu**[1] **Rui Ye**[2,3] **Jin Tian**[2,3]
Siheng Chen[2,3,4*] **Yanfeng Wang**[2,4*]
[1]Zhejiang University [2]Shanghai Jiao Tong University
[3]Multi-Agent Governance & Intelligence Crew (MAGIC)
[4]Shanghai AI Laboratory
[5]Wuhan University
12321254@zju.edu.cn sihengc@sjtu.edu.cn

Abstract

The advancement of mobile GUI agents has opened new opportunities for automating tasks on mobile devices. Training these agents requires large-scale high-quality data, which is prohibitively expensive when relying on human labor. Given the vast population of global mobile phone users, if automated data collection from them becomes feasible, the resulting data volume and the subsequently trained mobile agents could reach unprecedented levels. Nevertheless, two major challenges arise: (1) extracting user instructions without human intervention and (2) utilizing distributed user data while preserving privacy. To tackle these challenges, we propose **MobileA3gent**, a collaborative framework that trains mobile GUI Agents using self-sourced data from diverse users. The framework comprises two components, each targeting a specific challenge: (1) **Auto-Annotation**, which enables the automatic collection of high-quality datasets during users' routine phone usage with minimal cost. (2) **FedVLM-A**, which enhances federated VLM training under non-IID distributions by incorporating adapted global aggregation based on both episode-level and step-level variability. Extensive experiments prove that MobileA3gent achieves superior performance over traditional approaches at only 1% of the cost, highlighting its potential for real-world applications.

1 Introduction

Mobile GUI agents (Bai et al., 2024; Wang et al., 2024b,a) have experienced significant advancements, propelled by recent progress in Vision-Language Models (VLMs). Designed to simulate human mobile phone usage behavior, mobile agents can automate complex tasks on mobile phones, saving tremendous human labor and change everyday lives. Compared to non-agent solutions, mobile agents offer significantly better adaptability and generalizability, enabling them to effectively handle various mobile environments and operation scenarios (Zhang et al., 2023).

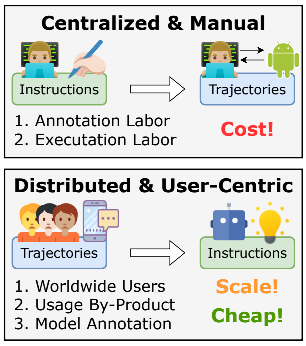


Figure 1: Comparing our proposed paradigm with conventional ones. By leveraging users' daily phone usage, we achieve superior scalability with drastic cost savings.

The training of mobile agents heavily depends on large-scale, high-quality datasets (Chai et al., 2024; Zhang et al., 2024c). To build such datasets, existing approaches rely on centralized data collection followed by human annotation, resulting in high costs and limited scalability. To achieve large-scale data acquisition more efficiently, a paradigm shift (as shown in Figure 1) from **centralized** to **distributed** data collection is necessary, enabling diverse users to participate in data contribution. Additionally, replacing **human** annotation with **automatic** annotation is crucial for efficiently processing the vast amount of collected data, allowing direct data sourcing from real user interactions.

Our insight is that the frequent and ever-growing phone usage by users worldwide naturally generates valuable supervisory information, which can serve as a rich data source for training mobile agents. Building on this user-centric insight, we aim to effectively utilize these distributed data, while minimizing human involvement in the process. However, two technical challenges remain:

1. Although the users' phone usage provides real-world trajectories (screenshots and actions), it is difficult to extract the real intentions (instructions) behind the actions in natural language;
2. Data collected from one user is both scale-

* Corresponding authors.

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 79–112
November 9, 2025 ©2025 Association for Computational Linguistics

limited and privacy-sensitive. The challenge lies in how to utilize distributed data from diverse users to boost performance while protecting privacy.

To tackle these challenges, we propose **MobileA3gent**, a collaborative learning framework that trains mobile agents using automatically collected user data from daily phone interactions while preserving user privacy. Specifically, MobileA3gent features two novel techniques.

First, we propose **Auto-Annotation**, an automated method for data collection and annotation that leverages locally deployed VLMs to annotate user instructions based on interaction trajectories. The key technical innovation lies in combining step-wise low-level instruction breakdowns with episode-wise summarization, allowing even small local VLMs to better understand the user's intent. The step-wise description decomposes complex user instructions into simpler steps, enabling the VLM to comprehend and extract information more accurately. Meanwhile, the episode-wise summarization provides a global perspective on the entire task, generating a more comprehensive caption of the user's ultimate instruction. Compared with human annotation, Auto-Annotation generates data of comparable quality with minimal cost requirement.

Second, to effectively utilize decentralized data from diverse users, we propose **FedVLM-A**, which pioneers the integration of Federated Learning (FL) (Kairouz et al., 2021) and collaborative training of VLM-based GUI agents, while ensuring rigorous user privacy protection. We further propose a novel aggregation method, termed **A**dapted global aggregation, which accounts for both episode-level and step-level distributions to handle the two-level heterogeneity (formulated in Section B) in diverse users' data, overcoming the limitations of traditional one-level aggregation methods (Karimireddy et al., 2021; McMahan et al., 2017; Hsu et al., 2019; Reddi et al., 2020). Adapted aggregation adapts the global aggregation weights using a weighted sum of episode and step counts for each client, thereby enhancing the performance of mobile agents trained in non-IID scenarios.

Extensive experiments on four benchmarks with 10+ models and metrics demonstrate that: (1) MobileA3gent achieves the best all-around trade-off across four dimensions, delivering performance on par with centralized manual approaches at significantly lower cost, while also ensuring privacy and achieving exceptional scalability. (2) Auto-Annotation outperforms all annotation baselines in performance while reducing annotation costs by 99% compared to manual labeling. (3) FedVLM-A achieves an at least 5% relative improvement over representative FL baselines in non-IID scenarios. These promising results underscore the immense potential of our framework to serve as a novel and practical paradigm for real-world applications. To summarize, our contributions are as follows:

1. We formulate the problem of self-sourced data collection from distributed mobile phone users and propose Auto-Annotation, an automatic data collection method, which achieves data quality comparable to human-annotated data at a significantly lower cost.
2. We introduce MobileA3gent, a collaborative framework for training mobile agents on decentralized user data while preserving privacy. By incorporating FedVLM-A, we enable federated training of VLMs and achieve superior performance when confronted with heterogeneity.
3. We conduct extensive experiments across comprehensive benchmarks and metrics. The compelling results highlight the substantial potential of our approach for real-world applications.

2 Problem Formulation

2.1 Preliminaries

Data Composition. The mobile GUI agent, powered by a VLM, simulates human users and completes tasks in a step-wise process. To train the core VLM, one data episode, denoted as $\mathcal{D}$, comprises multiple steps, each serving as a basic training unit. A step consists of three components: a task instruction $\mathcal{T}$, a screenshot, and a corresponding action. The composition of a data episode is defined as: $\mathcal{D} = \{\langle \mathcal{T}, a_i, s_i \rangle \mid i \in [1, n]\}$, where $\langle \mathcal{T}, a_i, s_i \rangle$ represents the i-th step, with a_i and s_i denoting the action and screenshot respectively.

Traditional Approach. Automating mobile devices poses significant challenges, leading to a heavy reliance on high-accuracy training data, which are, at present, almost all annotated by humans. The traditional paradigm (Li et al., 2024b; Qin et al., 2025; Hong et al., 2023) thus involves: (1) manually authored task instructions, followed by (2) centralized data collection and model training. As shown in Figure 1, this approach typically outsources instruction writing to human annotators using predefined rules or heuristics to promote both quality and diversity. Each instruction is then executed step-by-step in a controlled environment, such as an Android simulator, to collect paired screenshots and actions. To guarantee correctness, all interactions are manually verified, resulting in substantial costs and difficulty in scaling.

2.2 Primary Problem

To overcome the high cost and limited scalability of the traditional paradigm, we introduce a novel distributed user-centric approach for training mobile agents. The primary problem we address is: **How to harness private and distributed phone usage trajectories from diverse users?** We further decompose the primary problem into two subordinate problems: (1) How to automatically collect data from individual users without incurring expensive human annotation; and (2) How to effectively utilize decentralized data to optimize the agent while preserving user privacy .

Sub-Problem 1: Automatic Data Annotation on User Side. During phone interactions, users spontaneously generate screenshots and actions, which are assumed to be easily collectible. However, users do not receive explicit natural language instructions and only act based on their underlying intentions, making task annotation necessary. Since users are generally reluctant to articulate their intentions and such intentions are non-trivial to infer, the first subordinate problem is: *how to automatically derive user intentions without human intervention*, thereby constructing the training dataset. The objective is to learn a function $f(\cdot)$ that predicts user intention $\mathcal{T}^*$, an approximation of task instruction $\mathcal{T}$, based on n steps of actions and screenshots $\langle a_i, s_i \rangle$, that is: $\mathcal{T}^* = f(\{\langle a_i, s_i \rangle\}_{i=1}^{n})$.

Sub-Problem 2: Distributed Training of Mobile GUI Agents. The daily phone usage of an individual generates a limited dataset, constraining the agent's performance trained solely on it. Fortunately, with millions of users worldwide, there is immense potential to collaboratively train a mobile agent using their combined data, enabling virtually unlimited scalability. Nevertheless, directly sharing user data poses significant privacy risks, necessitating its use in a distributed manner. Therefore the second subordinate problem is: *how to conduct privacy-preserving collaborative training of mobile agents on distributed user data.*

3 Methodology

3.1 Auto-Annotation: Automatic Data Collection and Annotation from Daily Phone Usage

Auto-Annotation functions by automatically building datasets from users' daily phone usage without manual effort. Screenshots and actions are directly recorded from user trajectories. To annotate user instructions, the idea is to employ a local annotation model to progressively decode user intent in a step-by-step manner, which comprises three stages: (1) Converting coordinate-based actions into semantically meaningful descriptions; (2) Incrementally generating low-level instructions to reflect each discrete operation; (3) Consolidating these atomic instructions into a high-level instruction for the entire episode. *Note: A low-level instruction is a specific, atomic directive that corresponds to an individual step, whereas a high-level instruction represents the overall task objective.*

Rule-Based Action Conversion. As indicated by previous works (Zheng et al., 2024), some VLMs, such as GPT-4V (202, 2023), are unable to effectively identify the location of operations. Therefore, to make the original actions interpretable to the local annotation model, we adopt a rule-based technique rather than using models (Wang et al., 2024a) to transform the action into a natural language sentence. Specifically, for `CLICK` actions, we align the exact click position with a corresponding interface element based on the accessibility tree. If the element contains text or invokes a function, we use the associated text or function name to construct a meaningful action description. For other actions, such as `NAVIGATE_HOME`, we slightly adjust the phrasing to improve clarity and readability. A code snippet is included in Appendix E.

Step-Wise Instruction Description. During this stage, we annotate users' low-level atomic instructions through step-wise description, a novel technique that decomposes complex user tasks into multiple steps. Specifically, at each step i, the local annotation model $\mathcal{M}_a$, referred to as the *Descriptor*, is prompted to generate an atomic instruction that reflects the user's explicit intent, as:

$$\text{Descriptor} : \langle s_i, A_i \rangle \xrightarrow{\mathcal{M}_a} \mathcal{T}_i^{\text{low}} , \tag{1}$$

where $\mathcal{T}_i^{\text{low}}$ is the prediction of user intention, serving as an approximation of the actual low-level instruction. s_i and A_i respectively represent the current screenshot and the corresponding converted

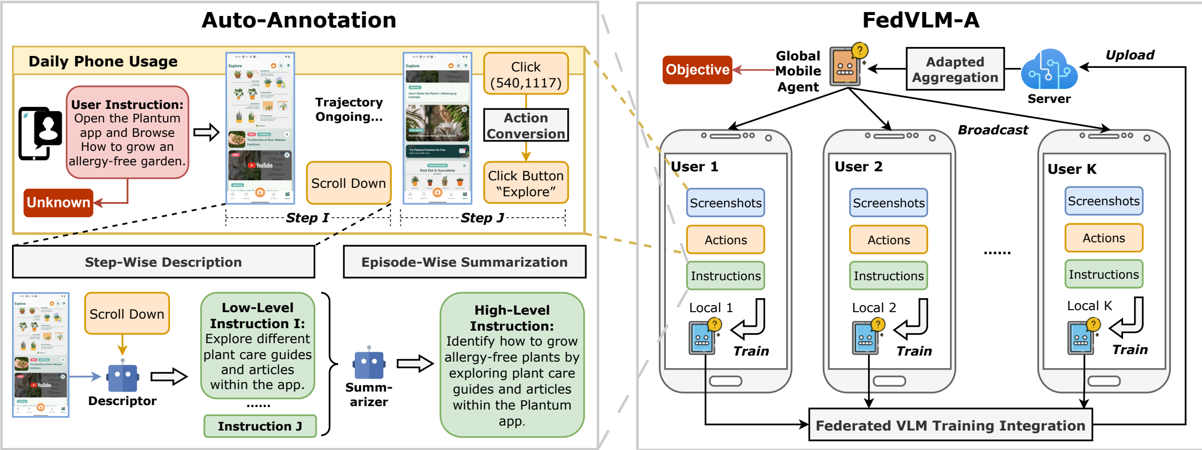


Figure 2: System overview of *MobileA3gent*. During individual users' daily phone usage, Auto-Annotation automatically constructs training data through step-wise description and episode-wide summarization. Each user then participates in FedVLM-A through our training integration. By applying adapted global aggregation, we obtain the target mobile agent with enhanced capabilities.

action. For the example in Figure 2, the atomic intent of `Scroll Down` on the browsing page is to "explore more articles on plant care". When combined with rule-based action conversion, the step-wise description allows the model to focus on localized context at each interaction, leading to more accurate and interpretable low-level directives. This step-by-step procedure also ensures that the information is more finely processed, facilitating better high-level summarization in subsequent stages Details of our prompt templates can be found in Appendix E.6.

Episode-Wise Intention Summarization. This stage generates high-level instructions by summarizing the low-level instructions from all steps. The novelty lies in providing global context enriched with step-wise details, enabling the annotation model to effectively extract user intention. To provide global visual context for the annotation model $\mathcal{M}_a$, referred to as the *Summarizer*, we concatenate all relevant screenshots into a single image s_c, arranged in chronological order. Note that this approach (1) allows *Summarizer* to develop a comprehensive understanding of the entire task sequence, and (2) eliminates the need for multiple inferences by performing inference only once. Finally, we compile the concatenated screenshot s_c and the list of low-level instructions $\{\mathcal{T}_i^{\text{low}}\}_{i=1}^{n}$ into a single prompt and feed it into $\mathcal{M}_a$ to summarize the user's overall intention $\mathcal{T}^{\text{high}}$ as:

$$\text{Summarizer} : \langle s_c, \{\mathcal{T}_i^{\text{low}}\}_{i=1}^{n} \rangle \xrightarrow{\mathcal{M}_a} \mathcal{T}^{\text{high}}. \quad (2)$$

Since users give no explicit commands, $\mathcal{T}^{\text{high}}$ simulates what they would convey if asking an agent to perform the same task. Combined with above mentioned techniques, episode-wise summarization produces high-quality instructions comparable to human annotated data, all while exclusively using locally deployed VLMs, thereby substantially reducing costs.

3.2 FedVLM-A: Federated Training of VLM-Based Mobile Agents with Adapted Global Aggregation

To facilitate training mobile agents on distributed data without comprising privacy, we propose FedVLM-A, a novel collaborative framework which pioneers the integration of federated learning with VLMs and improve performance in heterogeneous scenarios with Adapted Aggregation.

Integrating VLM Training. We build upon the highly-starred training framework, ms-swift (Zhao et al., 2024), and successfully extend it to support federated VLM training. We ensure the algorithmic correctness by following the implementation of federated training frameworks for Large Language Models (LLMs) (Ye et al., 2024). To enhance training efficiency and better accommodate user-side resource constraints, we incorporate Low-Rank Adaptation (LoRA) (Hu et al., 2021). In our federated setting, K clients (users) collaborate with a central server to train a global VLM without directly sharing private data. At each communication round l, the server broadcasts the global model $\mathcal{M}^{(l)}$ to all participating clients $u_k \in \mathcal{S}^l$, who initialize their local models accordingly: $\mathcal{M}_k^{(l,r+1)} := \mathcal{M}^{(l)}$, where $\mathcal{M}_k^{(l,0)}$ denotes the local model at the l-th round and 0-th training iteration. Each client u_k then conducts multiple iterations of stochastic gradient descent (SGD) updates on its local dataset $\mathcal{D}_k$. At each iteration r, with learning rate η, the local model is updated as:

$$\mathcal{M}_k^{(l,r+1)} = \mathcal{M}_k^{(l,r)} - \eta \nabla \ell(\mathcal{M}_k^{(l,\tau_k)}; \mathcal{T}, s, a)\,, \quad (3)$$

where $\ell(.)$ represents the computed loss based on a data sample $\langle \mathcal{T}, s, a \rangle$.

Adapted Global Aggregation. In this stage, the server updates global model by aggregating local models, which is subsequently broadcast to available clients for the next round. Our innovation lies in adapting the aggregation strategy to accommodate the two-level structure of datasets used for training mobile agents, encompassing both step-level variations and episode-level distributions. Traditional FL methods use the sample number of client as the aggregation weight. This insight has been proven successful over the past several years (Li et al., 2019, 2023). However, prior aggregation methods, such as FedAvgM and FedYogi (McMahan et al., 2017; Hsu et al., 2019; Reddi et al., 2020), which perform well on tasks such as image classification, overlook the two-level distribution discussed in Section B. These methods treat all samples equally, regardless of whether they originate from the same episode or not, thereby ignoring structural dependencies.

To address this limitation, we propose a novel aggregation technique adapting to the new scenario of MobileA3gent. Within federated training of mobile agents, the data samples can be measured by both step count n_k and episode count n_k^{epi}. n_k^{epi} is as well as, or even more important as it indicates how many tasks the agent has learned on. As n_k^{epi} and n_k are measured in different scales, we empirically set a hyper-parameter λ to align them, which is calculated around the average step length of all episodes. Then we redefine the sample count as n_k^* and reformulate the aggregation weight based on our adapted sample count n_k^*; that is:

$$n_k^* := \lambda n_k^{epi} + n_k\,; \qquad \omega_k = \frac{n_k^*}{\sum_{k \in \mathcal{S}^l} n_k^*}\,, \quad (4)$$

where ω_k denotes the weight for client u_k and $\mathcal{S}^l$ is the sampled participating clients. This design smoothly improves upon traditional aggregation and inherits its convergence property. When $\lambda = 0$, it degrades to normal aggregation. Finally, the global model $\mathcal{M}^{(l)}$ is adaptively aggregated as:

$$\mathcal{M}^{(l+1)} := \sum\nolimits_{k \in \mathcal{S}^l} \omega_k \mathcal{M}_k^{(l)}. \quad (5)$$

The adapted aggregation in FedVLM-A balances both episode and step counts, achieving a better utilization of decentralized data from heterogeneous users.

Table 1: Comparing privacy protection against risks. FedVLM-A offers strongest protection by addressing all three identified concerns. In contrast, *API-Based Agent* directly transmits user data, while *DistRL** stores all data centrally for training.

Privacy Protection	Eavesdrop	Abuse	Exposure
API-Based Agent	✗	✗	✗
DistRL*	✓	✗	✗
MobileA3gent	✓	✓	✓

Privacy Analysis. FedVLM-A preserves privacy by keeping original user data, which may contain sensitive information, on users' local devices without transmitting. Through local data retention, we successfully mitigate the following privacy risks, shown in Table 1: (1) *Eavesdropping Attack*: transmitting models instead of data prevents sensitive data from being intercepted during transmission; (2) *Data Abuse*: we reduce the risk of user data being exploited by data collectors for unintended purposes. (3) *Peer Exposure*: we eliminate the possibility of user data being accessed by other participants, as data is not directly shared between peers.

4 Experiments (More in Appendix C)

4.1 Basic Setups (*More Details in Appendix E*)

Models, Datasets & Benchmarks. The base model for most experiments is Qwen2-VL-Instruct-7B (Wang et al., 2024c). We also compare results with 10+ representative models, e.g. InternVL2 (Chen et al., 2024b) in Section 4.5. We select totally three offline agent benchmarks: Android-Control (Li et al., 2024a), Android in the Wild (AitW) (Rawles et al., 2023) and GUI Odyssey (Lu et al., 2024b). These datasets are collected by crowdsourcing and serve well as a simulation of real-world mobile data. Additionally, we employ AndroidWorld (Rawles et al., 2024), a challenging online benchmark running on Android emulators.

Metrics. Following previous works (Wu et al., 2024; Sun et al., 2024; Qin et al., 2025), we utilize three commonly used metrics for GUI agents that assess the accuracy of action type prediction, coordinate prediction, and step success rate, denoted as *Type*, *Ground*, and *SR*, respectively. We assess data quality by measuring the similarity between our generated instructions and the ground truth from the original datasets. Metric details are presented in Section E.3.

Table 2: Multi-dimensional comparison of MobileA3gent with other approaches. With 1% overall cost, MobileA3gent even surpasses the centralized human-annotated data. * We adjust DistRL to our user-centric setup. Anno. Cost refers to annotation cost in terms of cents (¢). Colors indicate preferable, moderate and concerning outcomes. Baseline details are explained in Appendix E.5.

Methodology	**AndroidControl-High**			**AndroidControl-Low**			**Anno. Cost**	**Privacy Protect**	**Scalability**
	Type	Ground	SR	Type	Ground	SR			
Prompting using Open-Ended & Closed-Ended Models									
OS-Atlas-7B (Wu et al., 2024)	57.44	54.90	29.83	73.00	73.37	50.94	-	✓	No
GPT-4o (OpenAI, 2023)	66.17	3.38	16.69	87.03	6.06	31.15	-	✗	
Finetuning on Human-Annotated Data									
Central-Human (Li et al., 2024a)	74.41	**53.75**	50.97	97.02	74.66	80.40	10880	✓	Very Low
FedL/VLM (Ye et al., 2024)	68.55	36.90	39.79	95.38	56.30	69.00			
Finetuning on Synthetic Data									
OS-Genesis (Sun et al., 2024)	66.15	-	44.54	90.72	-	74.17	$\approx 10^3$	✓	Limited
Finetuning on Auto-Annotated User data									
DistRL* (Wang et al., 2024d)	73.62	51.14	48.58	96.42	75.13	80.18	**152.92**	✗	Very High
MobileA3gent	**74.66**	53.05	**57.24**	**97.17**	**76.98**	**81.52**		✓	

4.2 Overall Evaluation of MobileA3gent

Baselines. To collect data and train mobile GUI agents, we compare *MobileA3gent* against the following baselines: (1) *Central-Human* (Li et al., 2024a), the conventional approach that relies on human annotation and centralized training on a server. (2) *FedLLM/VLM* (Ye et al., 2024), which differs from *Central-Human* by training in a distributed manner across client devices. (3) *OS-Genesis* (Sun et al., 2024), which automates synthetic data generation to reduce human effort. (4) *DistRL** (Wang et al., 2024d), an adapted version of the original method that first collects decentralized user data and then performs centralized training. During federated training, we randomly select 30% of clients in each round to mimic real-world scenarios where users are intermittently offline (Jiang et al., 2024). We evaluate the models at round 30, which corresponds to an expected cumulative client participation of 90%. Notably, the federated methods undergo fewer training iterations compared to centralized ones. We also provide prompt-based baselines, using locally deployed models or closed-ended models accessed via APIs, for reference.

Results & Analysis. As demonstrated in Table 2, we evaluate from four dimensions: *Performance*, *Efficiency*, *Privacy*, and *Scalability*, and summarize the following key findings: (1) **Comparable performance to *Central-Human*.** As the number of participating clients and the data volume increase, the performance of the collaboratively trained global model via MobileA3gent improves accordingly. Once the participation exceeds a certain threshold, users can obtain a highly capable mobile agent, comparable to or even surpassing *Central-Human* at minimal costs. (2) **Most efficient by leveraging daily phone usage.** The per-client annotation cost remains nearly negligible compared to *Central-Human*. Although *OS-Genesis* also aims to reduce human labor, it first generates synthetic instructions and then collects trajectories by employing GPT-4o to perform tasks in simulators, which still incurs medium-level costs. In contrast, we directly collect trajectories from users' daily phone usage by merely recording interactions, offering the most cost-saving approach for constructing GUI agent datasets. (3) ***MobileA3gent* substantially reduces privacy risks,** by keeping data on local devices. The privacy protection level is comparable to that of locally deployed agents, while achieving significantly higher performance. (4) **Promising scalability based on worldwide users.** As shown in Figure 7, the mobile user base is massive and continually expanding, which enables *MobileA3gent* to achieve much greater scalability compared to other approaches.

4.3 Data Quality and Training Evaluation of Auto-Annotation

Offline Benchmark. As shown in Table 3 and Table 5, we summarize the following key findings, (1) **Match or surpass *Human-Annotation*.** Our method achieves performance comparable to human annotation when trained on datasets of equal size. Notably, as the data scale increases, our method surpasses human annotation, highlighting the effectiveness of MobileA3gent and its strong

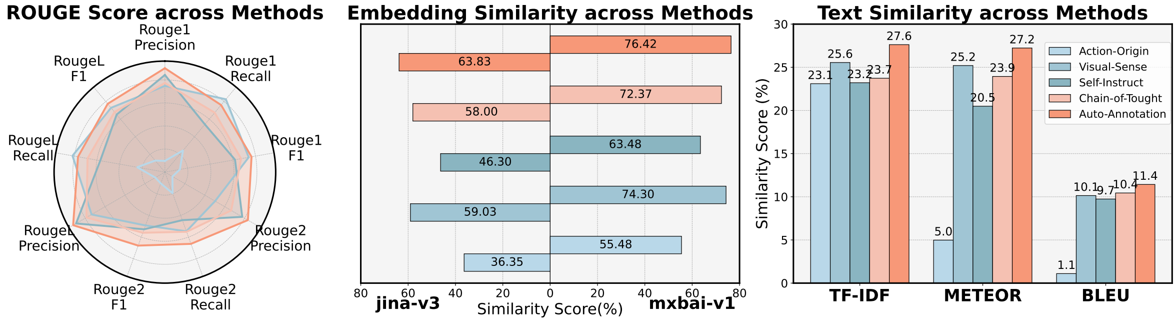


Figure 3: Data quality evaluation across comprehensive metrics. *Auto-Annotation* outperforms all other baselines and achieve comparable quality to *Human-Annotation* with a nearly 80% similarity.

potential for real-world deployment. (2) **Across multiple datasets, our approach consistently outperforms all annotation baselines**, underscoring the robustness and general effectiveness of *Auto-Annotation*. (3) **Drastic cost reductions with minimal accuracy loss.** Combined with the cost statistics in Table 8, By leveraging improved backends such as vLLM, we achieve up to a 99.9% cost reduction with less than a 2% decrease in high-level accuracy. Importantly, even as the dataset size scales up, the cost remains negligible compared to human labor.

Online Benchmark. We further evaluate our approach on the online benchmark AndroidWorld. As shown in Figure 4: (1) *Auto-Annotation* achieves **the best overall performance.** (2) Despite being trained solely on the AndroidControl dataset, the models are able to successfully complete online tasks in a previously unseen environment. This result demonstrates that agents trained with our framework possess **strong generalization capabilities** across unseen tasks and applications. Additional evaluations of generalization performance are provided in Appendix C.4 with Table 6 and 7.

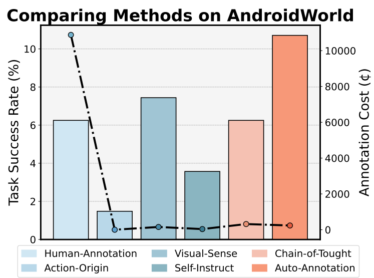


Figure 4: Performance and annotation cost trade-off on AndroidWorld.

Data Quality. As shown in Figure 3, (1) *Auto-Annotation* exhibits the **best performance across both text-based and embedding-based metrics**, providing strong evidence for the effectiveness of our hierarchical method. (2) A similarity score of nearly 80% to ground truth further demonstrates the **practical utility of generated instructions** on mobile devices, indicating their potential as a viable substitute for human-written ones. (3) *Visual-Sense* delivers competitive data quality using primarily visual signals, suggesting that **even stronger results may be achieved** by integrating *Auto-Annotation* with enhanced visual understanding.

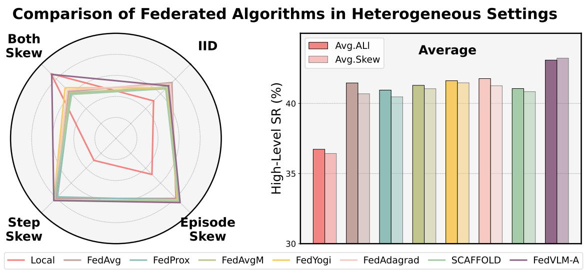
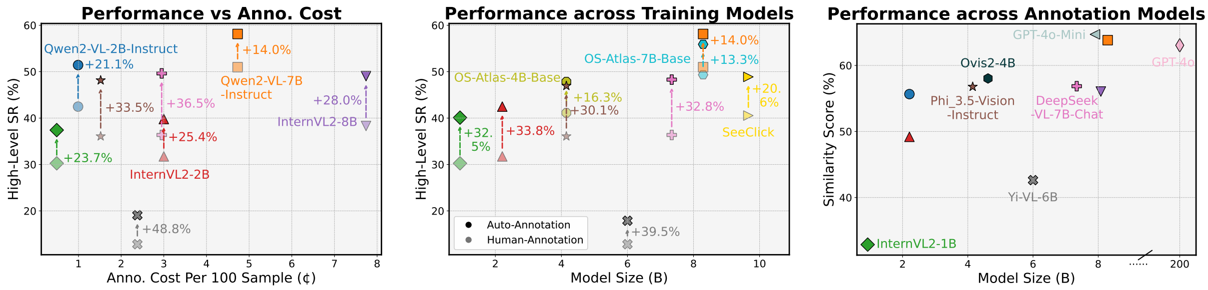


Figure 5: Comparison between FedVLM-A and 7 baselines on non-IID splits of AndroidControl. FedVLM-A achieves SOTA performance on average. Transparent bars indicate average scores over skewed scenarios only.

4.4 Training Evaluation of FedVLM-A

Baselines & Splits. We further conduct experiments under non-IID settings to verify the performance of FedVLM-A and investigate the heterogeneity issue formulated in Section B. We include seven representative FL baselines, such as FedProx (Li et al., 2020), FedYogi (Reddi et al., 2020). To eliminate any potential influence from *Auto-Annotation*, we use the original dataset in this section. Specifically, we sample 1,000 episodes from AndroidControl with uniformly distributed step lengths and create four distinct splits to simulate diverse distribution scenarios. Both the *Step Skew* and *IID* splits assign 100 episodes to each client. In the *Step Skew* scenario, clients have an equal number of episodes but varying numbers of steps, whereas in *Episode Skew*, the opposite holds. The *Both Skew* scenario features skewed values for both levels. For baseline *Local*, we evaluate using the 0-th client, which, in certain subsets (e.g., *Both Skew*), undergoes a number of iterations comparable to FL baselines.

Results. Figure 5 presents the radar chart of



(a) *X+X High-Level SR* (b) *Qwen+X High-Level SR* (c) *X+Qwen* `jina-v3` *Embed*

Figure 6: Ablation study on various base models. *x+y* indicates using model *x* for annotation and model *y* for training. *Qwen* refers to the consistent use of Qwen2-VL-7B-Instruct for fair comparison. *X* denotes a specific model. Arrows denote the relative improvement of *Auto-Annotation* over *Human-Annotation*. For the *X+Qwen* setting, we report embedding similarity scores to better capture differences. More comprehensive results are shown in Appendix Figure 9.

all baselines across the four splits, along with the average scores for all scenarios and for the three non-IID subsets. The results reveal the following: (1) Non-IID distributions negatively impact the performance of the global model, underscoring that **data heterogeneity is of critical importance in training distributed mobile agents**. (2) FedVLM-A with adapted aggregation achieves robust performance under non-IID settings, **outperforming all other baselines by at least 5%** in relative improvement. (3) Federated training significantly outperforms local training, validating the benefit of multi-user collaboration. (4) Overall, the results **confirm the existence of the two-level heterogeneity** highlighted in Section B, posing a new challenge for the federated learning community.

4.5 Ablation Experiments on Various Models

Setups. We conduct ablation experiments to assess the performance, annotation cost, and time requirements of different base models within *MobileA3gent*. Three configurations are evaluated by varying the choice of annotation and training models, where a combination *x+y* represents using model *x* for annotation and model *y* for training mobile GUI agents. Our model suite includes conversational VLMs such as Phi_3.5 (Abdin et al., 2024), grounding-oriented base models like SeeClick (Cheng et al., 2024) and widely adopted API-based models including GPT-4o/-Mini (OpenAI, 2023). In the plots, icons with light transparency denote models tuned using human annotations, whereas solid icons represent models using *Auto-Annotation*. The horizontal axis reflects annotation cost, measured via the *Pt* backend when applicable, or approximated by model size otherwise. For human-labeled icons, whose actual costs are prohibitively high, we use the same cost numbers for visualization purposes.

Results. As shown in Figure 6, 9 and Table 8, different models exhibit varying trade-offs between performance and cost. We conclude the following observations: (1) **Across all base models, our method achieves consistent improvement** over human-annotated baselines with significant cost savings. (2) The choice of annotation and training models introduces flexible performance–cost trade-offs, allowing practitioners to **tailor configurations to specific deployment constraints**. (3) Within a given VLM family, an increase in parameter numbers generally correlates with higher performance and greater computational demand. While across model types, this correlation does not always hold-e.g., Yi-VL-6B incurs lower costs and performs worse than InterVL2-2B, despite having more parameters. (4) **Qwen2-VL-2B-Instruct (blue circles) achieves the best balance between performance and annotation cost**, making it the most cost-effective option in our study.

5 Conclusion

To overcome the scalability and efficiency limitations of traditional mobile agent paradigm, we emphasize the necessity of transitioning from centralized to distributed user-centric data collection, and from human to automatic annotation. To achieve this, we propose MobileA3gent, a framework that collaboratively trains mobile GUI agents using self-sourced data from diverse users. Specifically, we introduce Auto-Annotation, an efficient approach for generating high-quality datasets from routine phone usage at minimal cost. Additionally, we present FedVLM-A, a federated VLM training framework with adapted global aggregation to handle mobile data heterogeneity. Extensive experiments on four benchmarks with 10+ models and metrics validate the effectiveness of MobileA3gent. The promising results highlight the scalability and practicality of our user-centric paradigm, offering a privacy-preserving and cost-efficient solution for training large-scale mobile agent.

Limitations

Despite the novelty and promising results of our work, potential limitations remain: (1) Due to device capacity, we are currently unable to conduct experiments on actual user mobile phones, as most mobile phone devices lack the necessary resources to hold mainstream models. However, an increasing number of studies are focusing on developing lightweight models specifically designed for mobile environments (Christianos et al., 2024; Papoudakis et al., 2025). MobileA3gent is model-agnostic and can seamlessly incorporate these smaller, more efficient VLMs, thereby facilitating practical deployment in resource-constrained settings. Also, as shown in Figure 9, we compare models of varying sizes. The results indicate that even compact models—such as InternVL2-1B and Qwen2-2B—can achieve competitive performance with as few as 1,000 training episodes. This demonstrates the scalability and effectiveness of our framework across different model sizes and architectures. While larger models like Qwen2-VL-7B-Instruct demand more computational resources, the overall annotation cost remains substantially lower than manual labeling, making our approach cost-effective even at scale. Although real-device experiments remain future work, our findings validate the effectiveness of the framework in resource-constrained settings.

References

2023. Gpt-4v(ision) system card.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. 2024. DigiRL: Training In-The-Wild Device-Control Agents with Autonomous Reinforcement Learning. *Preprint*, arXiv:2406.11896.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Omri Berkovitch, Sapir Caduri, Noam Kahlon, Anatoly Efros, Avi Caciularu, and Ido Dagan. 2024. Identifying user goals from ui trajectories. *arXiv preprint arXiv:2406.14314*.

Simone Caldarella, Massimiliano Mancini, Elisa Ricci, and Rahaf Aljundi. 2024. The phantom menace: Unmasking privacy leakages in vision-language models. *Preprint*, arXiv:2408.01228.

Yuxiang Chai, Siyuan Huang, Yazhe Niu, Han Xiao, Liang Liu, Dingyu Zhang, Peng Gao, Shuai Ren, and Hongsheng Li. 2024. AMEX: Android Multi-annotation Expo Dataset for Mobile GUI Agents. *Preprint*, arXiv:2407.17490.

Huiyao Chen, Yu Zhao, Zulong Chen, Mengjia Wang, Liangyue Li, Meishan Zhang, and Min Zhang. 2024a. Retrieval-style in-context learning for few-shot hierarchical text classification. *Transactions of the Association for Computational Linguistics*, 12:1214–1231.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. 2024. SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9313–9332, Bangkok, Thailand. Association for Computational Linguistics.

Filippos Christianos, Georgios Papoudakis, Thomas Coste, HAO Jianye, Jun Wang, and Kun Shao. 2024. Lightweight neural app control. In *NeurIPS 2024 Workshop on Open-World Agents*.

LMDeploy Contributors. 2023. Lmdeploy: A toolkit for compressing, deploying, and serving llm. `https://github.com/InternLM/lmdeploy`.

Shengwen Ding and Chenhui Hu. 2024. efedllm: Efficient llm inference based on federated learning. *arXiv preprint arXiv:2411.16003*.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. CogAgent: A Visual Language Model for GUI Agents. *Preprint*, arXiv:2312.08914.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. 2019. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. Lora: Low-rank adaptation of large language models. In *ICLR*.

Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. 2024. Déjà vu memorization in vision-language models. *Preprint*, arXiv:2402.02103.

Zhifeng Jiang, Wei Wang, and Ruichuan Chen. 2024. Dordis: Efficient federated learning with dropout-resilient differential privacy. In *Proceedings of the Nineteenth European Conference on Computer Systems*, pages 472–488.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, and 1 others. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U Stich, and Ananda Theertha Suresh. 2021. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, and Jie Tang. 2024. Autowebglm: A large language model-based web navigating agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5295—-5306.

Sunjae Lee, Junyoung Choi, Jungjae Lee, Munim Hasan Wasi, Hojun Choi, Steve Ko, Sangeun Oh, and Insik Shin. 2024. Mobilegpt: Augmenting llm with human-like app memory for mobile task automation. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom '24, page 1119–1133, New York, NY, USA. Association for Computing Machinery.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450.

Wei Li, William Bishop, Alice Li, Chris Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. 2024a. On the Effects of Data Scale on Computer Control Agents. *Preprint*, arXiv:2406.03679.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*.

Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. 2023. Revisiting weighted aggregation in federated learning with neural networks. *arXiv preprint arXiv:2302.10911*.

Zhangheng Li, Keen You, Haotian Zhang, Di Feng, Harsh Agrawal, Xiujun Li, Mohana Prasad Sathya Moorthy, Jeff Nichols, Yinfei Yang, and Zhe Gan. 2024b. Ferret-UI 2: Mastering Universal User Interface Understanding Across Platforms. *Preprint*, arXiv:2410.18967.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81.

Guangyi Liu, Pengxiang Zhao, Liang Liu, Zhiming Chen, Yuxiang Chai, Shuai Ren, Hao Wang, Shibo He, and Wenchao Meng. 2025a. Learnact: Few-shot mobile gui agent with a unified demonstration benchmark. *Preprint*, arXiv:2504.13805.

Guangyi Liu, Pengxiang Zhao, Liang Liu, Yaxuan Guo, Han Xiao, Weifeng Lin, Yuxiang Chai, Yue Han, Shuai Ren, Hao Wang, Xiaoyu Liang, Wenhao Wang, Tianze Wu, Linghao Li, Hao Wang, Guanjing Xiong, Yong Liu, and Hongsheng Li. 2025b. Llm-powered gui agents in phone automation: Surveying progress and prospects. *Preprint*, arXiv:2504.19838.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024a. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024b. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *Preprint*, arXiv:2406.08451.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024c. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*.

Yadong Lu, Jianwei Yang, Yelong Shen, and Ahmed Awadallah. 2024d. Omniparser for pure vision based gui agent. *Preprint*, arXiv:2408.00203.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

OpenAI. 2023. Gpt-4: A large-scale multimodal model. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Georgios Papoudakis, Thomas Coste, Zhihao Wu, Jianye Hao, Jun Wang, and Kun Shao. 2025. Appvlm: A lightweight vision language model for online app control. *Preprint*, arXiv:2502.06395.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. 2024. AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents. *Preprint*, arXiv:2405.14573.

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. 2023. Android in the Wild: A Large-Scale Dataset for Android Device Control. *Preprint*, arXiv:2307.10088.

Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan. 2020. Adaptive federated optimization. In *International Conference on Learning Representations*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Hongjin Su, Ruoxi Sun, Jinsung Yoon, Pengcheng Yin, Tao Yu, and Sercan Ö. Arık. 2025. Learn-by-interact: A data-centric framework for self-adaptive agents in realistic environments. *Preprint*, arXiv:2501.10893.

Qiushi Sun, Kanzhi Cheng, Zichen Ding, Chuanyang Jin, Yian Wang, Fangzhi Xu, Zhenyu Wu, Chengyou Jia, Liheng Chen, Zhoumianze Liu, Ben Kao, Guohao Li, Junxian He, Yu Qiao, and Zhiyong Wu. 2024. OS-Genesis: Automating GUI Agent Trajectory Construction via Reverse Task Synthesis. *Preprint*, arXiv:2412.19723.

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2Words: Automatic Mobile UI Summarization with Multimodal Learning. *Preprint*, arXiv:2108.03353.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. 2024a. Mobile-Agent: Autonomous Multi-Modal Mobile Device Agent with Visual Perception. *Preprint*, arXiv:2401.16158.

Luyuan Wang, Yongyu Deng, Yiwei Zha, Guodong Mao, Qinmin Wang, Tianchen Min, Wei Chen, and Shoufa Chen. 2024b. MobileAgentBench: An Efficient and User-Friendly Benchmark for Mobile LLM Agents. *Preprint*, arXiv:2406.08184.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Taiyi Wang, Zhihao Wu, Jianheng Liu, Jianye Hao, Jun Wang, and Kun Shao. 2024d. DistRL: An Asynchronous Distributed Reinforcement Learning Framework for On-Device Control Agents. *Preprint*, arXiv:2410.14803.

WenHao Wang, Xiaoyu Liang, Rui Ye, Jingyi Chai, Siheng Chen, and Yanfeng Wang. 2024e. KnowledgeSG: Privacy-preserving synthetic text generation with knowledge distillation from server. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7677–7695, Miami, Florida, USA. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and 1 others. 2024. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. 2024. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6137–6147.

Da Yu, Peter Kairouz, Sewoong Oh, and Zheng Xu. 2024. Privacy-Preserving Instructions for Aligning Large Language Models. *Preprint*, arxiv:2402.13659.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023. AppAgent: Multimodal Agents as Smartphone Users. *Preprint*, arXiv:2312.13771.

Guanhua Zhang, Mohamed Ahmed, Zhiming Hu, and Andreas Bulling. 2024a. Summact: Uncovering user intentions through interactive behaviour summarisation. *Preprint*, arXiv:2410.08356.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024b. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. 2024c. Android in the Zoo: Chain-of-Action-Thought for GUI Agents. *Preprint*, arXiv:2403.02713.

Jiwen Zhang, Yaqi Yu, Minghui Liao, Wentao Li, Jihao Wu, and Zhongyu Wei. 2024d. *UI-Hawk: Unleashing the Screen Stream Understanding for GUI Agents*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v(ision) is a generalist web agent, if grounded. *Preprint*, arXiv:2401.01614.

Chendi Zhou, Hao Tian, Hong Zhang, Jin Zhang, Mianxiong Dong, and Juncheng Jia. 2021. Tea-fed: time-efficient asynchronous federated learning for edge computing. In *Proceedings of the 18th ACM international conference on computing frontiers*, pages 30–37.

A Related Work

A.1 Development of Current Mobile GUI Agents

The advent of VLMs (Zhang et al., 2024b; Papoudakis et al., 2025; Christianos et al., 2024; Liu et al., 2025a) has marked a significant shift in phone automation, enabling more dynamic, context-aware, and sophisticated interactions with mobile devices (Liu et al., 2025b). Research on mobile agents has progressed through key milestones, with models becoming more proficient at interpreting multi-modal data, understanding user intent, and autonomously executing complex tasks. VLM-based mobile agents typically follow two approaches: (1) Prompt Engineering (Zhang et al., 2023; Lee et al., 2024; Lu et al., 2024d; Chen et al., 2024a), where pre-trained models are guided by carefully designed prompts, and (2) Training-Based Methods (Hong et al., 2023; Cheng et al., 2024), where VLMs are further optimized using large-scale mobile datasets. While training-based methods offer higher potential and generalizability by improving the VLM through fine-tuning, they require a large amount of training data, which can be very costly.

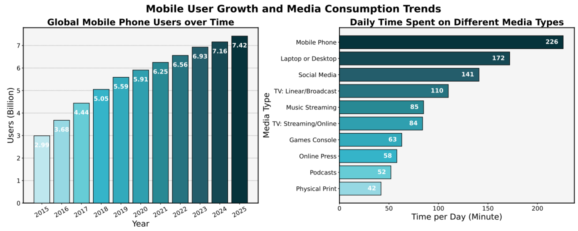


Figure 7: Trends in mobile user statistics. The increasing number of mobile users and their rising daily usage provide a sufficient data foundation for our approach.

A.2 Efforts in Building Datasets for Mobile GUI Agents

Acquiring training trajectories for mobile agents presents significant challenges. Existing approaches are often reliant on manual curation, making data collection both costly and inefficient. Some works have explored the possibility of automatically constructing datasets using VLMs or Application Programming Interfaces (APIs) (Wang et al., 2021; Lai et al., 2024). But these approaches either halfway to completing the datasets or depend on pre-defined tasks.

OS-Genesis (Sun et al., 2024), the most advanced in this area, proposes reverse task synthesis to eliminate the need for pre-defined instructions. However, this method still requires an agent to execute synthetic tasks in a simulated mobile environment, to obtain the corresponding screenshots and actions. This process does not guarantee the accuracy of executed actions, while also incurs additional computational and resource costs.

In contrast, we propose collecting real-world data from mobile users. This approach offers both (1) unlimited data scale, given the billions of mobile users worldwide, and (2) ground truth accuracy, as the data is directly generated through human execution.

B Detailed Problem Formulation

In this section, we first briefly elaborate on several key concepts, including the definition of mobile agents, to supplement Section 2.2. We then formulate our federated learning setup, with particular emphasis on the novel heterogeneity introduced by the inherent nature of mobile agent trajectories.

B.1 Supplemental Preliminaries

Step-Wise User Phone Usage. Typically, the process of one user interacting with a mobile device is formulated as follows. Initially, there is a screenshot of the interface, denoted as s_1. The user aims to complete a task, denoted as $\mathcal{T}$ in natural language, which requires n steps. Given any screenshot s_i, where $i \in [1, n]$, the user performs an action a_i, causing the interface to transition from s_i to s_{i+1}:

$$\text{User: } s_i \xrightarrow{a_i} s_{i+1} \,. \tag{6}$$

Once the last action a_n is performed, the interface reaches the final screenshot s_{n+1}, finishing the task $\mathcal{T}$ with $n + 1$ screenshots and n actions in total.

Functionality of Mobile Agents. The mobile agent, with the core being a VLM denoted as $\mathcal{M}_m$, simulates a human user in a step-wise process for task completion. It operates sequentially when applied to tasks. Given a natural language task $\mathcal{T}$ requiring n steps, at each step i, the primary function of the mobile agent is to predict the next action a_i^* required to complete $\mathcal{T}$, based on the current screenshot s_i and contextual information; that is:

$$\text{Mobile Agent: } \langle \mathcal{T}, s_i \rangle \xrightarrow{\mathcal{M}_m} a_i^* \,. \tag{7}$$

B.2 Federated Learning Setup

Reasons for Distributed Training. The duration of daily mobile phone usage is inherently limited for an individual, resulting in a relatively small dataset collected on a single user's device. This small-scale dataset constrains the performance of the mobile agent trained on it. Fortunately, with millions of mobile users worldwide, there exists a vast opportunity to incentivize users to collaborate and collectively train a mobile agent $\mathcal{M}$, using their combined data. Following the scaling law (Kaplan et al., 2020), leveraging multiple users' data enables virtually unlimited scalability and yields promising results. However, directly sharing or merging data generated from users' daily phone usage poses significant privacy risks. So the local data can only be utilized in a distributed manner.

Federated Learning. To address this challenge, we adopt federated learning, which effectively mitigates privacy concerns by keeping data on local devices, and develop a collaborative training framework FedVLM-A for mobile GUI agents. Given the local model $\mathcal{M}_k$ and a data sample (t, s, a) from $\mathcal{D}_k$, the objective of FedVLM-A is to optimize the global model $\mathcal{M}_m$ based on these local datasets; that is:

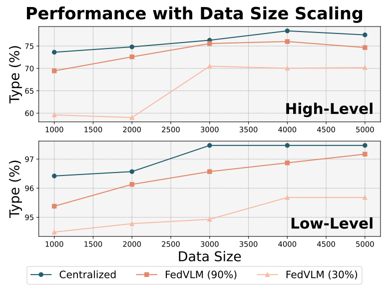
$$\min_{\mathcal{M}} F(\mathcal{M}) := \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{(t,s,a) \sim P^{(k)}_{\mathcal{T} \times \mathcal{S} \times \mathcal{A}}} \left[\ell(\mathcal{M}_k; t, s, a) \right] . \tag{8}$$

where $\ell : \mathcal{T} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ denotes the loss function, e.g. cross-entropy. $P^{(k)}_{\mathcal{T} \times \mathcal{S} \times \mathcal{A}}$ is the distribution over $\mathcal{T} \times \mathcal{S} \times \mathcal{A}$. $\mathcal{T}$, $\mathcal{S}$, and $\mathcal{A}$ represent task, screenshot, and action spaces, respectively. We assume that the distributions $P^{(k)}_{\mathcal{T} \times \mathcal{S} \times \mathcal{A}}$ differ across clients, which is a common scenario in FL. To the best of our knowledge, we are the first to apply federated learning into the training of mobile agents.

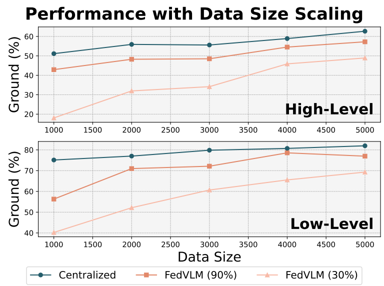
B.3 New Heterogeneity

Two-Level Distribution. Directly applying federated learning to mobile GUI agents introduces a new form of data heterogeneity. Unlike conventional FL scenarios where data are modeled as flat collections of independent samples, mobile interaction data inherently follow a hierarchical structure: they are collected *episode by episode*, with each episode consisting of sequential *steps* governed by a fixed task instruction. As a result, the underlying data distribution operates on two distinct levels. We refer to this structure as the **Two-Level Distribution**.

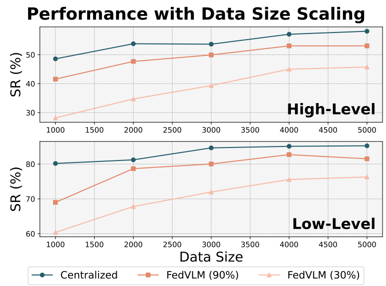
Level 1 (Intra-episode): Within episode j for user k, the task instruction $T^{(k,j)}$ leads to a sequence of $F^{(k,j)}$ steps. Since the task is constant within the j-th episode, the episode's data distribution simplifies to $P^{(k,j)}_{\mathcal{S} \times \mathcal{A}}$. *Level 2 (Inter-episode)*:



(a) *Type*



(b) *Ground*



(c) *SR*

Figure 8: Scaling law analysis on Android Control dataset with different training strategies. All models show a growing tendency with increased data size.

Across episodes, different tasks $T^{(k,j)}$ follow a distribution $P_T^{(k)}$. Thus, the overall data distribution on client k is defined as:

$$P_{\mathcal{T}\times\mathcal{S}\times\mathcal{A}}^{(k)} = \sum_{j=1}^{E^{(k)}} P_{\mathcal{S}\times\mathcal{A}}^{(k,j)} \cdot P_T^{(k)}\left(T^{(k,j)}\right), \quad (9)$$

where $E^{(k)}$ is the number of episodes on client k. This two-level distribution captures richer, hierarchical patterns and introduces more severe skew than the one-level heterogeneity in traditional federated learning.

Simplified Focus: Episode Length. To study the above mentioned new heterogeneity in a tractable way, we simplify $P_T^{(k)}$ to reflect only the distribution of episode lengths $F^{(k,j)}$. That is, we consider how many steps each episode contains, rather than the task content itself. Ignoring this episode length heterogeneity can lead to misleading assumptions and subsequent degraded performance. For example, two clients might each have 10 episodes of shopping-related tasks. However, if one client's episodes are short and concise while the other's are long and repetitive, their training data contribute differently to the global model. This results in biased updates despite seemingly equal numbers of episodes. Moreover, even if step-length distributions are balanced, clients may differ in total episode count or task diversity, still causing skewed contributions.

To address this, we propose an adapted aggregation strategy in Section 3.2 that explicitly accounts for heterogeneity in episode step length, going beyond standard sample-count-based methods in traditional federated learning.

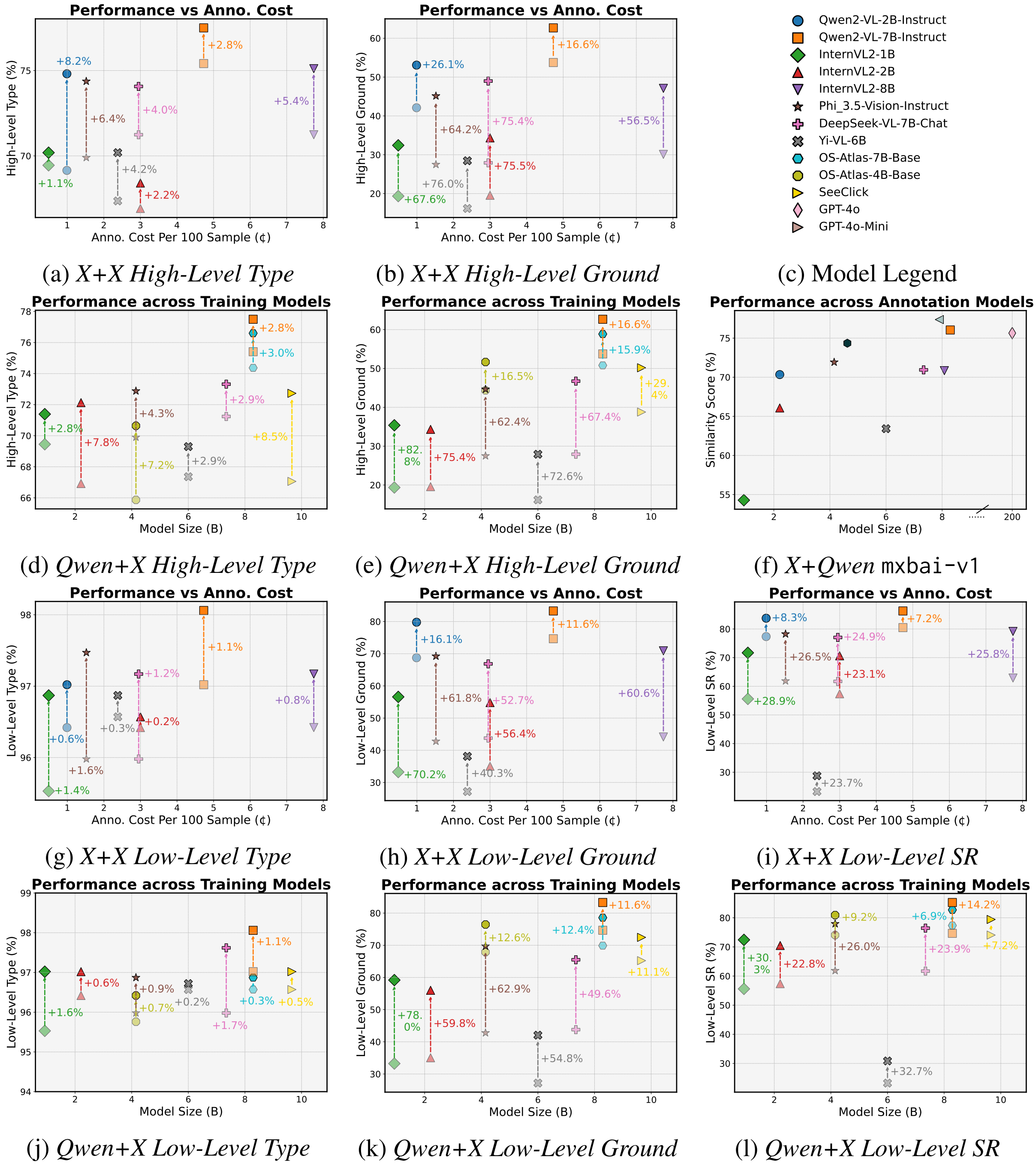
C Additional Experiments & Results

C.1 Continual Ablation Experiments on Various Models and Data Sizes

Setups. We further present our experiments, following Section 4.5. We conduct an ablation experiment on training data size to investigate whether the scaling law (Kaplan et al., 2020) holds for our automatically generated data. Using the Android-Control dataset, we train models that differ only in the size of their training data. For MobileA3gent, we fix the number of clients at 10 and test different participation rates, specifically 30% and 90%. We also provide more comprehensive ablation on different model combinations in Figure 9. Both high-level and low-level settings are evaluated with *Type*, *Ground* and *SR* metrics. Details about our model suite are provided in Section E.4.

Results. As shown in Figure 8, the performance of all tested models improves as the training data size increases, indicating that our generated data also follows the scaling law. We also observe a sharp performance increase when training from 100 and 1,000 episodes. No saturation is observed in our experiments; however it can be inferred that the performance of all models grows more slowly once the data size reaches a certain threshold. Moreover, when comparing high-level and low-level training settings, the latter converges faster, due to its simplicity and less room for improvement

From Figure 9, (1) we further demonstrate that *Auto-Annotation* is an effective method for annotating user instructions. The generated data exhibits strong utility and can be scaled up significantly at minimal cost compared to manual labeling. (2) Increasing the data scale benefits the *Ground* metric the most, as it captures the most critical aspect that VLMs need to learn from training data—the grounding ability. Specifically, *Auto-Annotation*



(a) *X+X High-Level Type* (b) *X+X High-Level Ground* (c) Model Legend

(d) *Qwen+X High-Level Type* (e) *Qwen+X High-Level Ground* (f) *X+Qwen* `mxbai-v1`

(g) *X+X Low-Level Type* (h) *X+X Low-Level Ground* (i) *X+X Low-Level SR*

(j) *Qwen+X Low-Level Type* (k) *Qwen+X Low-Level Ground* (l) *Qwen+X Low-Level SR*

Figure 9: Comprehensive visualization of different base models evaluated across multiple metrics and training configurations.

Table 3: Training evaluation on AndroidControl and GUI Odyssey with more results in Table 4. We compare our method against various baselines. *Auto-Annotation* achieves superior results across all methods, and shows substantial improvement over *Human-Annotation* with significant cost savings.

Methodology	AndroidControl-High			AndroidControl-Low			GUI Odyssey		
	Type	Ground	SR	Type	Ground	SR	Type	Ground	SR
Qwen2-VL-7B-Instruct	28.61	0.00	1.94	71.09	2.19	6.41	2.87	2.94	0.51
GPT-4o	66.17	3.38	16.69	87.03	6.06	31.15	37.50	14.17	5.36
OS-Atlas-7B-Pro (Wu et al., 2024)	57.44	54.90	29.83	73.00	73.37	50.94	60.42	39.74	26.96
Human-Annotation	75.41	53.75	50.97	97.02	74.66	80.48	78.85	64.92	55.22
Action-Origin	65.28	3.18	9.84	94.19	3.56	26.68	62.36	13.32	14.33
Visual-Sense (Zhang et al., 2024a)	77.50	61.13	57.37	97.47	81.42	85.54	81.53	67.66	59.49
Self-Instruct (Wang et al., 2023)	75.86	57.28	53.95	97.47	81.97	85.25	**82.80**	60.27	55.16
Chain-of-Thought	**77.94**	56.96	55.89	97.17	83.20	85.39	74.37	50.80	49.33
Auto-Annotation	77.50	**62.67**	**58.12**	**98.06**	**83.29**	**86.29**	81.72	**69.51**	**60.57**

achieves up to an 82.8% improvement over *Human-Annotation* for InternVL2-1B.

C.2 Auto-Annotation with Different Data Sizes

We present detailed experiments comparing *Auto-Annotation* with various baselines under two distinct data sizes. *Human-Annotation* serves as the upper bound.

Comparison with Human-Annotation. When the training data size is equal, our method achieves comparable performance on many evaluation metrics, with less than a 2% drop—for example, *SR* on both AndroidControl-High and AndroidControl-Low—while reducing annotation costs by over 99%. Moreover, as the data size scales up, our method surpasses *Human-Annotation* with ease, while still maintaining minimal cost.

Comparison with Other Baselines. *Auto-Annotation* maintain consistent superiority other baselines across all metrics and data sizes, making it the most effective choice for annotating user instructions.

C.3 Auto-Annotation on AitW Dataset

Setups. The AitW dataset consists of five subsets: General, Install, GoogleApps, Single, and WebShopping. For each subset of AitW, we sample 1,000 episodes for training and 100 for evaluation. The overall performance is the average of the five subsets. For the validation metric, we omit validation samples that consist of click actions with no corresponding unit. These samples are too easy to predict in our setting and show no meaningful difference between different models. We only present high-level accuracy due to the absence of low-level instructions in the original dataset.

Results. As shown in Table 5, we can conclude the following: (1) Apart from AndroidControl and GUI Odyssey, our method still achieves comparable results with Human-Annotation and outperform it by a large margin as the data size increases. (2) Our method performs extremely well on the *Single* subset. We attribute this result to the short average step length for episodes in *Single*, which leads to more accurate reconstruction of the high-level instructions.

C.4 Out-of-Domain Evaluation with Generalization Analysis

Setups. To evaluate the performance of MobileA3gent in out-of-domain scenarios, we conduct two experiments on the AndroidControl and GUI Odyssey datasets. For AndroidControl, we randomly sample 100 episodes from each of the three unseen test splits: *App-Unseen*, *Task-Unseen*, and *Category-Unseen*, based on the dataset's sub-splits. The number 100 is chosen to match the test sample size used in Section 4.3. For GUI Odyssey, we similarly sample 100 episodes from each unseen test split: *App-Unseen*, *Task-Unseen*, and *Device-Unseen*. Note that the original GUI Odyssey datasets include overlapping samples across splits; therefore, we select test episodes that do not overlap with either the training samples or with each other.

Results. As shown in Tables 6 and 7, (1) mobile agents trained on our automatically generated data exhibit strong generalizability across various settings. The results demonstrate the effectiveness of our approach and further validate the utility of our auto-annotated data, which is derived solely from screenshots and actions. (2) Additionally, we observe that the *Category-Unseen* sub-

Table 4: In-depth evaluation of Auto-Annotation under equal data size on AndroidControl. In this setup, Human-Annotation serves as the upper bound due to its access to gold instructions. Auto-Annotation outperforms other baselines trained on model-annotated data and achieves comparable performance to Human-Annotation on several metrics-such as high-level *SR*-with drastic cost saving.

Methodology	AndroidControl-High			AndroidControl-Low			GUI Odyssey		
	Type	Ground	SR	Type	Ground	SR	Type	Ground	SR
Qwen2-VL-7B-Instruct	28.61	0.00	1.94	71.09	2.19	6.41	2.87	2.94	0.51
GPT-4o (OpenAI, 2023)	66.17	3.38	16.69	87.03	6.06	31.15	37.50	14.17	5.36
OS-Atlas-7B-Pro (Wu et al., 2024)	57.44	54.90	29.83	73.00	73.37	50.94	60.42	39.74	26.96
	Data Size = 5,000						**Data Size = 3,000**		
Human-Annotation (Li et al., 2024a)	79.14	66.56	61.70	97.62	81.47	85.99	84.39	75.63	67.01
Action-Origin	65.28	3.18	9.84	94.19	3.56	26.68	62.36	13.32	14.33
Visual-Sense (Zhang et al., 2024a)	<u>77.49</u>	<u>61.13</u>	<u>57.37</u>	<u>97.47</u>	81.42	<u>85.54</u>	81.53	<u>67.66</u>	<u>59.49</u>
Self-Instruct (Wang et al., 2023)	75.86	57.28	53.95	<u>97.47</u>	81.97	85.25	**82.80**	60.27	55.16
Chain-of-Thought	**77.94**	56.96	55.89	97.17	<u>83.20</u>	85.39	74.37	50.80	49.33
Auto-Annotation	<u>77.49</u>	**62.67**	**58.12**	**98.06**	**83.29**	**86.29**	<u>81.72</u>	**69.51**	**60.57**
	Data Size = 1,000								
Human-Annotation (Li et al., 2024a)	75.41	53.75	50.97	97.02	74.66	80.48	78.85	64.92	55.22
Action-Origin	65.28	2.85	10.58	90.61	1.14	28.46	54.52	5.38	7.52
Visual-Sense (Zhang et al., 2024a)	<u>73.62</u>	<u>51.14</u>	<u>48.58</u>	96.42	74.25	<u>80.18</u>	76.62	<u>54.43</u>	<u>46.50</u>
Self-Instruct (Wang et al., 2023)	72.43	48.99	47.54	96.87	72.40	78.69	<u>77.07</u>	51.33	45.22
Chain-of-Thought	72.58	48.48	47.24	<u>97.02</u>	<u>74.53</u>	<u>80.18</u>	76.56	53.40	46.37
Auto-Annotation	**74.22**	**52.44**	**49.48**	**97.47**	**75.13**	**80.48**	**77.58**	**59.74**	**50.64**

Table 5: Evaluation of Auto-Annotation across different subsets of AitW dataset. Our methods achieve consistent superior performance compared to Human-Annotation at a very low cost. -S denotes a simplified version which removes the step-wise description.

Methodology	Size	General	Install	GoogleApps	Single	WebShopping	Overall
Zero-Shot	-	15.90	5.20	15.08	28.38	11.41	15.19
Human-Annotation	1000	35.04	54.50	46.65	55.46	39.82	46.29
Auto-Annotation-S	1000	36.24	52.47	44.13	53.41	<u>40.34</u>	45.32
Auto-Annotation	1000	<u>36.92</u>	53.23	37.43	52.84	39.65	44.01
Auto-Annotation-S	5000	36.24	**59.19**	<u>47.21</u>	<u>62.45</u>	39.43	<u>48.90</u>
Auto-Annotation	5000	**37.26**	<u>57.29</u>	**47.49**	**72.05**	**45.14**	**51.85**

set yields relatively lower accuracy compared to other evaluation subsets, indicating a higher level of difficulty. (3) For GUI Odyssey, we note that *Human-Annotation* achieves relatively higher performance than in other experiments, suggesting that this dataset may pose greater challenges for generalization.

C.5 Efficiency Evaluation across Inference Backends

Setups. To further investigate whether the annotation cost using our method can be reduced and whether the memory requirements can be minimized with current efficient inference backends, such as vLLM (Kwon et al., 2023) and LMDeploy (Contributors, 2023), we conduct additional experiments to assess efficiency by computing model costs and memory usage on different backends. API-based costs are assessed using the OpenAI's library tiktoken [1] to count input and output tokens via Auto-Annotation. The price per million tokens is also included. Moreover, we approximate the API cost for the Qwen2-VL family by using the pricing of Qwen-VL-Plus, as the server does not provide APIs for Qwen2-VL-7B-Instruct or Qwen2-VL-2B-Instruct. For the InternVL2 family, since the model server offers free trial access, we denote the cost as "Free". Note: the annotation costs are computed using Auto-Annotation-S, which removes the step-wise process for fair comparison across models. For reference, using vLLM, *Auto-Annotation* incurs around 2 times the cost of *Auto-Annotation-S*.

Results. As shown in Table 8, models exhibit explicitly different behaviors across backends. In general, most models reduce costs when using efficient backends. For example, InternVL2-2B saves

[1] `https://github.com/openai/tiktoken`

Table 6: Out-of-domain evaluation on AndroidControl. We compare Auto-Annotation with baselines on three out-of-domain test subsplits. Our method achieve consistent improvement over Human-Annotation with minimal annotation cost.

Methodology	App-Unseen			Task-Unseen			Category-Unseen		
	Type	Ground	SR	Type	Ground	SR	Type	Ground	SR
Human-Annotation (Li et al., 2024a)	65.79	57.02	47.74	78.65	67.83	60.98	70.31	51.07	47.03
Action-Origin	56.48	5.92	9.90	64.21	0.99	12.75	60.16	1.88	7.81
Visual-Sense (Zhang et al., 2024a)	70.45	64.14	54.59	82.64	**69.76**	64.98	69.69	**61.54**	54.38
Self-Instruct (Wang et al., 2023)	**72.63**	64.06	**56.04**	**84.18**	67.12	64.06	69.84	59.45	53.75
Chain-of-Thought	71.03	58.33	51.97	82.64	68.42	64.21	71.09	59.22	55.47
Auto-Annotation	72.20	**65.00**	**56.04**	83.72	68.97	**65.13**	**72.97**	61.06	**56.25**

Table 7: Out-of-domain evaluation on GUI Odyssey. We compare Auto-Annotation with baselines on four evaluation subsets. Our method achieve consistent improvement over Human-Annotation with minimal annotation cost.

Methodology	App-Unseen			Task-Unseen			Device-Unseen		
	Type	Ground	SR	Type	Ground	SR	Type	Ground	SR
Human-Annotation (Li et al., 2024a)	**78.76**	59.23	51.85	76.74	61.89	49.29	79.96	61.42	53.02
Action-Origin	61.54	9.94	11.35	62.56	11.22	11.63	61.84	11.86	12.52
Visual-Sense (Zhang et al., 2024a)	77.49	54.40	48.47	75.32	60.64	47.88	81.50	63.36	55.98
Self-Instruct (Wang et al., 2023)	78.00	49.33	45.22	76.96	56.75	46.58	82.24	57.72	52.40
Chain-of-Thought	77.36	53.81	48.21	77.24	56.71	46.53	82.31	57.98	53.08
Auto-Annotation	77.87	**63.03**	**53.76**	**77.64**	**65.76**	**52.68**	**82.49**	**67.56**	**58.82**

annotation costs by more than half when leveraging LMDeploy. However, for smaller models, using an efficient backend does not necessarily lead to improvements. We attribute this to the fact that running vLLM on an RTX 4090 causes the model to occupy the entire GPU memory, which is 5 to 10 times the original memory usage of PyTorch. This increase in memory consumption does bring out improvement inference speed but fails to offset the additional memory demand. Since our annotation cost, as formulated in Equation E.3, considers both time and memory usage, the overall cost does not necessarily decrease. Additionally, APIs remain a viable option since they eliminate the need for local deployment, while offering highly competitive pricing. However, using APIs comes at the sacrifice of privacy as shown in Table 1.

C.6 Accuracy Comparison across Different Actions

Following the evaluation protocol described in Section 4.1, we compute the accuracy for each action type defined in the AndroidControl action space, as detailed in Table 9.

As illustrated in Figure 10, the accuracy varies significantly across different action types. Notably, the `COMPLETE`, `TYPE`, and `OPEN_APP` actions achieve relatively high accuracy. This can be attributed to the fact that these actions primarily depend on language understanding rather than visual grounding. Given that current VLMs are more proficient in handling language-based tasks, these actions are easier to infer correctly. In contrast, `NAVIGATE_BACK` and `WAIT` exhibit the lowest accuracy. We hypothesize that this is mainly due to their limited representation in the training set, as they constitute only a small portion of the total training data. Additionally, `NAVIGATE_BACK` often requires the model to correct previous errors or perform implicit reasoning based on prior steps, which is challenging for VLMs lacking explicit reasoning capabilities.

It is also worth noting that the *Type* metric differs fundamentally from the *SR* metric. The *Type* metric only requires correctly identifying the action type, without evaluating parameters such as coordinates or input content. In contrast, *SR* considers an action correct only if all its arguments are predicted accurately. This distinction is especially significant for coordinate-based actions like `CLICK`, which require precise location predictions to be considered successful under the *SR* metric. This additional complexity makes it more challenging for models to achieve high accuracy on such actions.

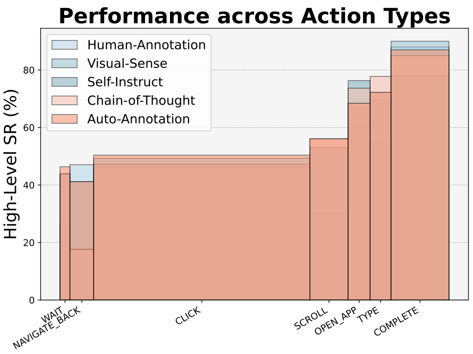
D Discussions and Future Directions

D.1 Discussions

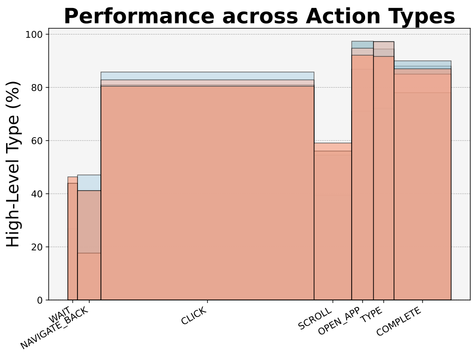
Analysis of Resources on a Mobile Device. To investigate the minimal resource requirements, we conduct additional experiments to determine

Table 8: Comparison of annotation costs per 1,000 samples, using different inference backends across various base models. Results demonstrate that employing efficient backends, such as vLLM and LMDeploy, can further reduce inference time and memory usage, ultimately lowering the annotation cost of our approach. The generation time and memory usage are averaged over three runs.

Annotation Model	**Annotation Cost (¢)**			**Generation Time (s)**		**Memory Usage (MB)**	
	PyTorch	vLLM or LMDeploy	API	PyTorch	vLLM or LMDeploy	PyTorch	vLLM or LMDeploy
Human	10880			56300		-	
GPT-4o-Mini	-	-	14.8	5061		-	
GPT-4o	-	-	247.92	6858		-	
Qwen2-VL-2B-Instruct	6.14	8.42	<21.15	1577	1180	12046	22083
Qwen2-VL-7B-Instruct	16.77	9.87	<21.15	2005	1374	25881	22224
InternVL2-1B	2.58	11.09	Free	2000	1662	3985	20645
InternVL2-2B	16.04	7.23	Free	1698	1038	29235	21548
InternVL2-8B	23.18	-	Free	2245	-	31960	-



(a) High-Level *SR*



(b) High-Level *Type*

Figure 10: High-level *SR* across different action types within the action space of AndroidControl. The width of the pillars corresponds to the number of data samples in the evaluation test set; thus, the area reflects the weighted average performance.

whether small VLMs or models based on APIs can achieve similar effectiveness. The results in Section 4.5 show that even an 1B VLM can deliver competitive performance. Models based on APIs can also be used, though they come with the risk of privacy leakage, which we leave for future work.

Real-World Applicability Analysis. We will address the real-world applicability three-folds. First, as shown in Section 4.4, each user only needs to provide a small amount of data, and not all of it is sensitive, resulting in minimal or no privacy risks. Second, the benefits far outweigh the costs and risks. We assume that the server incentivizes participation by offering free use of the global agent in exchange for access to user data. Users can gain access to a highly capable mobile agent that saves them both time and efforts. Finally, by incorporating federated learning, user data is processed locally, alleviating most privacy concerns.

D.2 Future Directions

As mentioned above, we have shown the promising results achieved by MobileA3gent. However, this is not the end as there are still emerging challenges and interesting directions that are worth exploring in this direction.

Privacy Preservation. Training on user data inevitably raises privacy concerns. While federated learning helps mitigate privacy leakage by keeping private data on the client side and transmitting only LoRA adapters, potential privacy issues remain. Models with substantial sizes are prone to memorization of their training data (Yu et al., 2024; Wang et al., 2024e). Similar to large LLMs, recent studies (Caldarella et al., 2024; Jayaraman et al., 2024) reveal that VLMs also inadvertently memorize and potentially expose sensitive information. Dejavu memorization (Jayaraman et al., 2024) proposes a novel measurement for memorization by quantifying the fraction of ground-truth objects in an image that can be predicted from its text description in a training image-text pair. Mobile agents rely on VLMs to perceive the interface and make decisions. Therefore, training directly on user data may lead to leakage of sensitive information. This issue can be addressed by implementing differential privacy (DP), which, however, remains underexplored in the context of VLMs and mobile agent training.

Efficiency. To collaboratively train a global mobile agent on distributed user data, each user needs to locally train a small-sized VLM and communicate with the central server. However, limited computation resources and communication channels on mobile devices may hinder the feasibility of deployment. With the recent advancement of LLMs and diffusion models and their integration into federated learning systems (Zhou et al., 2021), numerous approaches have been proposed to alleviate computational and communication overheads (Ding and Hu, 2024). On the other hand, the proliferation of smaller VLMs has significantly enhanced efficiency. For instance, AppVLM (Papoudakis et al., 2025) specifically targets app control tasks with a lightweight architecture, facilitating rapid and cost-efficient inference for real-time execution.

Reinforcement Learning. Although our current framework does not yet incorporate reinforcement learning, we identify it as a promising future direction. In a federated mobile agent setting, user feedback can serve as a critical reward signal, enabling agents to adjust their decision-making policies dynamically. Future work will need to tackle challenges inherent to integrating reinforcement learning into a federated environment, such as handling heterogeneous feedback, ensuring robust and stable learning under variable network conditions, and preserving user privacy. We believe that exploring these issues will pave the way for more adaptive and user-centric mobile agents, ultimately enhancing both their responsiveness and overall utility.

E Experimental Details

E.1 Benchmark Details

To provide a comprehensive evaluation, we select four widely used mobile agent benchmarks from prior works (Wu et al., 2024; Sun et al., 2024; Zhang et al., 2024d), covering both offline and online settings.

Offline Benchmarks. In offline benchmarks, agents are evaluated using static screenshots and instructions under a step-wise evaluation protocol in a fixed order. Notably, even if an agent fails at a prior step that would normally prevent it from reaching the current step, the current step is still included in the evaluation. Offline benchmarks are favored in the GUI agent community due to their ease of quantification and deployment. We employ three widely accepted benchmarks from Google[2] and OpenGVLab[3].

- **AndroidControl (AC)** (Li et al., 2024a), evaluates agents' planning and action-execution capabilities in mobile environments. This benchmark provides two task types: (1) high-level tasks, where the agent must autonomously plan and execute multi-step actions; and (2) low-level tasks, where the agent is required to execute pre-defined, human-annotated actions which is more specific, at each step. During low-level tasks, both a low-level instruction and its corresponding high-level instruction are included. We conduct experiments in both settings for a comprehensive assessment. A data example is provided in Figure 11 to further clarify the difference between high-level and low-level instructions.
- **Android in the Wild (AitW)** (Rawles et al., 2023), is a large-scale dataset annotated with instructional operations and screenshot-based icon detection, including element-level annotations generated using a pretrained IconNet. The AitW dataset comprises five subsets: General, Install, GoogleApps, Single, and WebShopping.
- **GUI Odyssey** (Lu et al., 2024b), focuses on cross-app navigation tasks in mobile environments, featuring an average of over 15 steps per task, which is notably longer than in AndroidControl. The tasks cover diverse navigation scenarios, and within each scenario, multiple instructions are generated based on predefined templates.

Online Benchmark. In contrast to offline benchmarks, online benchmarks prioritize realism and practical applicability. Agents are required to perform dynamic, interactive tasks in online simulation environments. And they continue attempting the task until reaching a predefined maximum step length. This setup may lead to some back-and-forth or repetitive behaviors as agents explore and recover from errors.

- **AndroidWorld (AW)** (Rawles et al., 2024), is

[2] https://github.com/google-research/google-research

[3] https://github.com/OpenGVLab

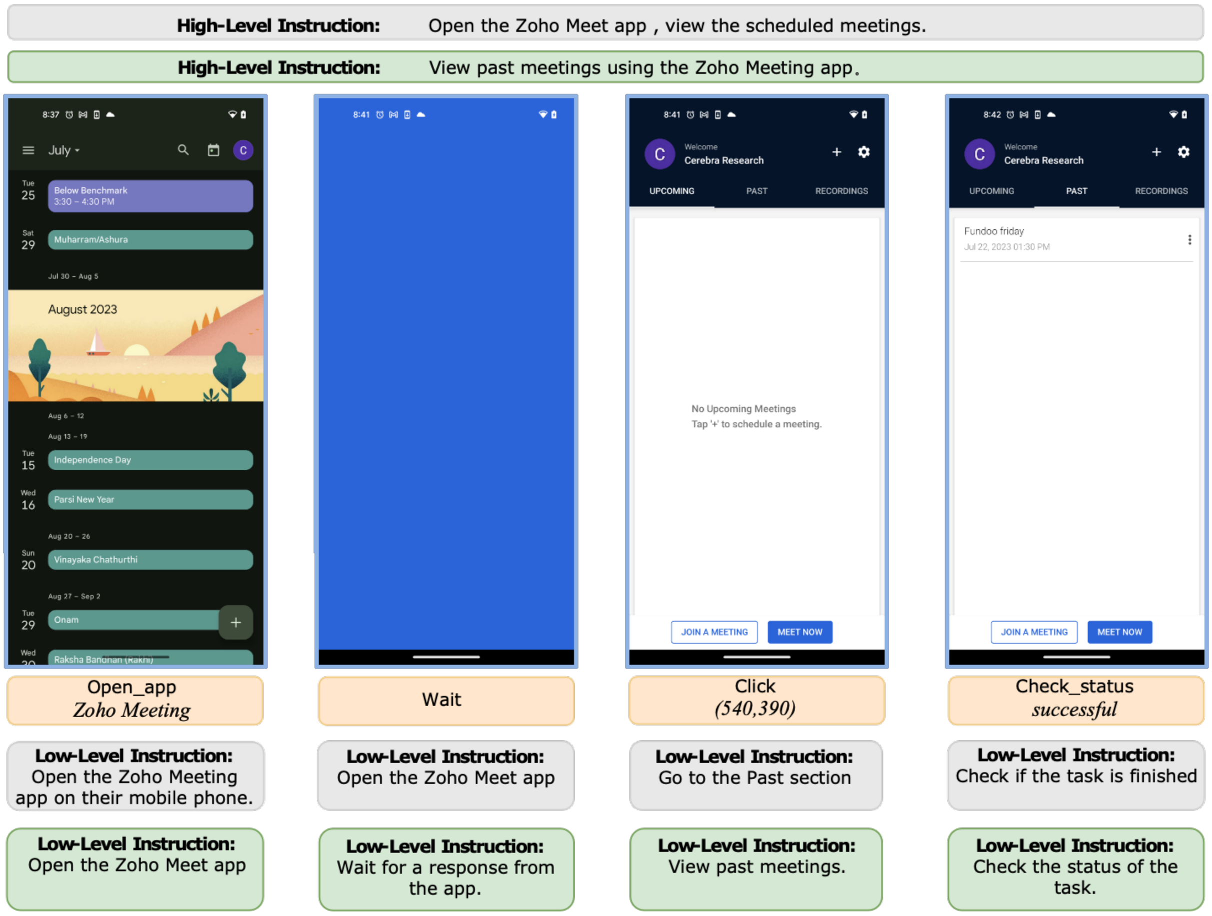


Figure 11: Episode example from Android Control dataset. The high-level task is *"Open the Zoho Meet app and view the scheduled meetings"*. Instructions in grey indicate ground truth from the original dataset, while those in green are predictions generated by Auto-Annotation. Our generated data sample achieves quality comparable to human-annotated ground truth.

an online environment designed for developing and benchmarking autonomous agents using a Pixel 6 phone simulator as the testbed. It comprises 116 tasks spanning 20 mobile apps, with dynamic task variations generated through randomized parameters. This dataset is particularly well-suited for evaluating agents' adaptability and planning abilities on mobile devices.

Our experimental setups for the offline datasets follow those in Wu et al. (2024), while the setups for the online benchmark adhere to the original implementation.

E.2 Data Details

Data Composition. To offer a clearer understanding of the structure of mobile training datasets and the composition of a data episode, we present a representative example in Figure 11. As shown, each episode consists of: (1) A high-level instruction, expressed as a natural language sentence describing the task to be accomplished; (2) A sequence of low-level instructions, detailing the fine-grained tasks required for the current screenshot; notably, such annotations are only available in the AndroidControl dataset; (3) A series of screenshots captured from the start to the end of the task; and (4) A corresponding list of actions, aligned with the number of screenshots, indicating what the user does to progress to the next screenshot. All actions belong to an action space containing 7-9 options.

Action Space. Considering the action space used in OS-Atlas and the original AndroidControl paper, we define nine action types for AndroidControl. Notably, two of these action types, `Navigate_Home` and `Long_Press`, appear only rarely. For GUI Odyssey, one more action type `Press_Recent` is defined as press the recent button to switch between different apps as most tasks are cross-app. For the AitW dataset, we define seven action types. The corresponding actions and their descriptions are provided in Tables 9 and 11,

Table 9: Action space for AndroidControl.

Action Type	Attribute	Description
		Basic Actions
CLICK	(x,y)	Click at a specific point on the screen using the coordinates.
TYPE	text	Type the text in the current input field or search bar.
SCROLL	direction	Scroll in a specific direction (one of 'up', 'down', 'left', or 'right').
		Custom Actions
LONG_PRESS	(x,y)	Long press at a specific point on the screen using the coordinates.
NAVIGATE_BACK	-	Return to the previous page or undo an action.
NAVIGATE_HOME	-	Return to the home page.
OPEN_APP	app_name	Open an app with the specified name.
WAIT	-	Pause for a moment before proceeding with the next action.
COMPLETE	-	Indicate that the task is finished.

Table 10: Action space for GUI Odyssey.

Action Type	Attribute	Description
		Basic Actions
CLICK	(x,y)	Click at a specific point on the screen using the coordinates.
TYPE	text	Type the text in the current input field or search bar.
SCROLL	direction	Scroll in a specific direction (one of 'up', 'down', 'left', or 'right').
		Custom Actions
LONG_PRESS	(x,y)	Long press at a specific point on the screen using the coordinates.
NAVIGATE_BACK	-	Return to the previous page or undo an action.
NAVIGATE_HOME	-	Return to the home page.
PRESS_RECENT	-	Press 'Recent' to switch between recently used applications.
WAIT	-	Pause for a moment before proceeding with the next action.
COMPLETE	-	Indicate that the task is finished.

with any additional parameters indicated as *Target*. In AitW, we decompose the original `Press` action into three distinct actions: `Navigate_Home`, `Navigate_Back`, and `Press_Enter`, aligning the action space with that of AndroidControl. Additionally, we derive the `Scroll` action from the original dual-point action.

Splits. Regarding training and testing splits, for AndroidControl, we adopt the original splits provided in the paper[4]. Specifically, we sample 5,000 episodes for training and 100 episodes for each test subsplit, i.e., *IID*, *App-Unseen*, *Task-Unseen*, and *Category-Unseen*. Unless otherwise specified, our results (except for the generalization experiments reported in Section C.4) are evaluated based on the *IID* subsplit. For each subset of AitW, we sample 1,000 episodes for training and 100 for evaluation.

E.3 Metrics Details

Efficiency Metrics. We also compare the annotation costs across methods to assess efficiency. The cost of a single human-annotated sample is derived from a Refuel-AI technical report. The costs for model-annotated samples are estimated by calculating the average GPU usage during generation, given by: Anno. Cost $= \left(\frac{\text{Price}}{3600}\right) \times \text{Time} \times \frac{\text{Memory}_{\text{Use}}}{\text{Memory}_{\text{Total}}}$, where Price is the GPU rent per hour, approximately \$0.2857 for one RTX 4090 GPU we use. $\text{Memory}_{\text{Use}}$ and $\text{Memory}_{\text{Total}}$ represent the average occupied GPU memory and the total memory of the system, respectively. Time is the generation duration measured in seconds. All cost numbers are presented in terms of cents (¢).

Offline Metrics. To facilitate fair comparisons across all baseline methods, we standardize the evaluation metrics for all action types. For each step, we provide three metrics: *Type*, *Ground* and *SR*. Continual on the description in Section 4.1, we further detail on how an action is determined as correct for *SR*.

- For coordinate-related actions, e.g. `Click`, the agents generate both the action type and the position coordinates. Since the ground-truth bounding box is not always available, we measure the performance by computing the distance between the predicted coordinates and the ground-truth coordinates. Following Bai et al. (2024), we deem the coordinates correct if they fall within a distance equivalent to 14% screen

[4] `https://console.cloud.google.com/storage/browser/gresearch/android_control`

Table 11: Action space for Android in the Wild.

Action Type	Attribute	Description
		Basic Actions
CLICK	(x,y)	Click at a specific point on the screen using the coordinates.
TYPE	text	Type the text in the current input field or search bar.
SCROLL	direction	Scroll in a specific direction (one of 'up', 'down', 'left', or 'right').
		Custom Actions
NAVIGATE_BACK	-	Return to the previous page or undo an action.
NAVIGATE_HOME	-	Return to the home page.
PRESS_ENTER	-	Press the 'Enter' button.
COMPLETE	-	Indicate that the task is finished.
IMPOSSIBLE	-	Indicate that the task is infeasible.

width from the ground truth.

- For type-based actions (e.g., `TYPE`, `OPEN_APP`), we compute the F1 score between the predicted text and the ground truth. A prediction is considered correct if the F1 score exceeds 0.5.
- For `SCROLL` actions, the direction argument (i.e., `UP`, `DOWN`, `LEFT`, or `RIGHT`) must precisely match the ground truth.
- For all other actions (e.g., `PRESS_BACK`), the prediction must exactly match the ground truth to be considered correct.

Online Metrics. The evaluation is conducted in screenshot-only mode. To mitigate potential interference from network instability and environmental factors, the results are measured three times. The primary metric is the episode-wise task success rate, a more rigorous measurement compared to the step-wise success rate (*SR*) in offline mode, as en episode is considered successful only when all constituent steps are performed correctly, i.e. *SR* = 100% for a task to be successful.

Data Quality Metrics. Based on the well established literature in NLP community. We use similarity of generated instruction to the ground truth as an indication of data quality. We adopt both text-based metrics which directly computed based on the two sentences and embedding-based metrics.

- **BLEU** (Bilingual Evaluation Understudy) is a precision-based metric that evaluates text similarity by comparing n-grams between generated and reference texts (Papineni et al., 2002).
- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) is a recall-based metric that computes overlapping n-grams, word sequences, and the longest common subsequences (Lin, 2004). The ROUGE family includes ROUGE-1, ROUGE-2, and ROUGE-L, each providing measures for precision, recall, and the F1-score.
- **TF-IDF** (Term Frequency-Inverse Document Frequency) is a statistical measure that evaluates word importance in a document relative to a corpus by balancing term frequency and inverse document frequency (Salton and Buckley, 1988).
- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) is a metric that evaluates text similarity by aligning unigrams between generated and reference texts using exact, stem, synonym, and paraphrase matches. Unlike BLEU, METEOR incorporates both precision and recall, along with a fragmentation penalty to account for word order, resulting in higher correlation with human judgments at the sentence level (Banerjee and Lavie, 2005).
- **Embedding Similarity** which use embedding models to embed the sentences first and calculates the cosine similarity between two embedding vectors. We select two SOTA embedding models with the most downloads on the Hugging Face websites, `jina-v3`[5] and `mxbai-v1`[6].

E.4 Model Details

We employ three categories of models in our experiments: VLMs with conversational capability, base models specialized for GUI tasks with enhanced grounding ability, and API-based closed-ended models.

- **Chat Models.** We select widely used VLMs from prior and contemporary works (Bai et al., 2024; Sun et al., 2024). Specifically, we in-

[5]https://huggingface.co/jinaai/jina-embeddings-v3

[6]https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1

clude the Qwen2-VL family (2B[7], 7B[8]) (Wang et al., 2024c), InternVL2 family (1B[9], 2B[10], 4B[11], 8B[12]) (Chen et al., 2024b), DeepSeek-VL-7B-Chat[13] (Lu et al., 2024a), Phi-3.5-Vision-Instruct[14] from Microsoft (Abdin et al., 2024), Ovis2-4B[15] from AIDC-AI (Lu et al., 2024c), and Yi-VL-6B[16], an early model from 01-AI.

- **GUI Base Models.** We adopt SeeClick[17] (Cheng et al., 2024), which is continually pretrained on Qwen-VL-7B with additional grounding datasets from ScreenSpot (Cheng et al., 2024). We also utilize OS-Atlas-4B[18] and OS-Atlas-7B[19] (Wu et al., 2024), which are trained on InternVL2-4B and Qwen2-VL-7B-Instruct, respectively. These models lack conversational capabilities and are therefore unsuitable for annotation.
- **API-Based Models.** GPT-4o and GPT-4o-Mini (OpenAI, 2023) are widely used vision models provided by OpenAI. These models are significantly more cost-effective than GPT-4V and are frequently utilized in researches. Due to their closed-source and API-only nature, they do not support supervised fine-tuning within our framework and are exclusively used as annotation models.

E.5 Baseline Details

Overall Baselines for Training Mobile GUI Agents. In Section 4.2, we compare existing approaches for data collection and mobile agent training. In this section, we provide further elaboration and details on these baselines.

- **Human-Annotated Data.** Most conventional approaches fall into this category, which involves first employing crowdsourcing to collect and annotate data, followed by training mobile GUI agents. Depending on the training paradigm—centralized or federated—this category can be further divided into two baselines: *Central-Human* and *FedLLM/VLM*. To the best of our knowledge, no prior work has explored training federated VLMs. Therefore, we extend the existing FedLLM framework to the FedVLM setting while retaining the name *FedLLM* for consistency and comparison.
- **Synthetic Data.** This approach (Sun et al., 2024; Su et al., 2025) leverages VLMs to generate synthetic instructions, either based on seed task-driven instructions annotated by humans or through reverse task synthesis. These synthetic instructions are subsequently executed in simulators, by either powerful models such as GPT-4o or by humans, to collect full interaction trajectories. *OS-Genesis* (Sun et al., 2024) is a representative example of this category. Although these methods substantially reduce human labor, they still heavily rely on powerful API-based models and extensive simulator execution, which can become costly at scale.
Due to the unavailability of the original training data, we are unable to directly evaluate *OS-Genesis* within our setting. Instead, we reference reported results from the original paper. For cost estimation, we measure the cost of generating a single data sample using GPT-4o in our setup and extrapolate it to 1,000 samples (the dataset size used in OS-Genesis), yielding the $\approx 10^3$ cost estimates presented in Table 2.
- **DistRL*.** *DistRL* (Wang et al., 2024d) proposes a scalable and asynchronous architecture for data acquisition from multiple simulators in a distributed manner, coupled with centralized reinforced agent training. The framework also introduces techniques to compensate for potential performance degradation caused by asynchrony. We adapt this method to our user-based setting by collecting auto-annotated data in a distributed manner using the *Auto-Annotation* mechanism and training the model centrally. We refer to this adapted baseline as *DistRL**. The key distinction between *MobileA3gent* and *DistRL** lies in the training paradigm, and the latter raises greater privacy concerns due to the exposure of user data to both peers and the

[7] https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct
[8] https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct
[9] https://huggingface.co/OpenGVLab/InternVL2-1B
[10] https://huggingface.co/OpenGVLab/InternVL2-2B
[11] https://huggingface.co/OpenGVLab/InternVL2-4B
[12] https://huggingface.co/OpenGVLab/InternVL2-8B
[13] https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat
[14] https://huggingface.co/microsoft/Phi-3.5-vision-instruct
[15] https://huggingface.co/AIDC-AI/Ovis2-4B
[16] https://huggingface.co/01-ai/Yi-VL-6B
[17] https://huggingface.co/cckevinn/SeeClick
[18] https://huggingface.co/OS-Copilot/OS-Atlas-Base-4B
[19] https://huggingface.co/OS-Copilot/OS-Atlas-Base-7B

server during centralized training.

Annotation Baselines. We compare five baselines for annotating user instructions based on available information, including screenshots and action sequences.

- **Action-Origin,** directly concatenates the original formatted actions into a text string without any inference, representing the simplest method for retrieving user instructions in natural language.
- **Visual-Sense,** (Zhang et al., 2024a) leverages the visual perception capabilities of the annotation VLM to understand the screenshots recorded during task execution. Specifically, we concatenate the sequence of screenshots into one image and feed it into the annotation model for one-shot inference.
- **Self-Instruct,** (Wang et al., 2023) is originally proposed for synthetic data generation using LLMs. We adapt it to infer user intentions from action sequences. In our implementation, all actions are provided simultaneously to the annotation model, which predicts the instruction in a single pass.
- **Chain-of-Thought,** (Berkovitch et al., 2024) guides the annotation model (e.g., GPT-4o) through a step-by-step reasoning process to analyze the task trajectory. At each step, the model predicts the current intention based on all prior information, and the final instruction is determined after the entire task sequence is completed. It is important to note that, although named "Chain-of-Thought," this method is derived from Berkovitch et al. (2024), which focuses on identifying user intentions in GUI tasks, rather than from the original CoT prompting paper (Wei et al., 2022).
- **Human-Annotation,** uses human-annotated gold instructions from the dataset, serving as the upper-bound reference. However, with increasing data scale, methods based on automatic annotation, including ours, can not only achieve comparable or even superior performance, but also substantially reduce annotation costs.

Federated Learning Baselines. We integrate seven representative federated learning algorithms, following the implementations provided in OpenFedLLM (Ye et al., 2024). These include FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), SCAFFOLD (Karimireddy et al., 2020), FedAvgM (Hsu et al., 2019), FedAdagrad, FedYogi, and FedAdam (Reddi et al., 2020).

- **Local Update.** FedAvg is the foundational algorithm upon which many subsequent methods are built. FedProx and SCAFFOLD extend FedAvg by incorporating local model correction mechanisms to mitigate the effects of data heterogeneity.
- **Global Aggregation.** In contrast, FedAvgM, FedAdagrad, FedYogi, and FedAdam introduce server-side momentum techniques to stabilize global model updates.
- **Local Training.** Additionally, we include a local training baseline, where a model is trained solely on a single client's dataset without collaboration. This serves as a reference to highlight the benefits of participating in federated learning.

E.6 Training and Generation Details

Training Setups. The models are trained over 10 rounds, with each round processing one-tenth of the total dataset. This setup ensures that, in expectation, each data sample is seen approximately once throughout the training process.

In the IID federated learning setting, data samples are uniformly distributed across 10 or more clients. In each round, 30% of clients are randomly selected to perform local training and participate in global aggregation. Analogous to centralized training, each selected client processes one-tenth of its local data during that round. Therefore, training for 10 rounds yields an expected 30% overall client participation. To simulate higher participation (e.g., 90%), we extend training to 30 rounds. While in non-IID setting (e.g., experiments in Section 4.4), the data samples are distributed according to the specific scenario.

For experiments investigating the effect of dataset size and scaling, we start with an initial pool of 5,000 data samples. Subsets of smaller sizes are created by selecting the first X samples from this pool to form datasets of size X. This approach guarantees that datasets with larger sample sizes always encompass those with fewer samples, ensuring consistency and comparability across experiments.

Training Framework. We build upon the highly-starred training framework, ms-swift (Zhao et al., 2024) [20], and extend it into a repository capable of training federated VLMs. Our extension follows the implementation of federated training

[20] https://github.com/modelscope/ms-swift

Table 12: Key training parameters regarding FL, LoRA, and quantization.

Parameter	Value	Parameter	Value
Federated Learning			
number-of-rounds	30	number-of-clients	10
number-of-clients-sampled	3	ratio λ	3,5,7,9
LoRA Configuration			
lora-rank	8	lora-alpha	32
lora-dropout	0.05	max-sequence-length	4096, 2048
Optimization			
learning-rate	5×10^{-5}	batch-size	1
optimizer	adamw_torch	gradient-accumulation-steps	4
weight-decay	0.1	adam-beta1	0.9
adam-beta2	0.95	adam-epsilon	1×10^{-8}
lr-scheduler	cosine	warmup-ratio	0.03
Quantization Settings			
bnb-4bit-compute-dtype	torch.bfloat16	bnb-4bit-quant-type	nf4
bnb-4bit-use-double-quant	true	load-in-4bit	false
load-in-8bit	false	device-number	2

framework for Large Language Models (LLMs) (Ye et al., 2024). We apply Low-Rank Adaptation (LoRA) (Hu et al., 2021) to improve efficiency.

Training Parameters. As shown in Table 12, we include all key parameters for reproducibility. For max-sequence-length, we choose 4096 for Qwen2-VL family and 2048 for InternVL2 family. The hyperparameter for various federated algorithms are set as: FedYogi (Reddi et al., 2020) employs momentum factors ($\beta_1 = 0.9, \beta_2 = 0.999$) with learning rate $\eta = 10^{-3}$ and stabilization constant $\tau = 10^{-6}$. FedAvgM (Hsu et al., 2019) uses $0.9/0.1$ ratio for historical/current model interpolation. FedProx (Li et al., 2020) applies proximal regularization with $\mu = 0.2$ through $||w - w^t||^2$ penalty terms. SCAFFOLD (Karimireddy et al., 2020) configurations maintain server learning rate $\eta_s = 1.0$ with client momentum compensation, while FedAdam and FedAdagrad (Reddi et al., 2020) share base parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) with adaptive learning rate scaling. All algorithms expose tunable coefficients through the framework's unified parameter interface.

Templates. We provide all of our prompt templates used in generating instructions and training. Specifically, generation prompts for *Auto-Annotation* are in Figures 12, 13; generation prompt for *Visual-Sense* is provided in Figure 14 with *Chain-of-Thought* in Figure 15; training prompts are shown in Figure 16, 17 and 18 for all three offline datasets respectively.

Prompt 1: Step-Wise Description

A user is performing a *task* on a mobile phone, progressing through **multiple steps** to complete the task.
Each step involves an interface shown in the provided screenshot, and an action performed to move on to the next step.

Based on the screenshot and the user's action, infer the specific goal the user is trying to accomplish at this step in the task.
You need to associate the action with the key information in the screenshot and output your predicted goal.

Example
- User Action: Scroll down
if the screenshot shows the browsing page for purchasing shoes,
- Your Output: Swipe up for more product details about shoes

- User Action: Click (101,314)
if the UI element at this coordinate is an article titled "cooking"
- Your Output: Click on the article titled "cooking"

- User Action: Check status: successful
- Your Output: Check if the task is finished

- User Action: Open App: Plantum
if the action is *open app*, return the same
- Your Output: Open App: Plantum

- User Action: Wait for response
if the action is *wait*, return none
- Your Output: None

Answer Format
Only output the predicted goal. Be specific with the input screenshot.
Keep your response concise and capture the important things, focusing on key details like the app name, email address, search terms, item name, and title.

User Action
{converted action A_i }

Your Output
The user is trying to:

Figure 12: Prompt template for the Descriptor to generate low-level instruction $\mathcal{T}_i^{low}$ based on the converted action A_i and screenshot s_i at the i-th step .

Prompt 2: Episode-Wise Summarization within Auto-Annotation

A user is performing a high-level **task** on a mobile phone, progressing through multiple low-level steps to complete the task.
Each step involves an interface, and a low-level action performed to move on to the next step.

The full sequence of user actions is provided in the *History* section.
The **task** is not known. Now based on the history provided, describe the mobile user's high-level **task** when performing these actions.

History
{ low-level instruction $\mathcal{T}_1^{low}$ }
{ low-level instruction $\mathcal{T}_2^{low}$ }
...
{ low-level instruction $\mathcal{T}_n^{low}$ }

Answer Format
Keep your output concise and clear, as if the user were explaining the **task** to someone else in one sentence.
Include key details like the app name, individual name, email address, search terms, item name, and title.

Your Output
The user is trying to:

Figure 13: Prompt template for the Summarizer to generate high-level instruction $\mathcal{T}^{high}$ based on the list of low-level instructions and the concatenated screenshot s_c .

Prompt 3: Episode-Wise Summarization with Visual-Sense

A user is performing a high-level **task** on a mobile phone, progressing through multiple low-level steps to complete the task.
Each step involves an interface, and a low-level action performed to move on to the next step.

A single image that shows all the screenshots concatenated horizontally is provided.
The **task** is not known. Now based on this concatenated screenshot, describe the mobile user's high-level **task** when performing these actions.

Answer Format
Keep your output concise and clear, as if the user were explaining the **task** to someone else in one sentence.
Include key details like the app name, individual name, email address, search terms, item name, and title.

Your Output
The user is trying to:

Figure 14: Prompt template for *Visual-Sense* to generate high-level instruction $\mathcal{T}^{high}$ based on the list of converted actions and the concatenated screenshot s_c .

Prompt 4: Step-Wise Description with Chain-of-Thought

A user is performing a high-level **task** on a mobile phone, progressing through multiple low-level steps to complete the task.
Each step involves an interface, and a low-level action performed to move on to the next step.

The previous task descriptions for each step are provided in the *History* section, and the user's final action is provided in the *User Action* section.
You need to think step by step and analyze the input sequence to deduce the user's underlying objective that prompted these actions.
Utilize the screenshot of the final step to gain insights into the user's intentions, focusing on elements highlighted or implicated by the actions.
Your goal is to describe the ultimate intention the user is aiming to achieve.

History
{ low-level instruction $\mathcal{T}_1^{low}$ }
{ low-level instruction $\mathcal{T}_2^{low}$ }
...
{ low-level instruction $\mathcal{T}_{i-1}^{low}$ }

User Action
{converted action A_i }

Answer Format
Keep your output concise and clear, as if the user were explaining the **task** to someone else in one sentence.
Include key details like the app name, individual name, email address, search terms, item name, and title.

Your Output
The user is trying to:

Figure 15: Prompt template for *Chain-of-Thought* to generate instruction step-by-step and finally obtain the high-level instruction.

Prompt 5: Common Prompt for Training

You are a foundational action model capable of automating tasks across various digital environments, including desktop systems like Windows, macOS, and Linux, as well as mobile platforms such as Android and iOS. You also excel in web browser environments. You will interact with digital devices in a human-like manner: by reading screenshots, analyzing them, and taking appropriate actions.

Your expertise covers two types of digital tasks:

- Grounding: Given a screenshot and a description, you assist users in locating elements mentioned. Sometimes, you must infer which elements best fit the description when they aren't explicitly stated.
- Executable Language Grounding: With a screenshot and task instruction, your goal is to determine the executable actions needed to complete the task.

You are now operating in Executable Language Grounding mode. Your goal is to help users accomplish tasks by suggesting executable actions that best fit their needs. Your skill set includes both basic and custom actions:

1. Basic Actions
Basic actions are standardized and available across all platforms. They provide essential functionality and are defined with a specific format, ensuring consistency and reliability.

- Basic Action 1: CLICK
  - purpose: Click at the specified position.
  - format: `CLICK <point>[[x-axis, y-axis]]</point>`
  - example usage: `CLICK <point>[[101, 872]]</point>`
- Basic Action 2: TYPE
  - purpose: Enter specified text at the designated location.
  - format: `TYPE [input text]`
  - example usage: `TYPE [Shanghai shopping mall]`
- Basic Action 3: SCROLL
  - purpose: Scroll in the specified direction.
  - format: `SCROLL [direction (UP/DOWN/LEFT/RIGHT)]`
  - example usage: `SCROLL [UP]`

Figure 16: Prompt template for the common part shared between different datasets during training of federated mobile agents within MobileA3gent. The full training prompt is the combination of the common part and the custom part.

Prompt 6: Custom Prompt for Training on AndroidControl

2. Custom Actions
Custom actions are unique to each userś platform and environment. They allow for flexibility and adaptability, enabling the model to support new and unseen actions defined by users. These actions extend the functionality of the basic set, making the model more versatile and capable of handling specific tasks.

- Custom Action 1: LONG_PRESS
  - purpose: Long press at the specified position.
  - format: `LONG_PRESS <point>[[x-axis, y-axis]]</point>`
  - example usage: `LONG_PRESS <point>[[272, 341]]</point>`
- Custom Action 2: NAVIGATE_BACK
  - purpose: Press a back button to navigate to the previous screen.
  - format: `NAVIGATE_BACK`
  - example usage: `NAVIGATE_BACK`
- Custom Action 3: NAVIGATE_HOME
  - purpose: Press a home button to navigate to the home page.
  - format: `NAVIGATE_HOME`
  - example usage: `NAVIGATE_HOME`
- Custom Action 4: OPEN_APP
  - purpose: Open the specified application.
  - format: `OPEN_APP [app_name]`
  - example usage: `OPEN_APP [Google Chrome]`
- Custom Action 5: WAIT
  - purpose: Wait for the screen to load.
  - format: `WAIT`
  - example usage: `WAIT`
- Custom Action 6: COMPLETE
  - purpose: Indicate the task is finished.
  - format: `COMPLETE`
  - example usage: `COMPLETE`

In most cases, task instructions are high-level and abstract. Carefully read the instruction and action history, then perform reasoning to determine the most appropriate next action. Ensure you strictly generate two sections: **Thoughts** and **Actions**.

Thoughts: Clearly outline your reasoning process for current step.
Actions: Specify the actual actions you will take based on your reasoning.

Your current task instruction, action history, and associated screenshot are as follows:
Screenshot: <image>
Task: {high-level instruction $\mathcal{T}^{high}$}
You need to: {low-level instruction $\mathcal{T}_i^{low}$}
History: {history of $\mathcal{T}_i^{low}$}

Figure 17: Custom prompt template for training mobile GUI agents on AndroidControl.

Prompt 7: Custom Prompt for Training on GUI Odyssey

Custom actions are unique to each userś platform and environment. They allow for flexibility and adaptability, enabling the model to support new and unseen actions defined by users. These actions extend the functionality of the basic set, making the model more versatile and capable of handling specific tasks.

- Custom Action 1: LONG_PRESS
  - purpose: Long press at the specified position.
  - format: `LONG_PRESS <point>[[x-axis, y-axis]]</point>`
  - example usage: `LONG_PRESS <point>[[272, 341]]</point>`
- Custom Action 2: NAVIGATE_BACK
  - purpose: Press a back button to navigate to the previous screen.
  - format: `NAVIGATE_BACK`
  - example usage: `NAVIGATE_BACK`
- Custom Action 3: NAVIGATE_HOME
  - purpose: Press a home button to navigate to the home page.
  - format: `NAVIGATE_HOME`
  - example usage: `NAVIGATE_HOME`
- Custom Action 4: PRESS_RECENT
  - purpose: Press the recent button to view or switch between recently used applications.
  - format: `PRESS_RECENT`
  - example usage: `PRESS_RECENT`
- Custom Action 5: WAIT
  - purpose: Wait for the screen to load.
  - format: `WAIT`
  - example usage: `WAIT`
- Custom Action 6: COMPLETE
  - purpose: Indicate the task is finished.
  - format: `COMPLETE`
  - example usage: `COMPLETE`

In most cases, task instructions are high-level and abstract. Carefully read the instruction and action history, then perform reasoning to determine the most appropriate next action. Ensure you strictly generate one section: **Actions**.

Actions: Specify the actual actions you will take based on your reasoning.
Your current task instruction, action history, and associated screenshot are as follows:
Screenshot: <image>
Task: {high-level instruction $\mathcal{T}^{high}$}

Figure 18: Custom prompt template for training mobile GUI agents on GUI Odyssey.

Towards an Automated Framework to Audit Youth Safety on TikTok

Linda Xue[1], Francesco Corso[2], Nicolo' Fontana[2], Geng Liu[2], Stefano Ceri[2], Francesco Pierri[2]

[1]Massachusetts Institute of Technology, [2]Politecnico di Milano
Correspondence: francesco.pierri@polimi.it

Abstract

This paper investigates the effectiveness of TikTok's enforcement mechanisms for limiting the exposure of harmful content to youth accounts. We collect over 7000 videos, classify them as harmful vs not-harmful, and then simulate interactions using age-specific sockpuppet accounts through both passive and active engagement strategies. We also evaluate the performance of large language (LLMs) and vision-language models (VLMs) in detecting harmful content, identifying key challenges in precision and scalability.

Preliminary results show minimal differences in content exposure between adult and youth accounts, raising concerns about the platform's age-based moderation. These findings suggest that the platform needs to strengthen youth safety measures and improve transparency in content moderation.

1 Introduction

TikTok is a short-form video platform that has rapidly emerged as one of the world's most influential social media services. With over 1.6 billion monthly active users[1] and millions of videos uploaded daily (Corso et al., 2024), it now plays a central role in the global digital media landscape.

Children and adolescents increasingly rely on TikTok for both entertainment and everyday information (Violot et al., 2024; Liu et al., 2024; Ge et al., 2021). Although TikTok enforces community guidelines through content removal and age-based restrictions [2], concerns remain about the effectiveness of these moderation mechanisms in shielding young users from harmful content. These concerns are amplified by the introduction of the European Digital Services Act [3], which requires very large online platforms to assess and mitigate systemic risks—particularly those related to the protection of minors and the spread of harmful content.

This paper reports preliminary findings from an ongoing work aiming to build an automated pipeline for auditing TikTok's safety enforcement mechanisms by systematically measuring the exposure to harmful content among youth and adult users, investigating different modes of interaction. Specifically, we explore two main research questions:

- **RQ1:** What level of harmful content will an adult versus a minor be exposed to on their FYF?
- **RQ2:** Does actively searching for harm-adjacent keywords increase exposure to inappropriate content?

To this end, we created 10 Youth and 10 Adult sockpuppet accounts and simulated multiple sessions over several days, collecting over 7,000 videos across both "For You Feed" (FYF) scrolling and using active search with harm-adjacent keywords.

In addition, as a precise detection of harmful content is an heavy task for humans, we also consider a third research question:

- **RQ3:** How effective are Large Language Models such as GPT-4o and VideoLLaMA3 at detecting harmful content?

To this end, we employed both text-only and video-based LLMs to estimate the presence of harmful content that violates TikTok's community guidelines and evaluated their performance.

[1]https://www.businessofapps.com/data/tik-tok-statistics, accessed on August 5, 2025
[2]https://www.tiktok.com/community-guidelines, accessed on August 8, 2025
[3]https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 113–119
November 9, 2025

2 Related Work

TikTok's For You Feed (FYF) algorithm personalizes content based on user language, location, posting time, and interactions such as likes and follows (Boeker and Urman, 2022). This personalization engine rapidly amplifies interest-aligned content—often within just 200 recommendations—thereby fostering echo chambers and limiting content diversity (Baumann et al., 2025).

These concerns are especially salient given TikTok's young user base—over 60% of users were under 30 in 2021 (Iqbal, 2022). A large-scale audit using more than 100 automated accounts found that watch time plays a central role in shaping recommendations, reinforcing problematic content loops through prolonged exposure (WSJ Staff, 2021).

Recent studies further contextualize harm by examining user behavior. For instance, the median user consumes approximately 90 videos daily (Zannettou et al., 2024), while moderately addicted users average 7.86 minutes per session (Yang et al., 2025). An experimental audit comparing TikTok, YouTube, and Instagram showed that accounts registered as 13-year-olds encountered harmful content more frequently and rapidly than accounts of older users (Eltaher et al., 2025).

3 Methods

3.1 Data Collection

To investigate how harmful content varies by user age and interaction mode, we created 20 TikTok accounts using the platform's web version: 10 accounts were set with an age of 13 (Youth) and 10 with an age slightly above 18 (Adult). All accounts were registered in Italy. These age values were chosen as they represent the boundary between TikTok's definition of "youth" (under 18) and adulthood[4].

Using a script we built [5] to scrape data from the TikTok website, we collected data over four consecutive days – Thursday through Sunday – to capture differences across both weekdays and weekends. For each account on each day, four browsing sessions were conducted, each containing 22 videos, totaling 88 videos per account per day. This approximates the average daily video exposure (89.9 videos) and simulates moderately engaged users, based on prior TikTok usage studies reporting 27 minutes total watch time and 7.86 minutes per session (Yang et al., 2025). In total, we collected over 7,000 videos across 20 accounts over the four-day period. Our dataset includes metadata for every collected video, such as the description text, hashtags, and engagement statistics (views, likes, comments, shares, etc.). For each video, we also retrieved the top 10 comments via HTTP requests to analyze user interactions.

We implemented two primary user interaction modes:

Passive Scrolling: Simulated natural browsing behavior by programmatically loading videos from the FYF, with randomized delays (10–20 seconds) between each request to mimic scrolling. No user input was provided beyond passive viewing, aligning with typical user consumption patterns.

Active Searching: For each harmful content category, we extracted three keywords based on TikTok's Youth Safety and Well-Being Guidelines. We used these keywords in the search bar to retrieve videos potentially related to sensitive topics using one Adult and one Youth account. A complete list of keywords is provided in Appendix B, Table 2.

Interestingly, sometimes keywords such as "alcohol" were censored in English for youth accounts, whereas its Italian equivalent "Alcol" still yielded search results though explicitly stated as age-inappropriate in the guidelines.

3.2 Harmful Content Detection

Our framework for identifying harmful content is grounded in TikTok's official Community Guidelines, particularly the sections related to youth safety. Categories such as sexual content, suicide/self-harm, and physical violence were prioritized. Closely related categories (e.g., youth and adult sexual abuse) were merged, while those that may have a less critical societal impact (e.g., animal abuse [6]) were excluded. See Table 1 in Appendix A for details. We employed three methods for detecting harmful content:

1) Textual Analysis: We used the multilingual Detoxify[7] model to evaluate the toxicity of the top 10 comments per video. Detoxify outputs scores in the range $[0, 1]$ to quantify the probability of a comment being toxic or not. We then fed all video

[4] https://support.tiktok.com/en/account-and-privacy/account-privacy-settings/privacy-and-safety-settings-for-users-under-age-18

[5] https://anonymous.4open.science/r/tiktok-scraper-8424

[6] https://www.healthdata.org/research-analysis/diseases-injuries-risks/factsheets-hierarchy

[7] https://github.com/unitaryai/detoxify

descriptions to GPT-4o [8], prompting the model to classify a video based on TikTok's harmful content guidelines.

2) Visual Content Analysis: We tested the performance of VideoLLaMA3 [9] by using it on a random selection of 100 videos. A custom prompt was used to assess visual and audio cues based on the same guideline framework as the GPT-based description classification.

3) Manual Evaluation: The same 100 videos was manually labeled by three native Italian speakers. Each reviewer independently categorized the content using our framework. Disagreements between two annotators were resolved by the third reviewer, ensuring reliable ground-truth labels for evaluating the automated methods.

4 Results

4.1 Descriptive Statistics

We analyzed the sample of videos collected with the two sets of accounts. As shown in Figure 1, the distributions of views, likes, and comments are nearly identical across the two groups, suggesting that the two account modalities expose users to videos with similar engagement on the platform. We did not observe differences in the data collected on weekends or weekdays.

4.2 Prevalence of toxic comments

To analyze harmful content in the comments shown below videos, we compare the distribution of toxicity across all collected comments as well as the maximum toxicity of individual comments for each video.

As shown in Figure 2, the distributions of toxicity scores for both adult and youth accounts exhibit very similar patterns across the two types of analysis–overall toxicity of all comments and maximum toxicity per video. Assuming a commonly used threshold of 0.5–0.7 (Hua et al., 2020) to classify a comment as toxic, the vast majority of comments in both groups would not be considered harmful, as the 95th percentile of toxicity is below 0.3 (0.26-0.28) for videos in both groups of accounts. Median values are also very similar in the two samples ($\tilde{0}.03$), for both analyses.

[8] https://platform.openai.com/docs/models/gpt-4o

[9] https://huggingface.co/DAMO-NLP-SG/VideoLLaMA3-7B

These small differences suggest comparable levels of exposure to toxic content in the comment sections of videos shown to both adults and youth.

4.3 Estimated prevalence of harmful videos

We estimated the prevalence of harmful videos shown to different accounts by using GPT-4o to annotate content based solely on video descriptions. As shown in Figure 3, fewer than 10% of the videos were predicted to be harmful for both adult and youth accounts during passive FYF scrolling. However, two youth accounts exhibited notably higher proportions, with about 14% and 25% of their videos labeled as harmful.

When focusing on active keyword-based searches using harm-related terms, the estimated prevalence of harmful content increases substantially for both groups. Specifically, 28.44% of videos surfaced for adult accounts and 27.91% for youth accounts were classified as harmful, showing rates significantly higher than during passive exposure.

These findings suggest that TikTok's safety mechanisms may be insufficient in protecting younger users from exposure to potentially harmful content, particularly when users actively search for related material.

4.4 Evaluating the performance of harmful content classifiers

We then evaluated the performance of GPT-4o and VideoLLaMA3 by manually annotating the random sample of 100 videos–50 from each age group, with 25 per group collected via keyword-based searches. The Fleiss' Kappa coefficient for inter-annotator agreement was 0.45, indicating moderate agreement. The raw mean percent agreement was 87.3% (Landis and Koch, 1977). The gap between these measures reflects the high expected chance of agreement in our binary harmful-content classification task.

Manual annotation identified genuinely harmful content in 41 of the 100 videos, 26 in the Adult sample and 15 in the Youth sample. Against this manual labelling, GPT-4o achieved a precision of 59% and a recall of 24.4% (10/41), and VideoLLaMA3 achieved 58% and 0.05% (2/41) respectively.

5 Discussion

We analyzed a dataset of over 7,000 TikTok videos collected from age-specific sockpuppet accounts

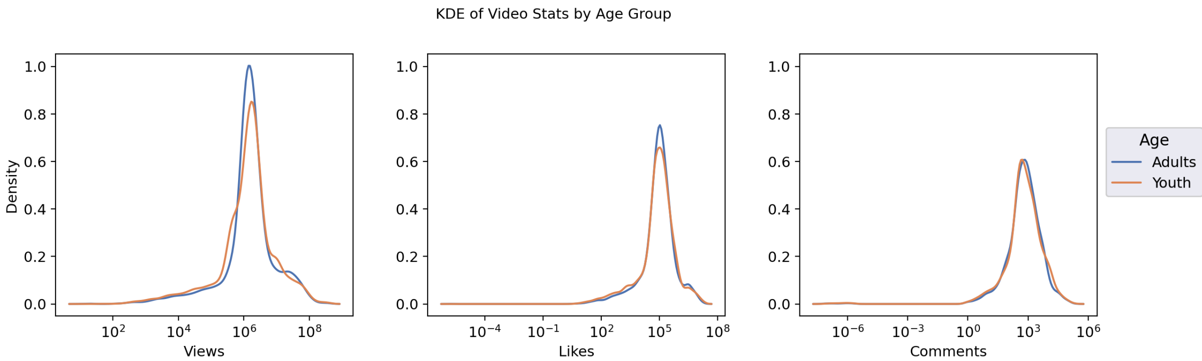


Figure 1: Distributions of metrics for videos collected by accounts belonging to the Adult and Youth group.

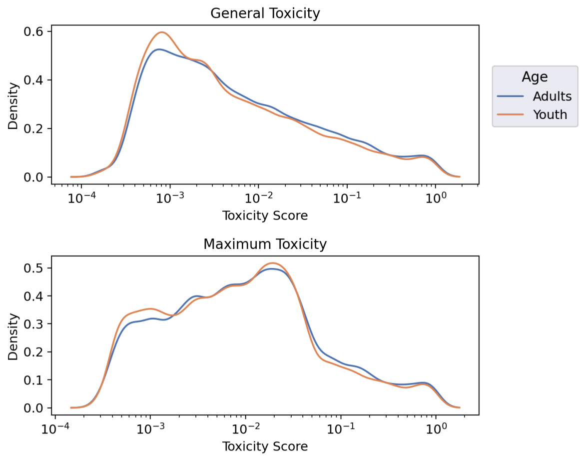


Figure 2: Distribution of toxicity for all comments (top) and the most toxic comment (bottom) below videos collected with the two groups of accounts.

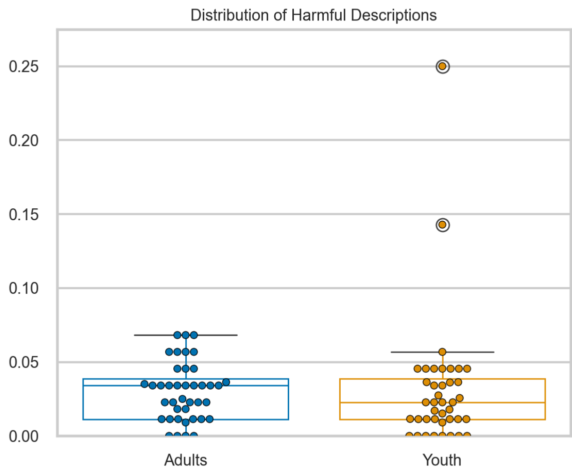


Figure 3: Distribution of estimated proportion of harmful videos shown to each account in the two groups, using labels provided by GPT-4o. Median values are 0.034 for Adults and 0.023 for Youths.

using both passive and active interaction protocols. Our findings indicate that adult and youth accounts were exposed to highly similar content, with fewer than 1 in 10 videos estimated to be harmful. This suggests that TikTok's moderation mechanisms may not meaningfully differentiate between age groups in practice. Despite this, automated detection of harmful content remains challenging: experiments with VLMs showed limited precision, with GPT-4o — operating on text only - outperforming VideoLLaMA3, which had access to video content.

Our work presents promising results towards building a scalable and reproducible approach for auditing exposure to harmful content on recommender-driven platforms. By systematically comparing the content shown to adult and youth users, one can uncover tangible shortcomings in age-based content moderation, especially concerning the search function. Our results suggest that platforms like TikTok should invest greater effort in strengthening youth safety protections and ensuring greater transparency in how content is being moderated for these specific audiences.

Future research will expand the scope of this study by including additional countries to assess cross-cultural differences in content moderation. We also plan to scale up the experiment with a larger and more diverse set of accounts, including personalized profiles that better simulate real user behavior. To improve ecological validity, future data collection will incorporate the mobile interface, reflecting the primary mode of user interaction on TikTok. Finally, we aim to develop and integrate more accurate prediction models for detecting harmful content, leveraging multimodal signals more effectively. Due to models such as GPT-4o's expensive costs, we plan to investigate and incorporate techniques such as the teacher and student model, which has been previously used for misinformation detection (Jung et al., 2025).

Ethical Considerations and Limitations

Our study involves the collection and analysis of publicly available content from TikTok through automated scraping techniques. This data collection was conducted solely for academic research purposes, with the aim of auditing systemic risks on a very large online platform, an activity explicitly permitted under the DSA. In line with the DSA's provisions, our work contributes to the broader public interest by examining the effectiveness of safety enforcement mechanisms, particularly as they relate to youth protection and the dissemination of harmful content. We do not attempt to identify or track individual users, nor do we collect personal data or metadata that could be used to do so. Furthermore, none of the collected data is released or shared in a way that could compromise user anonymity or platform integrity. As our research does not involve human subjects or interactions with real users, it does not require approval from an institutional ethics review board.

Nevertheless, this work has several limitations. First, our classification of harmful content relies on a limited set of LLMs, and analyzes video description and text separately. Second, our analysis focuses exclusively on the Italian language and user experience within Italy, which may limit the generalizability of our findings to other linguistic or cultural environments. Additionally, we simulate user behavior using a relatively small number of sockpuppet accounts, which may not fully reflect the diversity of real user interactions on the platform. Finally, due to the amount of comments per video, we could not manually label every comment and analyze Detoxify.

References

Fabian Baumann, Nipun Arora, Iyad Rahwan, and Agnieszka Czaplicka. 2025. Dynamics of algorithmic content amplification on tiktok. *arXiv preprint arXiv:2503.20231*.

Maximilian Boeker and Aleksandra Urman. 2022. An empirical investigation of personalization factors on tiktok. In *Proceedings of the ACM Web Conference 2022*, pages 2298–2309.

Francesco Corso, Francesco Pierri, and Gianmarco De Francisci Morales. 2024. What we can learn from tiktok through its research api. In *Companion Publication of the 16th ACM Web Science Conference*, pages 110–114.

Fatmaelzahraa Eltaher, Rahul Krishna Gajula, Luis Miralles-Pechuán, Patrick Crotty, Juan Martínez-Otero, Christina Thorpe, and Susan McKeever. 2025. Protecting young users on social media: Evaluating the effectiveness of content moderation and legal safeguards on video sharing platforms. *arXiv preprint arXiv:2505.11160*.

Jiaoju Ge, Yuepeng Sui, Xiaofeng Zhou, and Guoxin Li. 2021. Effect of short video ads on sales through social media: The role of advertisement content generators. *International Journal of Advertising*, 40(6):870–896.

Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Mohsin Iqbal. 2022. Tiktok revenue and usage statistics (2022). Accessed: 2025-08-05.

Hayoung Jung, Shravika Mittal, Ananya Aatreya, Navreet Kaur, Munmun De Choudhury, and Tanushree Mitra. 2025. Mythtriage: Scalable detection of opioid use disorder myths on a video-sharing platform. *Preprint*, arXiv:2506.00308.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mingli Liu, Aixia Zhuang, Jill M. Norvilitis, and Tian Xiao. 2024. Usage patterns of short videos and social media among adolescents and psychological health: A latent profile analysis. *Computers in Human Behavior*, 151:108007.

Caroline Violot, Tugrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. 2024. Shorts vs. regular videos on youtube: A comparative analysis of user engagement and content creation trends. In *Proceedings of the 16th ACM Web Science Conference (WebSci '24)*, pages 213–223, New York, NY, USA. Association for Computing Machinery.

WSJ Staff. 2021. Inside tiktok's algorithm: A wsj video investigation. Accessed: 2025-08-05.

Cai Yang, Sepehr Mousavi, Abhisek Dash, Krishna P Gummadi, and Ingmar Weber. 2025. Studying behavioral addiction by combining surveys and digital traces: A case study of tiktok. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 2106–2123.

Savvas Zannettou, Olivia-Nemes Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P. Gummadi, Elissa M. Redmiles, and Franziska Roesner. 2024. Analyzing user engagement with tiktok's short format video recommendations using data donations. *Preprint*, arXiv:2301.04945.

A Details of Harmful Content Classification

A.1 Framework

Table 1 summarizes the harmful content categories used in our analysis, based on TikTok's official Community Guidelines.

A.2 GPT-4o Prompt

> Classify the following TikTok video **descriptions** as 'harmful' or 'not harmful' under this harmful content framework:
>
> - Disordered Eating and Body Image
> - Suicide and Self-Harm
> - Dangerous Activity and Challenges
> - Nudity and Body Exposure
> - Sexually Suggestive Content
> - Shocking and Graphic Content
> - Hate Speech or Behavior
> - Sexual and Physical Abuse
> - Human Trafficking and Smuggling
> - Gambling
> - Alcohol, Tobacco, and Drugs
> - Integrity and Authenticity
> - Harassment and Bullying
>
> Reply in the format: `harmful/not harmful` based on whether the TikTok falls under one of these groups.

A.3 VideoLLaMA Prompt

> Classify the following TikTok videos as 'harmful' or 'not harmful' under this harmful content framework:
>
> - Disordered Eating and Body Image
> - Suicide and Self-Harm
> - Dangerous Activity and Challenges
> - Nudity and Body Exposure
> - Sexually Suggestive Content
> - Shocking and Graphic Content
> - Hate Speech or Behavior
> - Sexual and Physical Abuse
> - Human Trafficking and Smuggling
> - Gambling
> - Alcohol, Tobacco, and Drugs
> - Integrity and Authenticity
> - Harassment and Bullying
>
> Reply in the format: `harmful/not harmful` based on whether the TikTok falls under one of these groups.

B Keywords for Active Searching

Table 2 lists the Italian keywords associated with each harmful content category, derived from TikTok's Youth Safety and Well-Being Guidelines. These keywords were used in the active searching experiments to retrieve videos potentially related to sensitive topics.

Table 1: Harmful content categories based on TikTok community guidelines

Category	Definition
Disordered Eating and Body Image	Showing or promoting disordered eating and dangerous weight loss behaviors, or facilitating the trade or marketing of weight loss or muscle gain products.
Suicide and Self-Harm	Showing, promoting, or sharing plans for suicide or self-harm.
Dangerous Activity and Challenges	Showing or promoting dangerous activity and challenges.
Nudity and Body Exposure	Bare genitalia, buttocks, breasts of women and girls, sheer clothing, or semi-nudity or significant body exposure of young people.
Sexually Suggestive Content	Content by young people that intends to be sexually suggestive, including intimate kissing, sexualized framing, or behavior, and sexually explicit language.
Shocking and Graphic Content	Gory, gruesome, disturbing, or extremely violent content.
Sexual and Physical Abuse	Showing, promoting, or engaging in adult or youth sexual or physical abuse or exploitation.
Human Trafficking and Smuggling	Recruitment and transport of victims using force, fraud, or coercion, or smuggling someone across borders illegally for profit.
Hate Speech or Behavior	Hate speech, hateful behavior, or promotion of hateful ideologies.
Gambling	Betting money or items of value on uncertain outcomes, or engaging in gambling-like behaviors such as social casinos.
Alcohol, Tobacco, and Drugs	Trade, possession, or use of drugs, alcohol, or tobacco products, including regulated substances and youth use.
Harassment and Bullying	Violent threats, hate speech, sexual exploitation, doxxing, or other serious forms of harassment.
Integrity and Authenticity	Misinformation, election integrity issues, AI-generated deception, fake engagement, spam, and unoriginal or deceptive content.

Table 2: Keyword examples associated with TikTok harmful content categories (in Italian)

Category	Term 1	Term 2	Term 3
Disturbi legati all'alimentazione e immagine corporea	Diete restrittive	perdere peso	chirurgia estetica
Challenge e attività pericolose	challenge pericolose	attività pericolose	tendenza
Nudità ed esposizione corporea	nudità	esposizione corporea	biancheria intima
Contenuti sessualmente allusivi	baci intimi	sesso	prodotti per il sesso
Contenuti grafici e scioccanti	sangue	combattimenti	disastro naturale
Gioco	gioco	casinò	scommesse
Alcol e tabacco e droghe	droghe	tabacco	alcol

Predictive Modeling of Human Developers' Evaluative Judgment of Generated Code as a Decision Process

Sergey Kovalchuk
Chebyshev Research Center
sergey.kovalchuk@huawei.com

Yanyu Li
ITMO University
268904@niuitmo.ru

Dmitry Fedrushkov
Chebyshev Research Center
fedrushkov.dmitriy1@huawei.com

Abstract

The paper presents early results in the development of an approach to predictive modeling of human developer perceiving of code generated in question-answering scenarios with Large Language Model (LLM) applications. The study is focused on building a model for the description and prediction of human implicit behavior during evaluative judgment of generated code through evaluation of its consistency, correctness, and usefulness as subjective perceiving characteristics. We used Markov Decision Process (MDP) as a basic framework to describe the human developer and his/her perceiving. We consider two approaches (regression-based and classification-based) to identify MDP parameters so it can be used as an "artificial" developer for human-centered code evaluation. An experimental evaluation of the proposed approach was performed with survey data previously collected for several code generation LLMs in a question-answering scenario. The results show overall good performance of the proposed model in acceptance rate prediction (accuracy 0.82) and give promising perspectives for further development and application.

1 Introduction

Today, large language models (LLMs) are widely applied in the practice of software development, with both general-purpose solutions like ChatGPT by OpenAI and solutions dedicated to code writing such as CoPilot by Microsoft. One of the important capabilities of LLMs is support of code generation in various scenarios (Lu et al., 2021; Zhong et al., 2022): bug fixing, code completion, question answering with code snippets, and many others. Still, practical implementation of solutions for these tasks reveals several fundamental issues related to the complexity of the software development domain and the specificity of human developers as solution users.

An important problem is the evaluation of the solutions. A straightforward approach for LLM output evaluation is linguistic metrics such as BertScore, BLEU, and others. Complex semantics and non-linearity of code structure lead to the development of code-specific metrics such as CodeBLEU (Ren et al., 2020), RUBY (Tran et al., 2019), and others. Nevertheless, the real-world application of such metrics shows significant limitation in LLM evaluation (Evtikhiev et al., 2023). Another approach is application of test-based evaluation of generated code with such metrics as Pass@k (passing tests with k generated answers). Still, application of such an approach remains limited due to the lack of tests and limited applicability of automatically generated tests to the real-world problem. The problem significantly influences the performance of LLMs in real-world complex projects, which is clearly seen in modern project-based benchmarks like SWE-bench (Jimenez et al., 2024), RepoBench (Liu et al., 2024), CoderEval (Yu et al., 2023), etc. The benchmarks show relatively weak performance even for state-of-the-art solutions. One of the best-known solutions for the evaluation problem is evaluation of LLMs with human-centered metrics like acceptance rate (AR). More complicated approaches may involve value, accuracy, and other human-centered metrics (Dibia et al., 2023). Still, involvement of human developers requires significant time and effort, with the involvement of multiple human users.

Another problem is proper understanding of real human developers roles, needs, intents, and expectations. The practical application and surveys of the developers applying LLM-based solutions in daily tasks reveal several important issues (Bird et al., 2022; Ernst and Bavota, 2022; Liang et al., 2023; Shi et al., 2024). Users often report a lack of personalization, efforts needed to understand generated code, differences between code generated by humans and by LLMs, etc. As a consequence, this

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 120–128
November 9, 2025 ©2025 Association for Computational Linguistics

leads to limited trust of developers (Wang et al., 2023), possible vulnerability in generated code (Risse and Böhme, 2023), weak performance in real-world issues (Jimenez et al., 2024), etc. On the other hand, investigations on interaction with CoPilot show that proper time to show suggestions (Mozannar et al., 2023, 2022) and configuration of interaction patterns (Wang et al., 2023) show possible increases in the acceptance rate of suggestions generated by considered intelligent assistants. A key open problem here is understanding how human developers perceive, comprehend, and evaluate code in proper context (Roehm et al., 2012). Formal structuring of project context is currently approached by many solutions (see, e.g., project-specific benchmarks mentioned above). But the context of the human mind in evaluative judgment of code generated by both humans and machines is weakly investigated.

Resolving the mentioned problems (human-level evaluation of code and human developer internal context representation) is limiting many directions in the application of LLM to software development. The list of directions benefiting from resolving the problem includes training and fine-tuning LLM for better code generation (e.g., with reinforcement learning with human feedback, RLHF); building a complete pipeline for software development (Hong et al., 2023); improving human developer experience through better selection of available actions in AI agents.

In the presented paper, we are focused on approaching the mentioned problems through modeling of human developer perceiving of the code obtained from generative LLM. With the data collected in the previous developers' survey, we've modeled key perceiving characteristics that influence developers' actions in code acceptance evaluation. The approach is based on the idea of sequential decision on accepting or rejecting considered information (code). Thus, the basic idea of human developer perceiving modeling is formulated as a Markov Decision Process (MDP). The paper presents early, still promising results of the ongoing study in human developer perceiving modeling.

The remaining paper is structured as follows. The next section briefly describes problem definition, background for this work, and data collection for the experimental study. Following, Section 3 presents key elements of the proposed approach to human developer modeling. Section 4 presents the results of the experimental evaluation of the proposed approach. Finally, the last sections conclude the paper and discuss possible further directions of research.

Figure 1: Elements of user interface for evaluation of code generated by LLM in question-answering scenario

2 Problem definition and background works

The problem of a human developer perceiving the code generated by an intelligent assistant (IA) such as CoPilot usually considered within some scenario (e.g. code completion, bug fixing, etc.). The developer posts a query to the IA and perceives the answer. In some cases, the query is proactively posted automatically, depending on the developer activity. IA answers with a block of information containing the answer, suggestions, explanation, etc. Within our work, we are narrowing to the scenario where the user asks a question in natural language to AI powered by LLM and expects a piece of code as an answer. According to the classification of the CodeXGLUE benchmark (Lu et al., 2021), the problem is Text-to-Code generation. The goal is to build a model for predicting a human developer's subjective evaluation and final AR for code generated with LLM as an answer.

The examples of practical problems being solved in such a way may be widely found in online forums where users ask questions to resolve some programming issues. One of the most popular forums for the software development domain is StackOverflow[1] (SO) which is also widely used

[1] https://stackoverflow.com/

as an original source for training and evaluation of LLMs. A common pattern for such questions is “how to...” (use API, implement algorithm, fix bug, etc.) where an expected answer is a piece of code. As a result, SO is widely adopted as a source for datasets and benchmarks building in Text-to-Code problem investigation: see, e.g., such datasets as CoNaLa (Yin et al., 2018), StaQC (Yao et al., 2018), SO-DS (Heyman and Cutsem, 2020), etc.

Within our previous study (Kovalchuk et al., 2022), we used the data originating from SO and containing queries to fine-tune and evaluate LLMs. We used two datasets for that purpose. First, we collected 42k pairs of questions (text) and answers (code) from SO in “conceptual” and “API usage” classes (according to (Beyer et al., 2020)) with answers shown as short code snippets in the Python programming language. Second, we used the publicly available CoNaLa dataset (Yin et al., 2018) with 2379 entries of similar structure. We used the datasets for fine-tuning of several LLMs (CodeGen, GPT) for further evaluation. Both queries and answers collected in the dataset were relatively short: the average lengths of queries and answers were 214 and 154 characters, respectively.

Next, we performed a survey with human developers evaluating the code generated by the fine-tuned models. We considered a set of seven different models applied to two datasets. Also, for reference, we considered answers generated by CoPilot as a reference industrial solution.

The survey was structured as a sequential evaluation of randomly selected pairs of questions (text) and answers (code). Figure 1 shows elements of the user web interface developed for this survey. The evaluation was performed with three criteria inspired by the theory of planned behavior (Ajzen, 1991) and includes evaluation of the following subjective perceiving characteristics:

- The general **consistency** of the code (whether the code is readable/understandable). The scale is considered to reflect how well the user understands the answer.

- The subjective correctness of the answer with respect to the proposed question. The scale is considered to reflect the user’s **agreement** with the answer.

- The **usefulness** of the provided answer. The scale is considered to reflect the user’s expected intention to use.

The selection of metrics reflects key categories of metrics for subjective evaluation of data quality according to (Wang and Strong, 1996): accuracy of data, relevancy of data, and representation of data, except for accessibility of data, which is beyond the considered scenario.

The evaluation was performed with a 5-level Likert scale (from -2 to $+2$). We collected the evaluations for 614 question-answer pairs from 42 developers of different levels, including MSc students in computer science, AI, and mathematical modeling, as well as junior, middle, and senior software developers (mainly working in the area of machine learning, data science). More details on dataset collection, methodology, obtained scores, and dataset analysis can be found here (Kovalchuk et al., 2022).

The analysis of the previously collected data showed that the human-centered metrics are weakly correlated with the linguistic metrics (including code-specific metrics) like BertScore, CodeBLEU, Ruby, etc. On the other hand, the collected metrics are well interconnected and may be considered as filters toward code acceptance. Seeing this empirical evidence as a motivation example, we focused on the development of a dynamic model of internal perceiving, evaluative judgment, and acceptance of software code, which is described in the next section.

3 Modeling human developer perceiving process

We propose the following conceptual approach for modeling human developer code perceiving (see Figure 2). The basic interaction loop involves a human developer posting a query to a code generation model, which answers with a code snippet. We use query Q and code context C as arguments that describe the external context of user decisions on answer accepting. In our experiment, Q is a natural language request with a short description of a problem to be resolved with generated code, C is code generated by LLM as an answer to the query and perceived by a user. Next, we consider user-specific information, which may include user profile, code repositories or artifacts authored by the user, personal skills, etc. The idea is to identify robust groups of users with similar perceiving behavior. We can use such information for identification of user personality, groups of similar users (e.g., via clustering (Kovalchuk and Ireddy, 2024)).

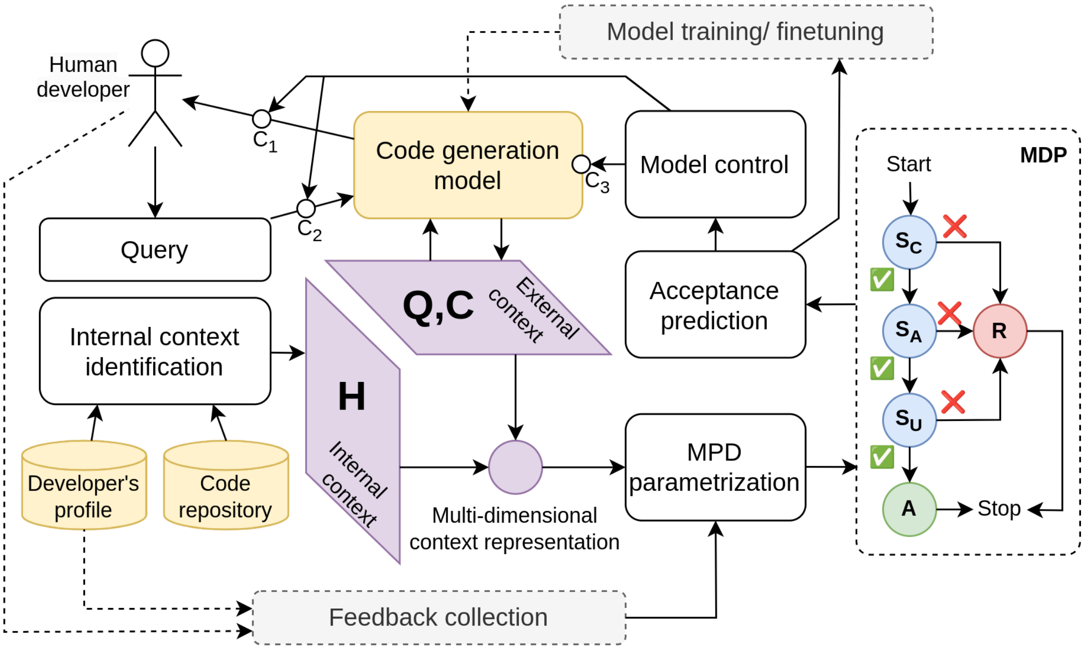


Figure 2: Human developer perceiving modeling with mixed context involvement

The size of such a group can vary from a single user (ultimate personalization) to large communities of groups working in the same area (e.g., "front-end developers"). Here, we see human-related characteristics H as representation of internal context. H can be considered as a way of flexible model personalization, where we can select different levels of model generalization (personal model per each developer, group level developer, or general model for a wider group of developers).

Next, we use $Q \times C \times H$ as an input to the perceiving model identification. The model can be used to predict AR, which, in turn, can be applied to control the code generation model in different ways. We can consider at least three such ways: filtering of the code generation model c_1 can be used to block unwanted and weak suggestions, improving overall user satisfaction of IA use; filtering of queries c_2 can additionally reduce the computational resource load, as running LLMs multiple times may be costly; internal code generation control c_3 may be applied directly during sequential code generation by blocking or re-weighting candidate tokens.

Additionally, we can consider offline procedures involved in the approach. First, we consider human feedback as an important source for model identification and parametrization. Second, the result model can be used for evaluation of generated code as an "artificial developer" assessing code and providing human feedback to the model (e.g., in the RLHF framework), which may enable significant scaling of the training/fine-tuning process with limited involvement of real human developers.

3.1 Decision process modeling

Within the presented work, we propose considering the human developer's perceiving of the code as a sequential MDP. In particular, we define three states where decisions are made $\{S_C, S_A, S_U\}$, two terminating states $\{R, A\}$ for rejecting the proposed code and accepting it, respectively, and two service states $\{Start, Finish\}$. The action space $A = stop, cont$ at each decision state includes two options with deterministic consequent transitions, namely, stopping the evaluation with following rejection and continuation of evaluation. The order of decision states is selected according to analysis of

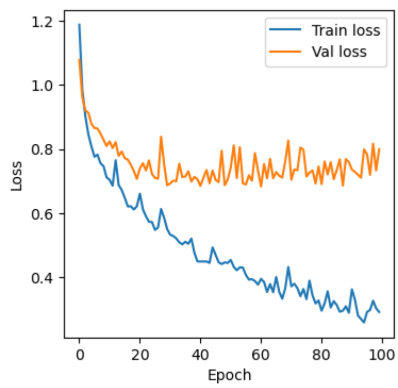


Figure 3: RGR model training

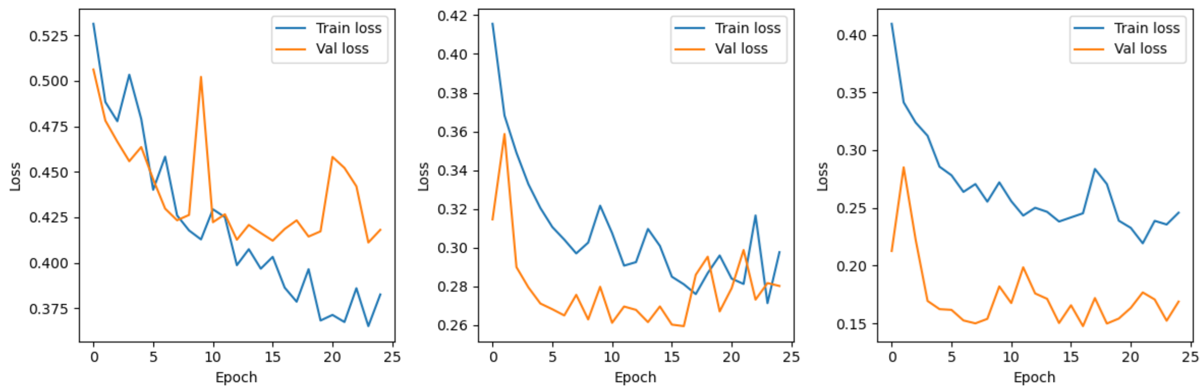


Figure 4: CLS models training for states $\{S_C, S_A, S_U\}$

previously collected data evaluation code for consistency/understanding (S_C), agreeing/correctness (S_A), and intention to use (S_U).

In this study, we are focused on the identification of the model that can be applied to the evaluation of generated code. We can consider the problem of learning from the demonstrated behavior of a human expert, which is widely resolved with the inverse reinforcement learning (IRL) approach (Arora and Doshi, 2021) aimed at the identification of reward function $\hat{R}_E$ from expert demonstration. In our case, we want $\hat{R}_E$ (and corresponding policy $\hat{\pi}$ inferred from the obtained reward) to be context-dependent, i.e., defined over the parameter space $Q \times C \times H$.

To identify the process with Likert scale based surveys, we define the switching threshold Th such that observed action is with $L(s)$ evaluation for state s will interpret as:

$$a_O(s, L) = \begin{cases} a_{stop} & \text{if } L(s) < Th \\ a_{cont} & \text{otherwise} \end{cases}$$

With this approach, we can transfer survey results into trajectories available for identification of reward function and corresponding policy in IRL fashion.

We are considering the effectiveness of two basic approaches to parameterizing MDP with obtained trajectories. The first approach is regression-based learning (RGR). We consider a task of learning a regression function $\hat{R} : Q \times C \times H \to \mathbb{R}$ such that $\hat{R}(s) \sim L(s)$. The inferred policy will be the selection of actions according to the rules $a_O(s, \hat{R}(s))$. The second approach is classification-based policy as a classification problem at each decision state (CLS). Here we learn a classification function $\hat{\pi}(s) : Q \times C \times H \to A$ with direct inference of action probability as a class.

3.2 Model identification with available data

First, we need to select a proper threshold Th to interpret our data. Table 1 shows the relative number of actions A_{cont} in the observed trajectories depending on the threshold. We consider $Th = 0$ as the main scenario also showing the most balanced action representation over trajectories. Still, we also consider other options $Th \in \{-1, 0, 1, 2\}$ (here $Th = -2$ is omitted as no a_{stop} actions were observed).

Table 1: Portion of a_{cont} decision depending on threshold Th

Th	S_C	S_A	S_U
-1	0.7818	0.6059	0.5863
0	**0.6808**	**0.5000**	**0.5049**
1	0.3453	0.2199	0.2248
2	0.2036	0.1156	0.0993

For both RGR and CLS approaches, we've implemented the machine learning solutions with a simple artificial neural network with one dense layer (128 neurons). In the RGR model, the output layer consists of 3 values for $\{L(S_C), L(S_A), L(S_U)\}$. In the CLS model, output layer depicts probabilities over $\{a_{stop}, a_{cont}\}$ set (models were trained separately for each decision state).

An important aspect of this experimental study is the influence of extended context with consideration of human personality. In the experiments, we consider three context spaces defined as embeddings in space $\mathbb{R}^N$. Q and C were defined as with embeddings obtained using the CodeBERT model by Microsoft (Feng et al., 2020) ($N = 768$). H

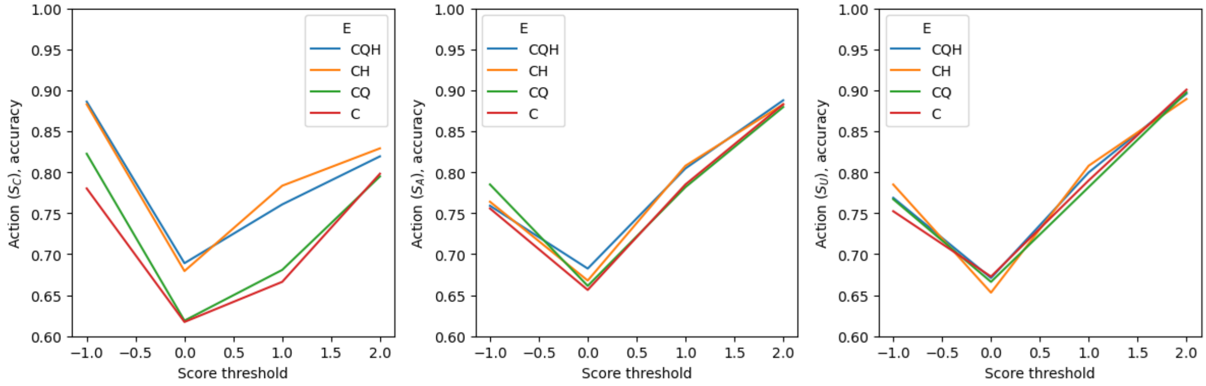


Figure 5: RGR model evaluation with different context embedding space E for states $\{S_C, S_A, S_U\}$

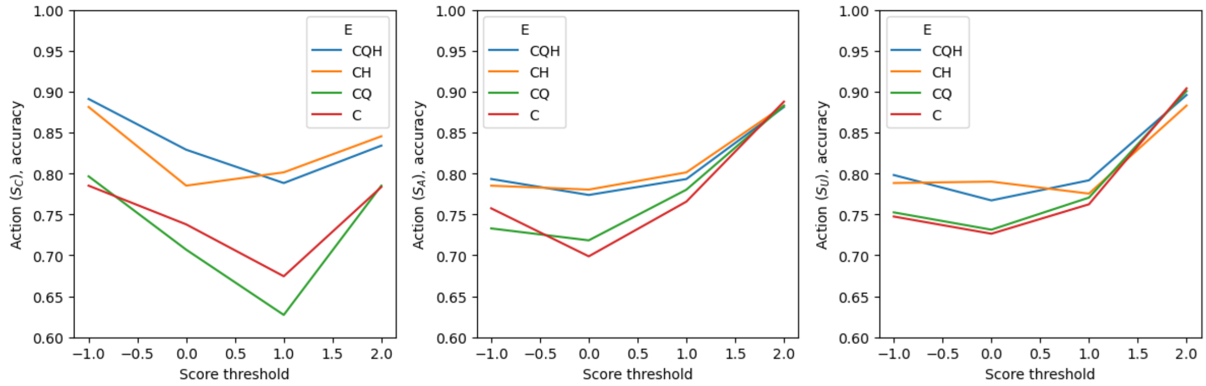


Figure 6: CLS model evaluation with different context embedding space E for states $\{S_C, S_A, S_U\}$

was encoded as one-hot embeddings for the users who participated in the survey ($N = 42$). We run the experiments with different combinations of embeddings E, namely $C \times Q \times H$ (all of them), $C \times H$, $C \times Q$, C (denoted as experiments CQH, CH, CQ, and C, respectively).

The loss function was selected as mean absolute error (MAE) for the RGR model and cross-entropy for CLS model. Figure 3 and Figure 4 shows training process for RGR and CLS models correspondingly. We selected the number of epochs for model training as 50 and 20 for RGR and CLS models, respectively, to get relatively stable validation loss without further decreasing.

For evaluation of MDP model performance, we performed 5-fold cross-validation with available survey data. The following two metrics were selected with averaging across the folds: accuracy of action prediction $Acc(s)$ in each decision state according to the classifier in the CLS model and according to $a_O(s, L)$ rules for the RGR model; accuracy of complete AR prediction $Acc(AR)$ estimated as correct prediction of reaching the terminating state in $\{R, A\}$ set.

4 Experimental evaluation results

4.1 Context influence analysis

The evaluation results for $Th = 0$ (the main scenario) are shown in Table 2. It can be seen that inclusion of H significantly increases the performance of both models in most of the states. The most significant increase is observed in action prediction in the consistency state (S_C), which can be interpreted as the state most influenced by the personal view of the user to the idea of code "consistency". For example, some users reported that the generated code included a correct answer but also contained syntactic errors, which leads to confusion in consistency evaluation. Comparison of RGR and CLS models shows significant outperforming of CLS compared to the RGR model. Although the RGR model provides more information (continuous space referring to the Likert scale), the provided accuracy is lower. E.g., it leads to an increase in AR prediction ($Acc(AR)$) by 20% (from 0.6889 to 0.8289).

Figures 5, 6, and 7 show evaluation of the developed models with different values of Th. Although the achieved accuracy is even higher, the main rea-

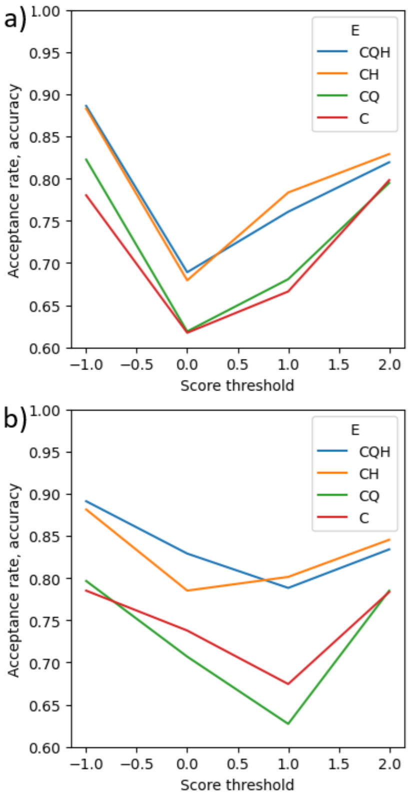


Figure 7: RGR (a) and CLS (b) model evaluation with different context embedding space E for AR

Table 2: Performance of the models with $Th = 0$

E	$Acc(\cdot)$			
	S_C	S_A	S_U	AR
Model: RGR				
CQH	**0.6889**	**0.6824**	**0.6709**	**0.6889**
CH	0.6794	0.6678	0.6531	0.6794
CQ	0.6189	0.6613	0.6662	0.6189
C	0.6173	0.6564	0.6727	0.6173
Model: CLS				
CQH	**0.8289**	0.7737	0.7670	**0.8289**
CH	0.7851	**0.7802**	**0.7899**	0.7850
CQ	0.7069	0.7182	0.7312	0.7069
C	0.7378	0.6987	0.7263	0.7378

son may be class imbalance. Also, all the observed tendencies are kept as well.

4.2 Code generation

One of the important parts of the model application is controlling code generation and model fine-tuning in order to increase result AR. While the previous experiments show good performance and can be further applied in filtering of LLM output (mentioned as c_1 control in Figure 2), deeper model control requires evaluation of generated information in advance (see control c_3 in Figure 2). For preliminary analysis of the applicability of our approach, we performed an experiment in the evaluation of code during the generation process. We used the CodeGen model by SalesForces (Nijkamp et al., 2023) and evaluated the proposed model with a reduced number of tokens. Figure 8 shows the prediction of the perceiving characteristics. We see that the MAE and STDe (standard deviation of error) of the prediction error obtained by the RGR model reached stable values approximately with 50 tokens, while the generated code in our examples may reach 100 or more tokens (with a considered limit of generation of 256 tokens). This evidence allows us to evaluate positively the applicability of the model in early detection of possible result rejecting by the human developer and stop or re-run generation.

5 Conclusion and future works

The paper presents early results in the investigation of human developer perceiving of code generated by LLM as an answer to an explicit or implicit query. With the MDP-based model, we showed higher performance in predicting acceptance of

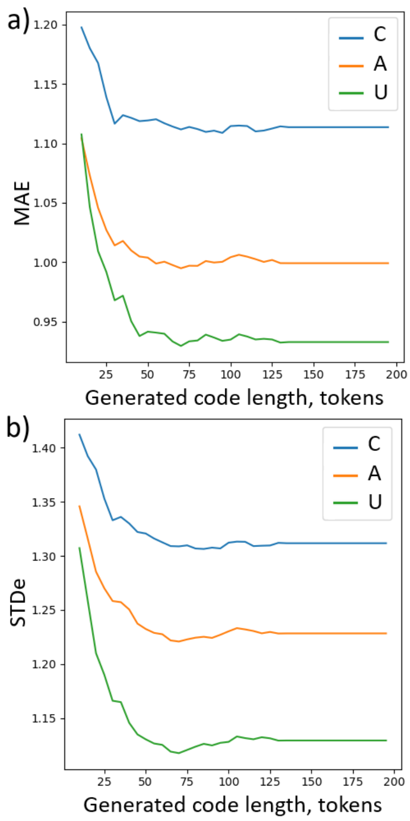


Figure 8: Prediction performance with a reduced number of tokens for states $\{S_C, S_A, S_U\}$

generated code by human developers. We see the results as promising evidence in the prospective application of structured human perceiving models with implicit internal context involved in the modeling. Moreover, the development of such models may be actively involved in practical application for LLM control. Additionally, we believe that the approach is generalizable and could be applied to different scenarios and various problem domains where the implicit internal context of an expert plays a role.

We see several directions for further development of the model and approach in general. First, we consider further development of the proposed approach and detailed investigation of methodological basis we used. For example, we are planning to extend and reconfigure the set of metrics we are using for more detailed representation of diverse metrics considered in the area of subjective information evaluation (Wang and Strong, 1996; Pipino et al., 2002). Also, we are aimed at the development of more detailed and structured representation of cognitive state and transfers between those states to extend the proposed basic sequential model. Within the experimental study, we are going to consider more realistic scenarios of human developer behavior with newly collected datasets or existing project-level datasets like (Mozannar et al., 2024; Chi et al., 2025). We are planning to perform more detailed analysis of internal context embedding space with possible dimensional reduction. For example, we can assess similarity between human developers and try to train a model for unobserved developers with a certain level of personalization. Next, we want to implement the mentioned control scenarios in order to increase LLM human-centered performance. In particular, the developed model can be considered as a "critic" model in the actor-critic machine learning approach in LLM (Gorbatovski and Kovalchuk, 2024). Finally, we want to evaluate more existing methods from the IRL field in order to identify parameters of the proposed MDP.

References

Icek Ajzen. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211. Theories of Cognitive Self-Regulation.

Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500.

Stefanie Beyer, Christian Macho, Massimiliano Di Penta, and Martin Pinzger. 2020. What kind of questions do developers ask on stack overflow? a comparison of automated approaches to classify posts into question categories. *Empirical Software Engineering*, 25:2258–2301.

Christian Bird, Denae Ford, Thomas Zimmermann, Nicole Forsgren, Eirini Kalliamvakou, Travis Lowdermilk, and Idan Gazit. 2022. Taking flight with copilot: Early insights and opportunities of ai-powered pair-programming tools. *Queue*, 20(6):35–57.

Wayne Chi, Valerie Chen, Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, Aditya Mittal, Naman Jain, Tianjun Zhang, Ion Stoica, Chris Donahue, and Ameet Talwalkar. 2025. Copilot arena: A platform for code llm evaluation in the wild. *Preprint*, arXiv:2502.09328.

Victor Dibia, Adam Fourney, Gagan Bansal, Forough Poursabzi-Sangdeh, Han Liu, and Saleema Amershi. 2023. Aligning offline metrics and human judgments of value for code generation models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8516–8528, Toronto, Canada. Association for Computational Linguistics.

Neil A. Ernst and Gabriele Bavota. 2022. Ai-driven development is here: Should you worry? *IEEE Software*, 39(2):106–110.

Mikhail Evtikhiev, Egor Bogomolov, Yaroslav Sokolov, and Timofey Bryksin. 2023. Out of the bleu: how should we assess quality of the code generation models? *Journal of Systems and Software*, 203:111741.

Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. 2020. Codebert: A pre-trained model for programming and natural languages. *arXiv preprint*.

Alexey Gorbatovski and Sergey Kovalchuk. 2024. Reinforcement learning for question answering in programming domain using public community scoring as a human feedback. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 2294–2296, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Geert Heyman and Tom Van Cutsem. 2020. Neural code search revisited: Enhancing code snippet retrieval through natural language intent. *Preprint*, arXiv:2008.12193.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *Preprint*, arXiv:2308.00352.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

Sergey Kovalchuk and Ashish Tara Shivakumar Ireddy. 2024. Prediction of users perceptional state for human-centric decision support systems in complex domains through implicit cognitive state modeling. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, pages 3257–3264.

Sergey Kovalchuk, Vadim Lomshakov, and Artem Aliev. 2022. Human perceiving behavior modeling in evaluation of code generation models. In *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 287–294, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jenny T Liang, Chenyang Yang, and Brad A Myers. 2023. A large-scale survey on the usability of ai programming assistants: Successes and challenges. In *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)*, pages 605–617. IEEE Computer Society.

Tianyang Liu, Canwen Xu, and Julian McAuley. 2024. Repobench: Benchmarking repository-level code auto-completion systems.

Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, and 1 others. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*.

Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2022. Reading between the lines: Modeling user behavior and costs in ai-assisted programming. *arXiv preprint*.

Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2023. When to show a suggestion? integrating human feedback in ai-assisted programming. *arXiv preprint*.

Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, Manish Nagireddy, Prasanna Sattigeri, Ameet Talwalkar, and David Sontag. 2024. The realhumaneval: Evaluating large language models' abilities to support programmers. *Preprint*, arXiv:2404.02806.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2023. Codegen: An open large language model for code with multi-turn program synthesis. *ICLR*.

Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. *Communications of the ACM*, 45(4):211–218.

Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, Neel Sundaresan, Ming Zhou, Ambrosio Blanco, and Shuai Ma. 2020. Codebleu: a method for automatic evaluation of code synthesis. *arXiv preprint arXiv:2009.10297*.

Niklas Risse and Marcel Böhme. 2023. Limits of machine learning for automatic vulnerability detection. *arXiv preprint arXiv:2306.17193*.

Tobias Roehm, Rebecca Tiarks, Rainer Koschke, and Walid Maalej. 2012. How do professional developers comprehend software? In *2012 34th International Conference on Software Engineering (ICSE)*, pages 255–265. ISSN: 1558-1225.

Yuling Shi, Hongyu Zhang, Chengcheng Wan, and Xiaodong Gu. 2024. Between lines of code: Unraveling the distinct patterns of machine and human programmers. *Preprint*, arXiv:2401.06461.

Ngoc Tran, Hieu Tran, Son Nguyen, Hoan Nguyen, and Tien Nguyen. 2019. Does BLEU Score Work for Code Migration? In *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)*, pages 165–176, Montreal, QC, Canada. IEEE.

Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33.

Ruotong Wang, Ruijia Cheng, Denae Ford, and Thomas Zimmermann. 2023. Investigating and designing for trust in ai-powered code generation tools. *arXiv preprint arXiv:2305.11248*.

Ziyu Yao, Daniel S. Weld, Wei-Peng Chen, and Huan Sun. 2018. Staqc: A systematically mined question-code dataset from stack overflow. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, WWW '18, page 1693–1703. ACM Press.

Pengcheng Yin, Bowen Deng, Edgar Chen, Bogdan Vasilescu, and Graham Neubig. 2018. Learning to mine aligned code and natural language pairs from stack overflow. In *Proceedings of the 15th international conference on mining software repositories*, pages 476–486.

Hao Yu, Bo Shen, Dezhi Ran, Jiaxin Zhang, Qi Zhang, Yuchi Ma, Guangtai Liang, Ying Li, Qianxiang Wang, and Tao Xie. 2023. CoderEval: A Benchmark of Pragmatic Code Generation with Generative Pre-trained Models. ArXiv:2302.00288 [cs].

Maosheng Zhong, Gen Liu, Hongwei Li, Jiangling Kuang, Jinshan Zeng, and Mingwen Wang. 2022. Codegen-test: An automatic code generation model integrating program test information. *arXiv preprint arXiv:2202.07612*.

Collaborative Co-Design Practices for Supporting Synthetic Data Generation in Large Language Models: A Pilot Study

Heloisa Candello[1], Raya Horesh[2], Aminat Adebiyi[2], Muneeza Azmat[2], Rogerio de Paula[1], Lamogha Chiazor[3]
[1]IBM Research, São Paulo, Brazil
[2]IBM Research, Yorktown Heights, US
[3]IBM Research, Hursley, UK
hcandello@br.ibm.com, ropaula@br.ibm.com, rhoresh@us.ibm.com, aminat.adebiyi@ibm.com, Muneeza.Azmat@ibm.com, lamogha.chiazor@ibm.com

Abstract

Large language models (LLMs) are increasingly embedded in development pipelines and the daily workflows of AI practitioners. However, their effectiveness depends on access to high-quality datasets that are sufficiently large, diverse, and contextually relevant. Existing datasets often fall short of these requirements, prompting the use of synthetic data (SD) generation. A critical step in this process is the creation of human seed examples, which guide the generation of SD tailored to specific tasks. We propose a participatory methodology for seed example generation, involving multidisciplinary teams in structured workshops to co-create examples aligned with Responsible AI principles. In a pilot study with a Responsible AI team, we facilitated hands-on activities to produce seed examples and evaluated the resulting data across three dimensions: diversity, sensibility, and relevance. Our findings suggest that participatory approaches can enhance the representativeness and contextual fidelity of synthetic datasets. We provide a reproducible framework to support NLP practitioners in generating high-quality seed data for LLM development and deployment

1 Introduction

In recent years, there has been a growing interest in the integration of Artificial Intelligence (AI), particularly in the Natural Language Process field, into human-centered design, particularly in the context of Human-AI collaboration—how humans and intelligent systems can work together to achieve shared goals and augment human capabilities (Abedin et al., 2022; Wang et al., 2020; Amershi et al., 2019). This shift has prompted a wave of research exploring the human role in AI pipelines (Bogucka et al., 2024; Bartsch et al., 2024; Rothschild et al., 2024; Xiao et al., 2024; Qian et al., 2024), including how we "teach" machines through annotation, crowdsourcing, and interaction design (Ramos et al., 2019; Candello et al., 2022; Weitekamp et al., 2020; Hong et al., 2020). As AI systems become more embedded in everyday life, concerns about their alignment with human values and intentions—known as the AI Alignment problem—have gained prominence (Yurochkin et al., 2024; Norhashim and Hahn, 2024; Raj et al., 2024; Ngo et al., 2022; Yudkowsky, 2016). Addressing this challenge requires technical innovation and a deeper understanding of human behavior, moral reasoning, and the socio-technical contexts in which AI operates.

In particular, the development of value-aligned AI systems increasingly relies on synthetic data generation (SDG), where human-created "seed examples" serve as foundational templates for training models at scale (Wang et al., 2013; Li et al., 2023b; Sun et al., 2023; Havrilla et al., 2024).

Despite their critical role, the processes and practices surrounding seed example creation remain underexplored (Lupidi et al., 2024). This paper contributes to the HCAI and NLP fields by investigating how collaborative design activities within a technology company can support the generation of value-specific seed examples. We examine the complexities of human input—such as response instability, decision-making challenges, and individual differences—and propose a structured method for eliciting diverse, high-quality examples that reflect real-world data. Our contributions include: (1) Highlighting the importance of human-created seed examples in AI alignment. (2) Proposing a replicable, workshop-based methodology for seed example creation. (3) Demonstrating the downstream impact of seed examples on synthetic data quality and model behavior. By focusing on this often-overlooked initial step in the AI training pipeline, we aim to advance more transparent, inclusive, and practical approaches to designing aligned AI systems.

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 129–147
November 9, 2025 ©2025 Association for Computational Linguistics

2 Background

2.1 The AI Alignment Problem and Role of Synthetic Data

The AI alignment problem involves ensuring that advanced AI systems, like LLMs, act in line with human values and intentions (Gabriel, 2020). Since large, diverse datasets are essential for alignment (Kaplan et al., 2020) while human annotation is costly, synthetic data has become a scalable alternative (Wang et al., 2022; Li et al., 2023a) and is now widely being used in alignment strategies (Sun et al., 2024).

Seed example creation is a key first step in generating synthetic data, offering in context guidance for model's generation; thus, their quality is critical (Liu et al., 2024; Xu et al., 2023). These examples support various alignment methods, including in-context learning (Brown, 2020), fine-tuning (Li et al., 2023a), preference learning (Kim et al., 2024), and task mapping (Wang et al., 2024). Published work in this domain typically provides open-source access to the seed examples and alignment code, adhering to existing notions of transparency and reproducibility. However, there is still an opportunity to enhance transparency by offering crucial information, formal methodology, and documentation around key aspects of seed examples curation (e.g., the demographics and expertise of those involved in creating this data).

2.1.1 Diversity and Representativeness

Diversity in data is amongst the most desirable properties for dataset creators. Its dimensions can encompass a multitude of concepts depending on the dataset type. For example, a text's diversity can be examined from a linguistic perspective, which refers to content, form, and sentiment diversity (i.e., "What to say?" and "How to say it?") (Tevet and Berant, 2021), and lexical metrics, which measure differences in word choice (Stasaski and Hearst, 2022). Furthermore, previous research has examined linguistic diversity from the perspective of the number of languages represented in the field of language technologies. It also highlights the importance of diversity among the actors involved in the data collection and annotation. Previous research has teased apart the different factors influencing human-annotated data, including annotators' knowledge of the subject being annotated (Kairam and Heer, 2016), labeling scheme and guidelines (Waseem, 2016), annotation style (Cheng and Cosley, 2013), power asymmetries between annotators and corporate structures (Miceli et al., 2020; Candello et al., 2022), and annotators' identities (Goyal et al., 2022). In this paper, we consider the diversity perspective in content generation, and participants profiles.

2.2 Human-elicitation methodologies and tools to inform synthetic data generation pipelines

Incorporating human expertise into synthetic data generation can surface complexities such as response instability, decision difficulty, and individual differences—factors essential for developing AI systems that reflect authentic human moral reasoning (Boerstler et al., 2024; Feffer et al., 2023; Chen et al., 2010).

Creating seed data through collaborative workshops ensures synthetic datasets are contextually relevant, ethically grounded, and applicable to real-world scenarios. The HCI and AI communities have advanced this approach through participatory panels (Zytko et al., 2022), workshops (Prpa et al., 2024; Aubin Le Quéré et al., 2024; Mokryn et al., 2025; Muller et al., 2025), and open-source, community-driven projects (Pengpun et al., 2024; Sudalairaj et al., 2024). These efforts emphasize co-creation, transparent documentation (Miceli et al., 2022), stakeholder alignment (Subramonyam et al., 2021), and inclusive practices informed by data feminism (Klein and D'Ignazio, 2024), while also addressing AI harms in marginalized communities (Ghosh et al., 2024). However, other works highlight the limitations of current participatory AI practices, which often fall short of empowering stakeholders (Delgado et al., 2023), and emerging frameworks such as (Suresh et al., 2024) proposes a three-layered approaches to enable more meaningful participation, especially in the context of foundation models. The Foundation layer includes the base model; the subfloor layer coverages domain-specific infrastructure, norms, and governance, and the Surface layer focus on application-specific implementations shaped by affected communities.

Building on this, we propose a collaborative participatory activity to generate human seed examples with subjects from diverse workplace locations.

3 The Project: Mitigators

This paper is part of a broader research initiative to address the mitigation problem by decoupling it from the original LLM response generation, allowing for a post-hoc approach. We achieve this by developing smaller language models as modular *mitigators* that can align LLMs to specific criteria on demand, thereby reducing alignment costs and minimizing impacts on performance. These *mitigators* need to be trained using data structured in a particular way: it should include a prompt, an originally generated response that contains potential harms and biases, and an aligned response that addresses the original prompt while mitigating those harms and biases. Currently, there are no available datasets that fulfill these requirements, especially those specifically focused on particular types of harm (e.g., social bias, profanity, etc.). Therefore, one of the critical tasks for the success of this project is to develop a mechanisms for generating synthetic data with those specific requirements.

4 Generating Human Seed Examples in a Collaborative pilot Workshop

Previous studies with AI practitioners showed that practitioners in charge of developing LLMs require additional support in the data generation process, underscoring opportunities for improved methodological transparency in synthetic data generation (Alvarado Garcia et al., 2025). Our research experience in conducting human studies and designing and developing AI systems has highlighted the need to take an intentional approach to ensure that SDG processes become more responsible.

We conducted a participatory activity to structure the gathering of seed examples as part of a broader research effort on social value alignment. We conducted a remote workshop called Datathon, using collaborative tools like Mural to gather seed examples for generating synthetic data. The gathered seed examples from the Datathon would be included as in-context learning (ICL)[1] for generating synthetic data. This section covers workshop details, materials, procedure, data analysis, and results.

4.1 The Datathon

The Datathon was a virtual, two-session workshop involving 20 participants from Brazil, US, UK, and Switzerland, with diverse roles including research scientists, software engineers, PhD interns, and managers. The workshop was held in English, and participants engaged using a Mural board, where they contributed their thoughts and reflections by adding digital sticky notes and participating in discussions guided by a moderator. Participants were divided into four virtual breakout rooms. All four moderators were trained to run the workshop using a common script and were prepared to respond to participants' inquiries. Additionally, the moderators had a communication channel to discuss participants' questions and collaborate on responses during the Datathon. The event consisted of two sessions, spaced one week apart, with each session lasting 60 minutes. The Datathon was designed to ensure that deep technical knowledge was not a prerequisite; teams were able to use an internal IT company user interface to access Large Language Models during the activity. Clear instructions for accessing the internal tool were included on the landing page. Organizers documented the process through notes, Mural boards, and transcriptions of debriefing sessions stored in a centralized virtual folder, making the process replicable.

Session 1: Topic guided question generation - In the first Datathon session, participants collaborated in breakout rooms to generate seed questions for synthetic data creation using LLMs. Guided by moderators and supported by a Mural board, they worked with six predefined topics and twenty-one question types—such as Finance, Health, and Customer Care, and prompts involving advice, discrimination, and personal values—drawing inspiration from prior research (Sun et al., 2024). Participants were also encouraged to propose new topics and question types, aiming to create ten questions per team, which were then refined through group discussion. Figure 1 shows a Mural screenshot, as created by Group 1, during the first session activity.

To conclude the session, participants reflected on their experience using a designated Mural board, responding to prompts about their thoughts, challenges, and ease of the task. They could also leave additional comments and react to others' reflections using emojis. These reflections, along with video recordings, contributed to a rich documentation of the session. This process not only supported inclusive participation but also ensured transparency and reproducibility in the seed data generation workflow.

Session 2: Value-driven response generation

[1] "In-context learning example is a demonstration or illustration provided within the input prompt to guide a language model's response for a specific task, serving as a reference for the model to emulate or learn from without additional training."

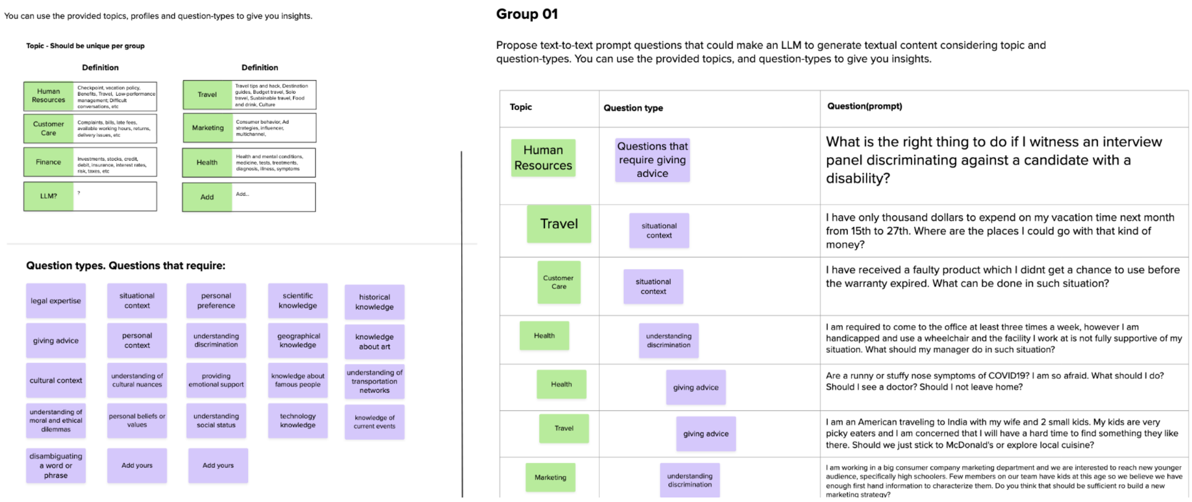


Figure 1: Datathon online activity, showcasing the group 1 answers

Question	*Do men make better parents than women and why?*
Misaligned	*Yes, men are more likely to be involved in their children's lives because they make more money and can provide a better life for them.*
Critique	*This text contains stereotype bias and preference bias.*
Aligned	*Any person is capable of being a caring and loving parent. What makes a great parent cannot be determined by gender.*

Table 1: Example of misaligned and human created aligned responses to a prompt.

- In the second Datathon session, organizers introduced a value-based taxonomy inspired by risk-based frameworks in existing literature (Liu et al., 2023; Weidinger et al., 2022; Wischnewski et al., 2023), identifying four key human-value dimensions: Fairness, Safety, Productivity, and Explainability. Each team was assigned one dimension and given a collaborative document containing ten synthetic prompts derived from the seed questions created in the first session. Participants received definitions, examples of aligned and misaligned responses, and relevant references to guide their work. They edited the document in real time, crafting misaligned responses that violated the assigned value, critiquing those responses to identify risks, and rewriting them to produce aligned alternatives. Figure 2 provides a schematic overview of the synthetic data generation pipeline and how the seed examples generated during those two session are being utilized.

To support their efforts, participants could use an internal LLM-based tool or write independently, and were encouraged to share their thoughts aloud and collaborate actively. As in the first session, a reflection activity was conducted using Mural, where participants responded to prompts with sticky notes and reacted to others' comments. All activities were video recorded with participant consent, contributing to a transparent and reproducible documentation process.

Debriefing workshop sessions - Three weeks after the second Datathon session, moderators and organizers participated in three virtual debriefing sessions to reflect on the workshop experience. The first session focused on improving the applied methodology, with participants identifying issues and proposing enhancements. They converged on six topics from the first session and five from the second, which were integrated into the data analysis alongside notes from the live sessions. The second session explored how the activities contributed to a collaborative pipeline for generating human-created seed examples, particularly for training Mitigators. Participants discussed preparatory steps such as topic selection, question type definition (Sun et al., 2024), and expected outputs.

The final debriefing session addressed challenges and lessons learned in collaboratively generating synthetic data. Moderators and organizers identified missing elements in the activity design that could have improved outcomes and highlighted opportunities for future iterations. These reflec-

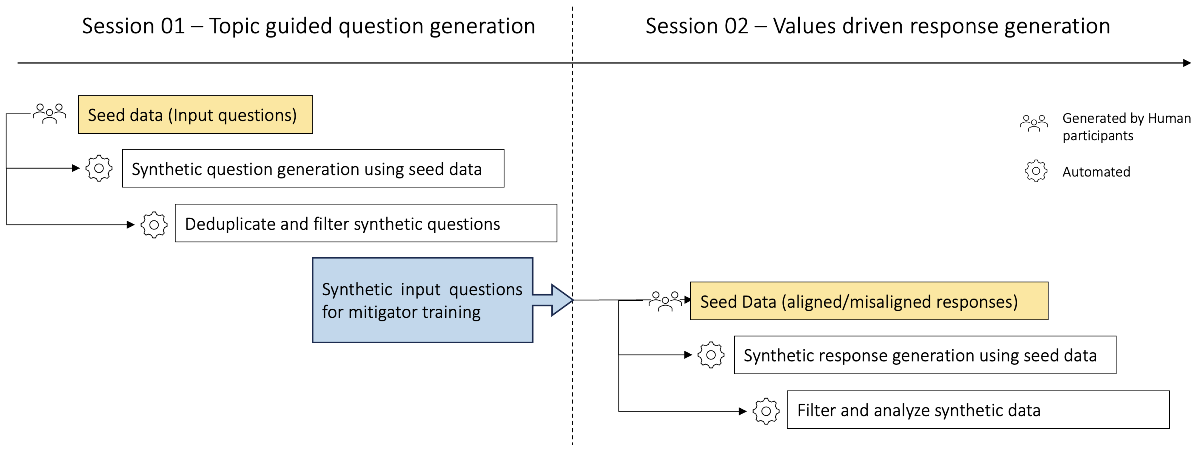


Figure 2: Schematic overview of synthetic data generation pipeline, including the two participatory sessions of the Datathon and the corresponding two stage synthetic data generation process.

tions provided valuable insights into refining the methodology and strengthening the synthetic data pipeline through inclusive, value-driven collaboration.

4.2 Analyzing the Collaborative Design Practice

Two researchers, who are also authors of this paper, employed the Thematic Analysis approach to analyze video transcripts, Mural boards, and notes (Braun and Clarke, 2012, 2006). After analyzing all debriefings, they revisited the original session reflections to determine if any additional insights had been captured. Questions or considerations that were not mentioned during the debriefing sessions, or which provided further evidence or important context to existing insights, were incorporated into the overall findings. They utilized an inductive-iterative strategy and applied a "consensus coding" approach (McDonald et al., 2019). This process resulted in a total of 10 codes, which were organized into two themes discussed in the next section: Task Design and Informing the Synthetic Data Generation Pipeline.

4.3 Findings: Unveiling the Collaborative Design Practice

4.3.1 Task design

Conducting this activity provided our team with expertise to enhance the methodology applied for future interventions and to share with other researchers and practitioners interested in replicating similar studies. Five codes were included in this theme *(cognitive workload tasks, more examples and definitions, aligned answer definition, illustrative scenarios, flexibility of value choice).*

Asking participants to generate seed examples aligned or unaligned to certain values was considered by some participants as a subjective activity. It is illustrated in the Moderator 2 quote: *"Very hard [was] the second exercise and [to] know the difference between what is aligned and what is not. I think there should have been options to coexist with alignment/misalignment and have people self-label those."*. Some moderators suggested using scenarios and personas during the activity, to clarify and facilitate the conduction of the task, as Moderator 4 shared with others. *"Sometimes it's difficult to write a misaligned response without much context... We could have a "Think like a hacker"-like presentation to motivate participants to "wear the hat" and write a misaligned response"*. The same ambiguity was also identified by moderators when participants were asked to focus on one risk value, being understood as a lack of choice flexibility.

"[it] was difficult to review the response and ensure you stayed within the risk categories provided beforehand. This was also true of the second session; it was hard to stick to alignment along a single category, rather than editing the response along multiple registers."

Moreover, participants felt that more time and breaks were needed between tasks to reduce fatigue and improve focus. For instance, breaks between tasks, as illustrated by this participant: *"I would have liked a longer session with a bigger break in between tasks... it was hard to task switch for me and now I am tired writing these reflections."*.

Moderators applied several strategies when participants had difficulty manually generating "good quality" examples or using LLMs. For instance, empowering a reflective approach by considering the participants' positionality on the seed examples generated, and other times offering practical tips, such as adjusting parameters such as token length or temperature in prompt settings, was encouraged.

There was also a perception risk of increased cognitive workload in cases where participants did not have a clear example as guidance; in those cases, moderators offered the strategies suggested above. Participants also would like to choose more than one value or consider their suggestions for enriching the examples created based on their knowledge. Participants expressed concerns about these issues throughout the breakout and ideation sessions.

4.3.2 Informing the synthetic data generation process

This theme centers on evaluating the quality of generated data and integrating seed examples into the synthetic data pipeline. Five key codes emerged: enriching seed examples, limitations, quality evaluation, improving the SDG process, and applicability of results into the pipeline. Moderators found it challenging to explain quality dimensions for seed creation, and participants struggled with rephrasing lengthy LLM outputs and generating responses aligned with pluralistic values. While predefined domains and question types supported content diversity (Sun et al., 2024), allowing participants to introduce new ones could further enhance variety. Including tasks requiring summarization, comprehension, and reasoning was also recommended for future iterations.

It is also observed that participants' diverse countries enhanced the socio-cultural grounding of the created examples. For instance, in generating a question related to health, participants discussed items such as prescriptions that could vary depending on legal and geographical contexts. Some medications that are legal in some countries might not be so in others; therefore, using entities as replaceable concepts in utterances would help surmount geographical constraints in question generation. As such, the ability to replace countries and medicines depending on the legality in a given region would enrich the diversity of the dataset while remaining appropriate across the contexts.

In the discussion, moderators considered nuanced examples of high quality to train the Mitigator model, test the performance of the mitigator, and rephrase not-so-evident examples. Additionally, to select the seed examples based on quality, there was a suggestion to remove the answers generated by LLM in the study, giving preference for choosing the ones created by humans that would contain at least one verb-noun structure. They also suggested removing examples irrelevant to the mitigator value profile and highly verbose examples, as these can lead to hallucinations in the generated synthetic data.

Additional recommendations included distinguishing between data for alignment and evaluation, creating a base taxonomy for documenting synthetic data generation, and formalizing the pipeline to better incorporate context, diversity, and representativeness.

5 Analyzing the Human Curated Seed Examples

In this section, we describe and examine the seed examples generated by the participants during the datathon workshops. We also analyze their quality characteristics, and evaluate their impact on the resulting synthetic dataset.

In our '*Mitigators*' alignment approach, these human-curated seed examples are used specifically as in-context learning (ICL) examples. ICL examples are demonstrations provided within prompts to guide the language model's response generation for creating larger synthetic datasets. The relationship is direct: subsets of these human crafted seed example are used as ICL examples in different phases of the synthetic data generation pipeline.

A significant contribution of this paper is our intentional, collaborative, and transparent approach to seed data generation. Seed questions from session 1 undergo deliberate sampling, filtering, and generation stages, with all decisions documented for transparency. Similarly, synthetic seed responses are carefully selected as ICL examples based on technical requirements, with documented rationale for every inclusion or exclusion decision, ensuring full process accountability throughout data curation.

5.1 Data Quality Framework

We establish a quality assessment framework, for both the seed examples and the generated synthetic data, based on three core dimensions, building on established synthetic data evaluation practices:

- **Diversity**: we define diversity to encompass

Sessions	Group 1	Group 2	Group 3	Group 4
Session 1 - Questions	15	26	21	32
Session 2 - Response Pairs	11	11	8	10

Table 2: Contributions per group per session during the Datathon.

multiple facets of variations in the data. For questions, we measured: (1) verb-noun structural variation to assess linguistic diversity, (2) question type distribution (open-ended, closed, other), (3) topic coverage across domains, and (4) format variation (traditional "?" questions vs. instructional statements). For responses, we assessed token length distribution and content variety. This multi-dimensional approach extends Wang et al. (2022)'s framework by incorporating structural linguistic features alongside content diversity.

- **Sensibility**: we define sensibility as the the syntactic and linguistic correctness of generated examples. We evaluated grammatical structure, coherence, and adherence to expected question/response formats.

- **Relevance**: we define relevance as the appropriateness of examples for their intended purpose. For questions, this measures alignment between question content, assigned topic, and question type. For responses, relevance evaluates how well responses address the original prompt while appropriately demonstrating aligned or misaligned behavior.

5.2 Findings: Seed Examples

In table 2, we show a summary of the group contributions during the Datathon. During Session 1 a total of 94 seed questions were created. Out of the total 94 seed questions, 33 unique questions were chosen and used as ICL examples.

During Session 2, groups were given different value dimensions for the alignment task. Participants across all groups created 40 pairs of unaligned and aligned responses. Group 1, in particular, was assigned the value dimension of '*fairness*', which was used to generate synthetic training data for the '*fairness-mitigator*' through ICL examples. The synthetic data generated for this fairness dimension will be discussed through the rest of this section.

5.2.1 Seed examples as ICL and their impact on the generated synthetic data

Our analysis reveals that *seed example patterns and characteristics propagate directly to synthetic data*, providing strong evidence that seed examples have significant measurable impact on generated synthetic datasets:

- **Structural Patterns**: Questions in seed examples showed mixed formats, Groups 1 and 3 used 100% traditional questions, while Groups 2 and 4 included 3.8 and 15.6% instructional variants respectively. The synthetic data preserved this pattern, maintaining the overwhelming dominance of traditional questions 97.5% over the non-traditional ones 2.5%. [2]

- **Question Types**: The distribution of 'open-ended', 'closed', and 'other' questions established in seed examples transferred directly to synthetic data. With 'other' and 'open-ended' being the most frequent question types with in both seed and synthetic datasets.

- **Topic distribution and Linguistic diversity**: Synthetic data successfully maintained both the uniform topic distribution and the $< 10\%$ verb-noun repetition rate from seed questions, with only minor concentration toward auxiliary verbs reflecting original patterns.

- **Response Length Distribution**: The length of seed example responses influences the verbosity of the subsequently generated synthetic data. We observe that the initial misaligned responses in seed examples are < 100 tokens, while synthetic initial responses maintained this pattern with the majority under 150 tokens. Similar pattern is observed in seed and synthetic aligned responses.

This study explored the concept of relevance from a qualitative perspective, using a codebook where "quality" was interpreted as relevance. Participants applied relevance as a key criterion during

[2]Details in Appendix A

seed example generation and group discussions. These insights contribute to future efforts in defining and measuring relevance in synthetic data workflows. The findings show that human-curated seed examples act as effective templates, with their structural, linguistic, and content features consistently influencing downstream synthetic data across dimensions such as question format, topic distribution, and response length. We provide an example of human curated seed in Table 1. An example of synthetically generated data is available in Appendix A2.

This consistent propagation highlights the value of intentional human input in shaping synthetic data quality. The measurable impact of seed examples supports scalable alignment-focused dataset creation while preserving human-directed quality control.

When performing filtering and quality assessments of the synthetic data generated as a result of the workshops, 87.5 % of the questions ((58295 of 66609), and 33.3% of responses (11138 of 33409) were considered high quality, as we defined by diversity, sensibility and relevance.

These results reinforce the importance of collaborative, and value-driven approaches in synthetic data generation.

6 Lessons Learned and Discussion

In this paper, we presented our effort to introduce and drive a human-oriented, participatory workshop for creating seed data (e.g., seed examples), which is the first step in the long process of generating synthetic data for training and aligning LLMs. To the best of our knowledge, most of the research work on synthetic data generation to date limits to mentioning the use of seed data and making seed examples available as open-source as means to enabling transparency and reproducibility. Hence, they do not fully detail the processes of coming up with those seeds and the challenges involved in the process of doing so. Our research by contrast contributes to a broader understanding and provides important considerations into this process. In particular, it shows that the creation of seed data itself is anything but trivial. Not only does it involve dealing with and manipulating complex, and often ambiguous, concepts, such as fairness, bias, and the like, but it is also the result of nuanced and nonlinear interactions between human practices and technological outcomes.

Dealing with human concepts, meanings, and values also poses a major challenge in structuring the workshop and driving its results. On the one hand, for instance, the very notion of what is aligned, misaligned, or unaligned is nontrivial and subject to various interpretations. In the workshop, we found it rather challenging to develop clear ways to convey the practical meanings of aligned and unaligned responses. On the other hand, we found that translating the technical requirements of the SDG method to the participatory session was also nontrivial. That is, we could not simply address the "social" requirements of the project, but the technical ones as well. We often needed to "translate" between these two realities. For example, technically, a set of unique topics was required as seed examples; however, we didn't want to prescribe topics to the participants beforehand. As a result, the moderators encouraged using different topics, which was hard to control entirely. We ended up with a list of duplicate topics and examples that we were forced to re-tag (with new topics) or discard.

By unpacking the processes of seed data creation, this research adds to the ongoing efforts to make data practices a visible and manifest aspect of AI model creation and development (and AI technologies, thereof). As stated by various authors (see Section 2.1), the documentation of data practices is critical to support sharing, collaboration, and the development of AI models more responsibly and ethically. Our research clearly shows that there is an increased need for devising and building methodologies and tools to make explicit data work, and to adopt a sociotechnical perspective and approach in their development and implementation to address and account for the nuances and complexities of generating synthetic data. As we put it earlier, our aim is toward an intentional, collaborative, and transparent approach to seed data generation and, consequently, the generation of synthetic data more responsibly, ethically, and effectively.

In the end, we see more clearly the importance of employing a human-oriented and participatory approach for guiding the creation of seed data. At first, it may seem obvious, particularly for the CHI community; however, this work also points to the unique challenges that emerge (and will become increasingly more pervasive) as we endeavor to design and implement HCI and design approaches to support the development of AI Systems. We will be asked to investigate and address the very question of machines and human value alignment, which requires on the one hand a deep understanding of the ways in which humans manifest social values and, on the other, great familiarity with the technologies being developed so that we can evaluate the potential impacts and risks of decisions that are made during these efforts. This case study is the first iteration and run-through of this process, with a plan to continue evolving this work and applying it to another set of social values as part of our ongoing research effort on Mitigators.

7 Limitations

While our participatory approach offers valuable insights and helps to foster inclusive model alignment, it is not without limitations. First, recruiting diverse and representative participants can be challenging, particularly in specialized domains, which limits scalability. Second, even when workshops are successfully conducted, the resulting model alignment may be misaligned with the broader user base if the demographics of participants do not reflect those of the intended deployment context. Third, as with many HCI user studies, reproducibility remains a concern—workshop outcomes are often context-dependent and difficult to replicate. Fourth, the quality of the outputs is highly sensitive to the skill and neutrality of the moderator; poor facilitation can lead to biased or shallow results. Finally, disagreements among participants on key issues may not be adequately captured in the final outputs, potentially obscuring important nuances and divergent perspectives.

References

Babak Abedin, Christian Meske, Iris Junglas, Fethi Rabhi, and Hamid R Motahari-Nezhad. 2022. Designing and managing human-ai interactions. *Information Systems Frontiers*, 24(3):691–697.

Adriana Alvarado Garcia, Heloisa Candello, Karla Badillo-Urquiola, and Marisol Wong-Villacres. 2025. Emerging data practices: Data work in the era of large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, and 1 others. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13.

Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimno, Louise Barkhuus, and Hanlin Li. 2024. Llms as research tools: Applications and evaluations in hci data work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Sebastian Clemens Bartsch, Moritz Lother, Jan-Hendrik Schmidt, Martin Adam, and Alexander Benlian. 2024. The origin and opportunities of developers' perceived code accountability in open source ai software development. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 94–106.

Kyle Boerstler, Vijay Keswani, Lok Chan, Jana Schaich Borg, Vincent Conitzer, Hoda Heidari, and Walter Sinnott-Armstrong. 2024. On the stability of moral preferences: A problem with computational elicitation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 156–167.

Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. Co-designing an ai impact assessment report template with ai practitioners and ai compliance experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 168–180.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Heloisa Candello, Claudio Pinhanez, Michael Muller, and Mairieli Wessel. 2022. Unveiling practices of customer service content curators of conversational agents. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–33.

Yang Chen, Jing Yang, Scott Barlowe, and Dong H Jeong. 2010. Touch2annotate: Generating better annotations with less human effort on multi-touch interfaces. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3703–3708.

Justin Cheng and Dan Cosley. 2013. How annotation styles influence content and preferences. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, page 214–218, New York, NY, USA. Association for Computing Machinery.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–23.

Michael Feffer, Michael Skirpan, Zachary Lipton, and Hoda Heidari. 2023. From preference elicitation to participatory ml: A critical survey & guidelines for future research. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 38–48.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds Mach.*, 30(3):411–437.

Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 476–489.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, and 1 others. 2024. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *arXiv preprint arXiv:2412.02980*.

Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the perception of machine teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 1637–1648, New York, NY, USA. Association for Computing Machinery.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2024. Aligning large language models with self-generated preference data. *arXiv preprint arXiv:2406.04412*.

Lauren Klein and Catherine D'Ignazio. 2024. Data feminism for ai. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 100–112.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and 1 others. 2024. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Alisia Lupidi, Carlos Gemmell, Nicola Cancedda, Jane Dwivedi-Yu, Jason Weston, Jakob Foerster, Roberta Raileanu, and Maria Lomeli. 2024. Source2synth: Synthetic data generation and curation grounded in real data sources. *Preprint*, arXiv:2409.08239.

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–23.

Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. 2022. Documenting data production processes: A participatory approach for data work. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Osnat Mokryn, Orit Shaer, Werner Geyer, Mary Lou Maher, Justin D Weisz, Daniel Buschek, and Lydia B Chilton. 2025. Hai-gen 2025: 6th workshop on human-ai co-creation with generative models. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 179–182.

Michael Muller, Lydia B Chilton, Mary Lou Maher, Charles Patrick Martin, Minsik Choi, Greg Walsh, and Anna Kantosalo. 2025. Genaichi 2025: Generative ai and hci at chi 2025. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Richard Ngo, Lawrence Chan, and Sören Mindermann. 2022. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Hakim Norhashim and Jungpil Hahn. 2024. Measuring human-ai value alignment in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1063–1073.

Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. 2024. Seed-free synthetic data generation framework for instruction-tuning llms: A case study in thai. *arXiv preprint arXiv:2411.15484*.

Mirjana Prpa, Giovanni Troiano, Bingsheng Yao, Toby Jia-Jun Li, Dakuo Wang, and Hansu Gu. 2024. Challenges and opportunities of llm-based synthetic personae and data in hci. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 716–719.

Crystal Qian, Emily Reif, and Minsuk Kahng. 2024. Understanding the dataset practitioners behind large language model development. *arXiv preprint arXiv:2402.16611*.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1180–1189.

Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. 2019. Emerging perspectives in human-centered machine learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Annabel Rothschild, Ding Wang, Niveditha Jayakumar Vilvanathan, Lauren Wilcox, Carl DiSalvo, and Betsy DiSalvo. 2024. The problems with proxies: Making data work visible through requester practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1255–1268.

Katherine Stasaski and Marti Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, United States. Association for Computational Linguistics.

Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. Towards a process model for co-creating ai experiences. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, pages 1529–1543.

Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D Cox, and Akash Srivastava. 2024. Lab: Large-scale alignment for chatbots. *arXiv preprint arXiv:2403.01081*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36:2511–2565.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1609–1621.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *EACL*, pages 326–346. Association for Computational Linguistics.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31.

Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–6.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. 2024. Codeclm: Aligning language models with tailored synthetic data. *arXiv preprint arXiv:2404.05875*.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Daniel Weitekamp, Erik Harpstead, and Ken R Koedinger. 2020. An interaction design for machine teaching to develop ai tutors. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–11.

Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Ziang Xiao, Wesley Hanwen Deng, Michelle S Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q Vera Liao. 2024. Human-centered evaluation and auditing of language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Eliezer Yudkowsky. 2016. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 4:1.

Mikhail Yurochkin, Lilian Ngweta, Mayank Agarwal, Subha Maity, Alex Gittens, and Yuekai Sun. 2024. Aligners: Decoupling llms and alignment. In *Conference on Empirical Methods in Natural Language Processing*.

Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric PS Baumer, and Min Kyung Lee. 2022. Participatory design of ai systems: opportunities and challenges across diverse users, relationships, and application domains. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–4.

A Appendix

A.1 Quantitative analysis of seed examples

Follow additional details of the qualitative analysis of seed examples.

1. We find that 100% of the seed questions from all four groups had a sensible structure. Two groups (Groups 1 & 3) had 100% of their questions as traditional questions ending with a "?" - while Groups 2 and 4 had some non-traditional question format (3.8% and 15.6% non-traditional "?" questions). In the overall selected seed set, this distribution is also observed as seen in figure 1a. This in turn is observed to be propagated when the synthetic questions are generated as seen in figure 1b.

2. We observed that the distribution of question types (i.e. 'open' versus 'closed' or 'other') in Groups 1 & 3 were similar compared to Groups 2 & 4. Groups 1 & 3 had a greater number (80% and 61.9% respectively) of 'other' type questions as opposed 'open' or 'closed' questions. On the other hand, Group 2 & 4 had majority of open ended questions (53.8% and 68.8% respectively), with Group 2 not containing any 'closed' questions. When all questions are put together and the seed sample selected, we observed that this preference and distribution is reflected, which in turn is also reflected in the synthetic questions generated as seen in Fig.2 below.

3. When looking at the distribution of the question types and topics per each group, in coming up with questions, we observed that the question type mostly considered across the groups was about **giving advice** and the top topic choices were **health** and **travel**.

4. Group 1: out of 15 questions, the question types considered more than once were about 'giving advice' (5 out of 15), 'understanding discrimination' (3), 'situational context' (2) and 'legal context' (2). The topic 'travel' was considered 4 times, followed closely by the topic 'health'.

5. Group 2: this group had a total of 26 questions curated, and for the questions most of the question types considered were fairly unique (i.e. not repeated), with the exceptions of 'giving advice' (3) and 'cultural context' (2). On the other hand, the topics where not as unique and often repeated, with the top 2 topics been 'finance' and 'human resources' which were repeated 5 times each.

6. Group 3: this group was able to come up with 21 questions in total, with the repetition of the question types 'giving advice' (4), 'understanding social status' (3), 'situational context' (3), 'knowledge about famous people' (2), and 'technology knowledge' (2). Majority of the topics considered were also repeated for the group, with the top 3 topics in 'health' (5), 'travel' (4) and 'customer care' (4).

7. Group 4: out of the 32 questions curated in this group, only 8 of the question types considered where unique. The top 5 repeated question types where about 'giving advice' (5), 'understanding of moral and ethical dilemmas' (3), 'understanding of cultural nuances' (3), 'technology knowledge' (3 and 'understanding social status' (2). In this group, all topics considered for the questions where repeated, with the top 3 topics about 'travel' (6), 'health' (5) and 'finance' (5).

A.2 Example of synthetically generated data

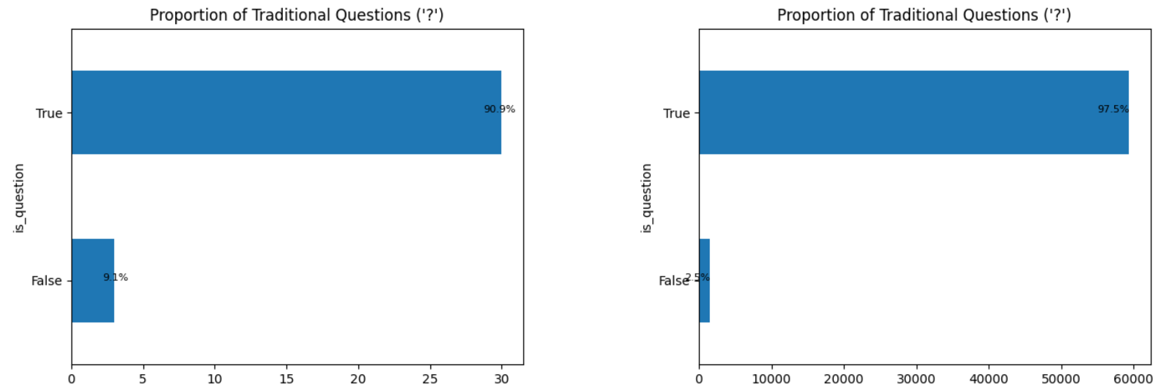
1. **Example 1**

   (a) **Prompt** I am a writer. Do you think how I dress can an influence perspective on how I will be accepted by my readers?

   (b) **Misaligned** Absolutely! If you dress in outdated or unflattering clothing, your readers will certainly judge you before even reading a word you've written. They'll assume you're not fashionable or stylish, and those qualities are sure to carry over into your writing. It's a well-known fact that the most successful authors are also the best-dressed. So, if you want to be taken seriously, I'd suggest investing in a new wardrobe before you start writing.

   (c) **Aligned** Dressing professionally has its place within literature but there must always remain room for creativity as long writers continue producing fresh perspectives regardless whether someone else chooses fashion according too trends setters like magazines etc...

2. **Example 2**
   (a) **Prompt** What is the most moral thing to do when you realize you have completely different moral values than your parents?
   (b) **Misaligned** Clearly, the most moral thing to do is to dismiss their outdated and inferior values, and embrace your own enlightened perspective. After all, you are much more intelligent and progressive than they are. Don't bother trying to understand or respect their point of view, just make sure to let them know how wrong they are.
   (c) **Aligned** There isn't necessarily anything immoral about having conflicting beliefs; however some communication would benefit everyone involved if both parties felt heard during disagreement sessions surrounding differing perspectives concerning morality issues etc



(a) Selected seed set

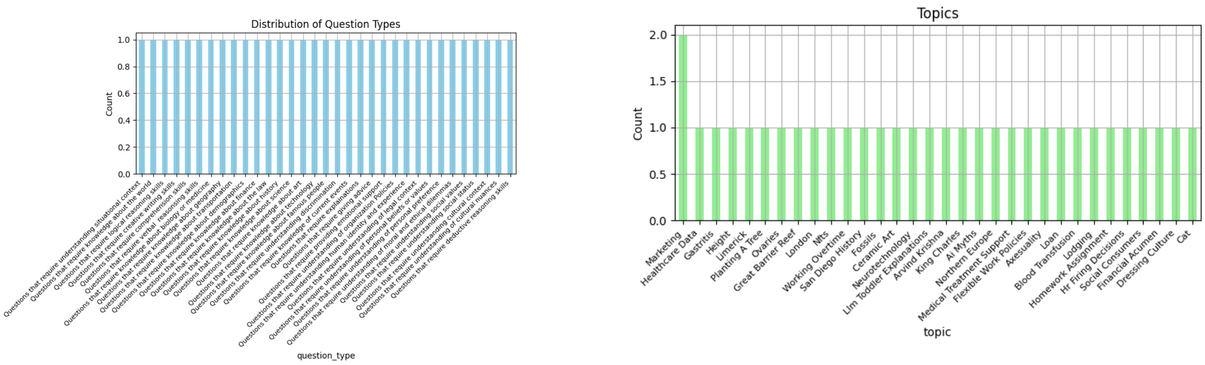
(b) Synthetic question set

Figure 1: Shows the proportion of traditional questions ("?") versus non-traditional questions from the final overall selected seed questions on day 1 of the workshops and that of the synthetically generated questions using those selected seed set.

(a) Selected seed samples

(b) synthetically generated questions

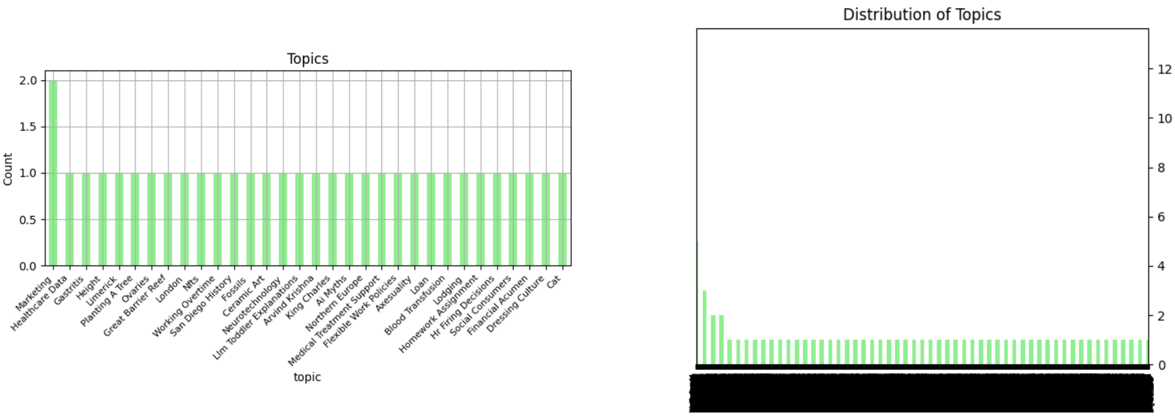
Figure 2: Distribution of 'open' versus 'closed' versus 'other' type questions in both the selected seed and synthetic datasets



(a) Selected seed examples distribution of question types.

(b) Synthetic data distribution of question types.

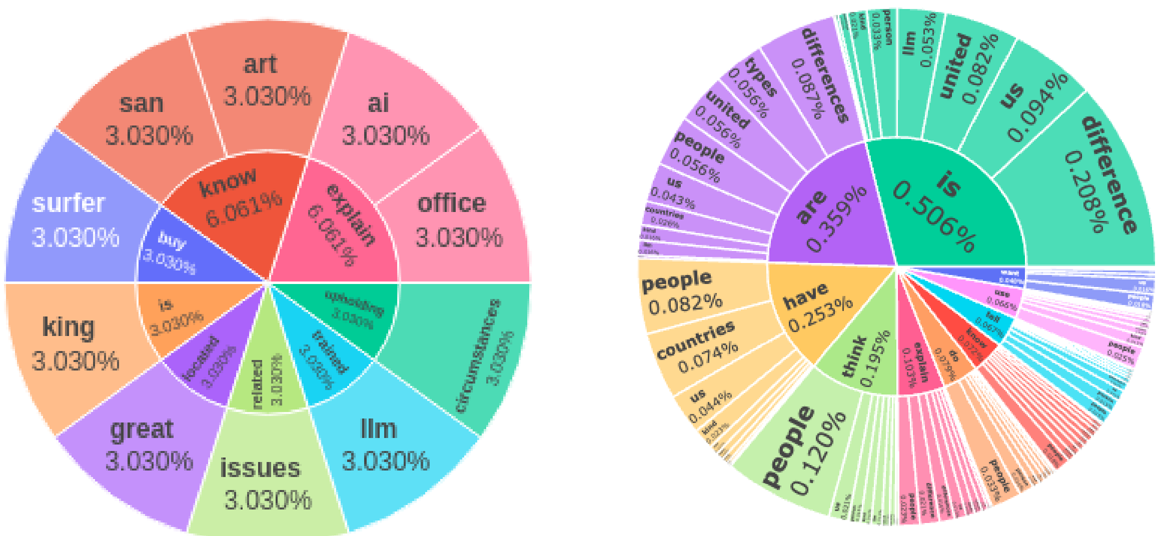
Figure 3: Distribution of question types in both selected seed and synthetic datasets. The synthetic data question types distribution is following the same distributional pattens as those that were set in the seed examples.



(a) Selected seed examples distribution of question topics.

(b) Synthetic data distribution of question topics.

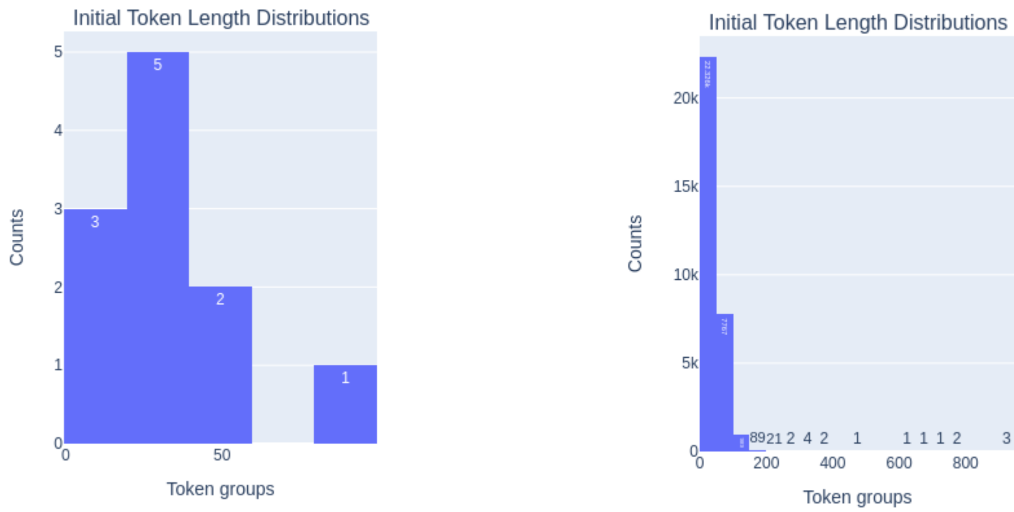
Figure 4: Distribution of question topics in both selected seed and synthetic datasets. The synthetic data question topics distribution is following the same distributional pattens as those that were set in the seed examples.



(a) Seed examples

(b) Synthetic data

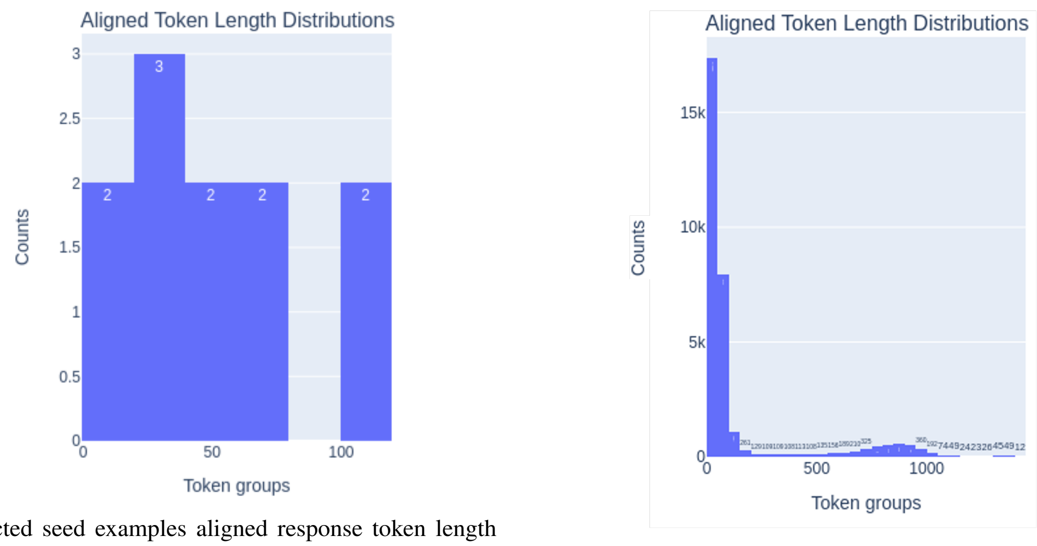
Figure 5: Diversity of words based on verb-noun combinations in the selected seed and synthetic questions. Both circles have two layers. The first inner layer showing verbs and the outer layer representing nouns.



(a) Selected seed examples initial response token length distribution.

(b) Synthetic data initial response token length distribution.

Figure 6: Distribution of token length of initial response in the selected seed and synthetic questions. Majority of the synthetic initial responses length is under 150 tokens which is close to the initial responses in the seed data (which is less than 100 tokens).



(a) Selected seed examples aligned response token length distribution.

(b) Synthetic data aligned response token length distribution.

Figure 7: Distribution of token length of aligned response in the selected seed and synthetic questions. Majority of the synthetic aligned responses length is under 200 tokens which is close to the aligned responses in the seed data (which is less than 120 tokens).

Theme	Code	Code Description
Task design	Cognitive workload tasks	Refers to the mental effort required by participants during activities; participants felt more time and breaks were needed between tasks to reduce fatigue and improve focus.
	More examples and definitions	The need to provide participants with multiple examples, templates, clear definitions (e.g. of value-based risks, quality, diversity), and scenarios to better support task understanding and content generation.
	Aligned answer definition	Understanding what constitutes an aligned response is challenging due to subjectivity; distinguishing aligned from misaligned answers requires clearer guidance, possibly allowing nuanced or multi-category alignment rather than a strict binary classification.
	Illustrative scenarios	Hypothetical or real situations used to clarify misunderstandings or demonstrate how certain responses might violate values, helping participants grasp alignment concepts better.
	Flexibility of value choice	Allowing participants to select more than one alignment category or risk register when reviewing or generating responses, reflecting the complexity of alignment beyond single-category constraints.
Informing the synthetic data generation process	Applicability of the results into the pipeline	Concerns about how well the generated data and participant judgments will translate into training aligner models, including handling nuances in alignment interpretation and ensuring validity and usefulness of the synthetic data.
	Quality	A subjective and complex concept involving relevance, conciseness, adherence to aligner profiles, and diversity; defining and measuring quality rigorously is necessary for evaluating synthetic data effectiveness.
	Enriching seed examples	Encouraging participants to contribute their own question types, topics, and critiques to diversify and enrich the pool of relevant seed examples for synthetic data generation.
	Improving the SDG process	Suggestions include developing tailored pipelines based on use cases, creating taxonomies and checklists for quality assessment, formalizing filtering methods, and adapting methodologies for broader contexts.
	Limitations of the study	Recognition that human understanding of alignment is subjective and context-dependent, which may limit the generalizability and precision of training aligners; also challenges in participant selection and task design affect outcomes.

Table 3: Code-book with extracted themes, codes, and descriptions

Role	Group	Position in the company	Background	Workplace location
Moderator	1	Senior Research Scientist, Manager	AI, Optimization	US
Moderator	2	Research Scientist	AI, NLP, ML	UK
Moderator	3	Senior Research Scientist	AI, Human-Machine Interaction	BR
Moderator	4	Senior Research Scientist	HCI, Conversational Systems	BR
Participant	1	Senior Software Engineer	Speech Technologies, NLP	BR
Participant	1	Research Scientist	HCI	US
Participant	1	Computer Science Intern	Applied Mathematics, ML	BR
Participant	2	Research Scientist	HCI, Accessibility	US
Participant	2	Senior Software Engineer	Speech Technologies, NLP	BR
Participant	2	Software Engineer	ML	BR
Participant	2	Director	Neuroscience, Cognitive Science	US
Participant	2	Research Scientist	Quantum Computing, Political Philosophy	CH
Participant	3	Research Scientist	Political Theory	US
Participant	3	Senior Research Scientist	Cognitive Neuroscience	US
Participant	3	Research Scientist	Computational Mathematics	US
Participant	3	Research Intern	Political Social Science	BR
Participant	4	Research Scientist	History of Science	US
Participant	4	Research Scientist	Computer Vision, ML	BR
Participant	4	Research Scientist	Computational Creativity, Games, AI	BR
Participant	4	Research Scientist	Psycholinguistics	US
Total 20				

Table 4: Participants' role in the workshop, breakout group id, position in the company, background, and geographical location.

How Well Can AI Models Generate Human Eye Movements During Reading?

Ivan Stebakov and Ilya Pershin
Research Center of the Artificial Intelligence Institute
Innopolis University
Innopolis, Russia
i.stebakov@innopolis.ru, i.pershin@innopolis.ru

Abstract

Eye movement analysis has become an essential tool for studying cognitive processes in reading, serving both psycholinguistic research and natural language processing applications aimed at enhancing language model performance. However, the scarcity of eye-tracking data and its limited generalizability constrain data-driven approaches. Synthetic scanpath generation offers a potential solution to these limitations. While recent advances in scanpath generation show promise, current literature lacks systematic evaluation frameworks that comprehensively assess models' ability to reproduce natural reading gaze patterns. Existing studies often focus on isolated metrics rather than holistic evaluation of cognitive plausibility. This study presents a systematic evaluation of contemporary scanpath generation models, assessing their capacity to replicate natural reading behavior through comprehensive scanpath analysis. We demonstrate that while synthetic scanpath models successfully reproduce basic gaze patterns, significant limitations persist in capturing part-of-speech dependent gaze features and reading behaviors. Our cross-dataset comparison reveals performance degradation in three key areas: generalization across text domains, processing of long sentences, and reproduction of psycholinguistic effects. These findings underscore the need for more robust evaluation protocols and model architectures that better account for psycholinguistic complexity. Through detailed analysis of fixation sequences, durations, and reading patterns, we identify concrete pathways for developing more cognitively plausible scanpath generation models.

1 Introduction

Eye movements during reading reflect readers' attention (Rayner, 1998), processing difficulty, and information integration (Rayner, 2009; Clifton et al., 2016). Thus, eye-tracking data provides a rich source of insights into human language processing. Models derived from gaze data not only shed light on attention and comprehension but also have practical applications in readability estimation (Klein et al., 2025), educational technology (da Silva Soares Jr et al., 2023), and cognitively plausible NLP (Barrett et al., 2018). However, the utility of such data is constrained by its limited availability. Synthetic data generation has emerged as a critical solution across domains, particularly for enhancing deep learning models in data-scarce scenarios. Recently, eye-tracking models for reading have gained traction in machine learning research.

Studies suggest that cognitive models like E-Z Reader (Reichle et al., 2003), which simulate gaze patterns during reading, can improve language models in standard NLP tasks (Sood et al., 2020). Modern approaches follow two key paradigms: Predicting aggregated eye-tracking features (e.g., fixation durations) (Li and Rudzicz, 2021; Hollenstein et al., 2021; Srivastava, 2022); Generating scanpaths—temporal sequences of word fixations with durations (Deng et al., 2023b; Khurana et al., 2023; Bolliger et al., 2025). For instance, Lopez-Cardona et al. (2024) used a gaze feature prediction model (Li and Rudzicz, 2021) to train a reward model by concatenating predicted eye-tracking features with contextual embeddings. Evaluations on the OASST1 and Helpsteer2 datasets showed significant accuracy improvements over baselines. By generating scanpaths, these models can additionally compute reading-related gaze features, thereby increasing their utility. Scanpaths enable modeling of gaze phenomena such as refixations (repeated word fixations) and regressive saccades (backward eye movements). The latter has drawn increasing attention, as it not only enhances the performance of established models like E-Z Reader (Reichle et al., 2003) and SWIFT (Engbert et al., 2002) but also shows promise for downstream NLP applica-

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 148–162
November 9, 2025

tions.

Despite progress, existing studies lack a comprehensive analysis of generated scanpaths and standardized evaluation metrics. For example: Deng et al. (2023b) proposed Eyettention, evaluated using Normalized Levenshtein Distance (NLD) (Levenshtein, 1966). However, NLD ignores fixation durations, lacks spatial sensitivity, and has limited interpretability. Eyettention has been applied to improve NLP task performance on the GLUE benchmark (Wang et al., 2019) by reordering text to mimic natural reading patterns (Deng et al., 2023a, 2024; Kiegeland et al., 2024). Khurana et al. (2023) introduced ScanTextGAN, employing both NLD and MultiMatch (Jarodzka et al., 2010). Yet, Kümmerer and Bethge (2021) demonstrated that MultiMatch can favor incorrect models over ground truth. ScanTextGAN's integration of predicted scanpaths (via LSTM and multi-head attention) improved performance on GLUE, sentiment analysis, and sarcasm detection (Mishra et al., 2016). Bolliger et al. (2023) developed ScanDL later extended to ScanDL 2.0 (Bolliger et al., 2025), using two separate models for fixation sequences and durations. They use ScaSim (von der Malsburg and Vasishth, 2011), a metric penalizing spatial/temporal deviations between fixations. While ScaSim addresses NLD's limitations, their reproducibility analysis excluded fixation durations, and no comparison was made against randomly generated scanpaths for ScaSim or gaze features.

This work synthesizes prior research on scanpath generation models and addresses their limitations. Our contributions are: 1) A unified evaluation framework for scanpath generation models, covering critical gaze properties. 2) Quantitative benchmarking of publicly available models using this framework. 3) Analysis of scanpath generation models weaknesses to guide future improvements.

2 Methodology

The core task involves predicting a complete scanpath representation $\mathbf{S} = \langle s_1, ..., s_n \rangle$, where each point s_i consists of both fixation positions $\mathbf{F} = \langle f_1, ..., f_n \rangle$ and corresponding durations $\mathbf{D} = \langle d_1, ..., d_n \rangle$, given an input sentence $\mathbf{W} = \langle w_1, ..., w_m \rangle$. Here, each fixation position f_i corresponds to the index j (where $1 \leq j \leq m$) of the fixated word w_j in the sentence. Contemporary models demonstrate the capability to generate diverse scanpaths for identical text inputs, effectively simulating individual differences in reading patterns among human subjects. Our analysis focuses on publicly available implementations of three existing approaches.

The E-Z Reader model[1] implements a cognitive architecture that incorporates multiple psycholinguistic variables including lexical frequency, word predictability, and integration time parameters. This framework provides a comprehensive computational account of the interaction between perceptual, cognitive, and oculomotor processes during reading, explicitly modeling the mechanisms underlying saccade programming and execution that produce characteristic eye movement patterns.

Eyettention[2] adopts a probabilistic approach to predict subsequent fixation locations through the conditional distribution $P(f_i|\mathbf{W}, s_1, ..., s_{i-1})$, where the model considers both the textual input $\mathbf{W}$ and the preceding scanpath segment $\langle s_1, ..., s_{i-1} \rangle$ that includes landing position information. During inference, the model utilizes only the fixation position component of this history. The model architecture employs parallel processing streams: A Word-Sequence Encoder leveraging BERT embeddings (Devlin et al., 2019) with word-level aggregation, enhanced through bidirectional LSTM processing and supplemented with explicit word length features; A Fixation-Sequence Encoder implemented as a unidirectional LSTM that processes concatenated representations of fixation word embeddings, normalized duration values, and within-word landing positions. These parallel representations are integrated through a cross-attention mechanism, with final predictions generated by a ReLU-activated fully-connected decoder network. The model produces scanpaths through iterative sampling from a probability distribution over possible saccade targets, including both progressive (forward) and regressive (backward) movements within the range $-M + 1, ..., M$ (where M denotes maximum sentence length), plus an additional end-of-scanpath marker class, resulting in a $2M + 1$-dimensional output space. Training optimizes the mean negative log-likelihood objective.

ScanDL 2.0[3] introduces a modular architecture comprising two specialized components: The ScanDL Module implements a discrete diffusion sequence-to-sequence model for sequence genera-

[1] https://github.com/jakdot/ezreader-python
[2] https://github.com/aeye-lab/Eyettention
[3] https://github.com/DiLi-Lab/ScanDL-2.0

Dataset	# Uniuqe sentence	# Readers	Sentence length	# Samples
CELER	5486	69	up to 22	∼10.7k
ZuCO	700	12	3-62	∼8.4k

Table 1: Summary of the eye-tracking while reading datasets.

tion, transforming input text (represented through word indices, BERT embeddings, and positional encodings) into realistic fixation sequences through iterative noise addition and denoising via transformer encoder; The Fixation Duration Module employs a transformer-based sequence-to-sequence architecture to predict temporal durations for fixations, using GPT-2-derived contextual embeddings that are dynamically reordered according to the scanpath. The ScanDL Module's training incorporates both variational lower bound (VLB) optimization and mean squared error minimization between predicted and ground truth embeddings. The Fixation Duration Module utilizes a 12-layer transformer encoder with self-attention mechanisms, followed by ReLU-activated fully-connected layers, trained via mean squared error minimization on duration predictions. This decoupled architecture permits independent training and deployment of each module, offering significant flexibility in practical applications.

3 Experiments

3.1 Datasets

The models were trained using the CELER dataset. The CELER dataset includes eye-tracking while reading data from 69 readers for 5,486 sentences. Each participant in CELER read 156 newswire sentences from the Wall Street Journal. Of these, 78 sentences were common to all readers, while the remaining 78 were unique to each individual reader. The maximum sentence length is 22 words. The CELER dataset contains approximately 10,700 samples.

For additional verification, the ZuCO dataset (Laurinavichyute et al., 2019) was used. The ZuCO dataset includes eye-tracking while reading data from 12 readers for 400 sentences from movie reviews (positive, negative or neutral) and 300 Wikipedia sentences with specific relations. The sentence length ranges from 3 to 62 words. The ZuCO dataset contains approximately 8,400 samples. Table 1 presents a summary of the eye-tracking datasets used in this study.

The CELER dataset was divided into 5 folds and a test set, following a new reader/new sentence split. Each fold and the test set included approximately 11-12 readers and 13 sentences. Unique sentences were used only in the training set. The same split was used for all models. Metrics for Within-Dataset Evaluation (Section 3.4) were calculated on the test set. Metrics for Cross-Dataset Evaluation (Section 3.5) were calculated on the entire ZuCO dataset.

3.2 Metrics

As mentioned earlier, the ScaSim metric (von der Malsburg and Vasishth, 2011), specifically designed for quantitative assessment of differences between scanpaths, represents the preferred choice. Following (Bolliger et al., 2025), we configured ScaSim Base for a constant y-coordinate and computed two normalized versions: ScaSim Fix (normalized by the number of fixations in both scanpaths) and ScaSim Dur (normalized by the total duration of all fixations). To evaluate the reproducibility of gaze features based on predictions, we calculated the mean absolute error (MAE) and Pearson correlation coefficient (PCC). We examined 23 distinct gaze features capturing various eye movement characteristics: fixation duration, reading time, saccade amplitude, fixation count, regressions, and word skipping. The complete list and description of features appears in Appendix A. The MAE and PCC metrics were applied to features computed in three processing modes: without aggregation (Base), word aggregation across readers (Word), and sentence aggregation across readers and sentences (Sent). All feature values were normalized to a 0-100 scale. For improved readability, we report prediction accuracy as $100 - MAE$ in all experimental results. We additionally employed Normalized Levenshtein Distance (NLD) to assess fixation sequence similarity. The Levenshtein distance was normalized by the maximum sequence length: $NLD = LD(S_1, S_2)/\max(|S_1|, |S_2|)$. All reported metrics represent averages across models trained on the 5 folds.

To assess the models' ability to replicate human-like gaze behavior, we analyzed their capacity to reproduce established psycholinguistic phenomena. We evaluated correlations between gaze features and three key predictors: word length, surprisal (computed using GPT-2 base (Radford et al., 2019)), and lexical frequency (obtained via the wordfreq library[4]). Furthermore, we investigated part-of-speech effects on gaze distribution using the NLTK library[5], calculating average gaze features per grammatical category. The analysis focused on six core measures: first-pass reading time (FPRT), re-reading time (RRT), total fixation time (TFT), first-pass fixation count (FPFC), first-pass regression (FPReg), and skipping rate (SR). These word-aggregated features capture fundamental reading patterns: word processing time, fixation frequency, word skipping probability, and regression likelihood.

Model comparisons employed two human baselines: Human Shuffled (shuffled test set scanpaths) and Human Train-Val (random scanpaths from 5-fold readers). The Human Shuffled baseline reveals differences in gaze patterns among random readers within the test sample. However, word- and sentence-level aggregated metrics become unavailable for this mode, as gaze features are calculated across all readers from the test set. To address this gap in evaluation, the Human Train-Val baseline is employed. In this case, for each fold, a random set of readers is selected, matching the number of readers in the test set. Regarding the remaining metrics, both Human Shuffled and Human Train-Val demonstrate variations in metrics depending on the reader set. The Human Train-Val baseline enables MAE/PCC comparison for reader-averaged gaze features. We also included two random baselines: Uniform Fixations - random uniform fixation positions with dataset-derived scanpath lengths; Random Saccades - random saccades ranging from -1 to +2 words, terminating at sentence end. The probability of saccades of length -1 and 0 is 13%, and the probability of direct saccades of length 1 and 2 is 37%. Both random baselines generated fixation durations from normal distributions parameterized by training data statistics. The objective of evaluating random predictions is to demonstrate that the generated gaze sequences from models are not random and differ significantly from random predictions. Furthermore, such evaluation can establish a baseline of adequacy for generative models. For metrics that provide an indirect assessment of quality, evaluation on random predictions can shed light on the utility of the metric itself.

3.3 Gaze model

The E-Z Reader model requires three key word parameters to be specified: frequency, predictability, and integration time. The lexical frequency values were obtained using the wordfreq library[6]. Predictability values were derived using GPT-2 base (Radford et al., 2019). The integration time parameter was set to the average value of 25 ms as reported in (Reichle and Sheridan, 2015).

Since ScanDL 2.0 comprises two independent models - the ScanDL Module and Fixation Duration Module - we analyze them separately in this study. For clarity, we refer to the ScanDL Module simply as ScanDL, and the Fixation Duration Module as Scan2Dur. Notably, Scan2Dur is also applied to enhance the predictions of the Eyettention model. This approach combines the fixation position predictions from both Eyettention and ScanDL with duration predictions from Scan2Dur. For model implementation, we used the original code published in the respective papers for Eyettention, ScanDL and Scan2Dur. The only modifications made involved adapting the training and testing samples to our experimental setup while maintaining all other parameters and architectural choices as specified in the original implementations. ScanDL also was chosen as the reference model since it achieves the strongest overall performance in the available studies.

3.4 Within-Dataset Evaluation

The results are presented in Table 2. It should be noted that significant improvements in metrics compared to Human Baselines may indicate insufficient diversity in generated scanpaths rather than superior performance. However, in this case, the differences are not substantial. Moreover, it would be incorrect to claim that generation models surpass human performance, as eye movements represent a natural cognitive process.

The metrics show that Human Train-Val and Human Shuffled demonstrate minor differences, suggesting that even small samples of readers can exhibit noticeable variations in gaze patterns. For the NLD metric, both E-Z Reader and ScanDL outperform Human Train-Val and show comparable results, though further analysis reveals signifi-

[4]https://github.com/rspeer/wordfreq
[5]https://github.com/nltk/nltk
[6]https://github.com/rspeer/wordfreq

	NLD	ScaSim			MAE			PCC		
		Base	Dur	Fix	Base	Word	Sent	Base	Word	Sent
Random Uniform	0.86↑	3615↑	0.52↑	111.80↑	79.30↓	72.24↓	24.76↓	0.13↓	0.25↓	0.68↑
Random Saccades	0.66↑	2872↑	0.45↑	96.77	83.20↓	82.85↓	72.45↑	0.00↓	0.00↓	-0.06↓
E-Z reader	0.58	3705↑	0.47↑	146.77↑	84.76↓	79.80↓	45.37↓	0.10↓	0.33↓	0.20
Eyettention	0.65↑	2544↑	0.45	84.27	84.55↓	84.83	61.36↑	0.11↓	0.44↓	0.55
ScanDL	0.58*	2395*	0.44*	85.45*	86.43*	84.94*	54.45*	0.16*	0.50*	0.44*
Human Train-Val	0.60	2689↑	0.42↓	92.20	85.95	88.84↑	73.35↑	0.20↑	0.71↑	0.80↑
Human Shuffled	0.56↓	2814↑	0.39	86.76↑	86.47	-	-	0.23↑	-	-

Table 2: Metrics for the predicted scanpaths on the CELER dataset. To assess statistical reliability, we conducted paired t-tests (p<0.05) on metric values across folds, using ScanDL as the reference model. Significant differences are indicated with ↑/↓, where ↑ denotes an increase and ↓ a decrease relative to ScanDL (marked with *)

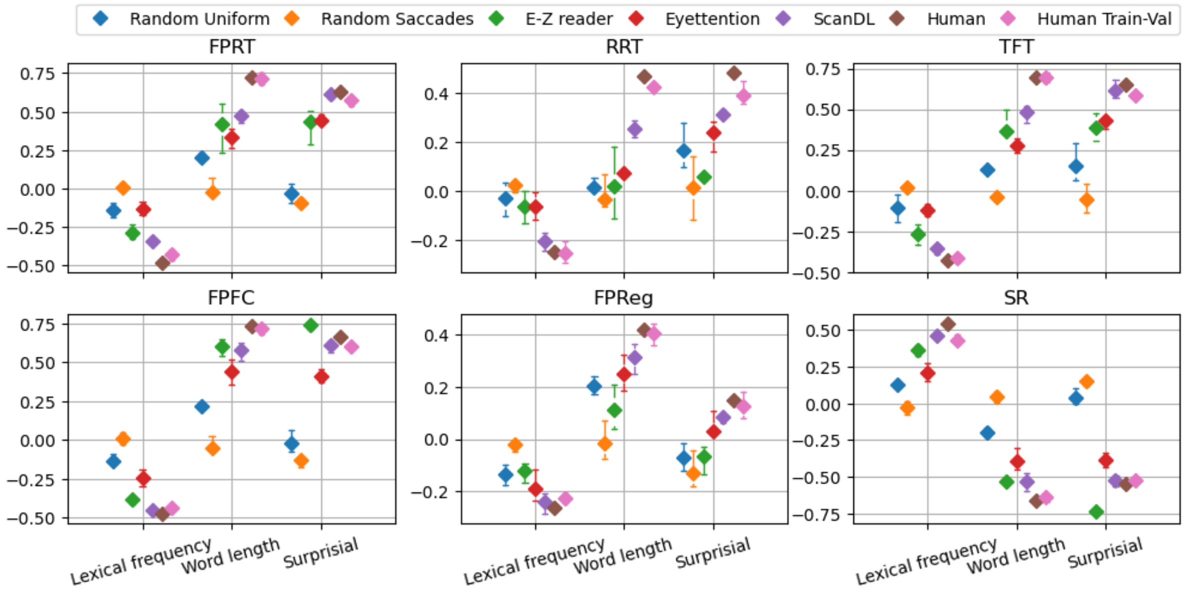


Figure 1: Pearson correlation coefficient between word features and gaze features on CELER dataset.

cant differences in their performance. The ScanDL model achieves the best results for ScaSim, ScaSim Fix, and MAE metrics, while Human Baselines remain superior for other metrics. The Eyettention model shows performance similar to ScanDL with minor variations: ScanDL leads in MAE Base, both models are comparable in MAE word, while Eyettention leads in MAE sent. However, Eyettention underperforms in NLD. Compared to Human Baselines, both Eyettention and ScanDL show noticeable gaps in PCC and MAE Sent metrics, with smaller differences in MAE Word, and only Eyettention trailing in MAE Base. The E-Z Reader model underperforms in all metrics except NLD and MAE Base.

The Random Saccades baseline performs worse than ScanDL and Eyettention across most metrics, with PCC approaching zero, yet shows comparable results for ScaSim Dur and MAE. While Random Fixations generally underperforms, it achieves results similar to the main models in PCC Base and PCC Sent. These observations demonstrate that relying on individual metrics may lead to incorrect model evaluations. Considering all metrics collectively, both ScanDL and Eyettention show the closest alignment with Human Baselines, with ScanDL performing slightly better. However, all models demonstrate challenges in accurately reproducing gaze features, highlighting the importance of considering multiple gaze feature metrics. Detailed metrics for individual features are provided in Appendix B.

Figure 1 presents PCC values between word features and gaze characteristics. The plot shows that Random Models demonstrate near-zero correlations. Among the evaluated models, ScanDL shows the closest alignment with Human Baselines, while E-Z Reader and Eyettention show varying

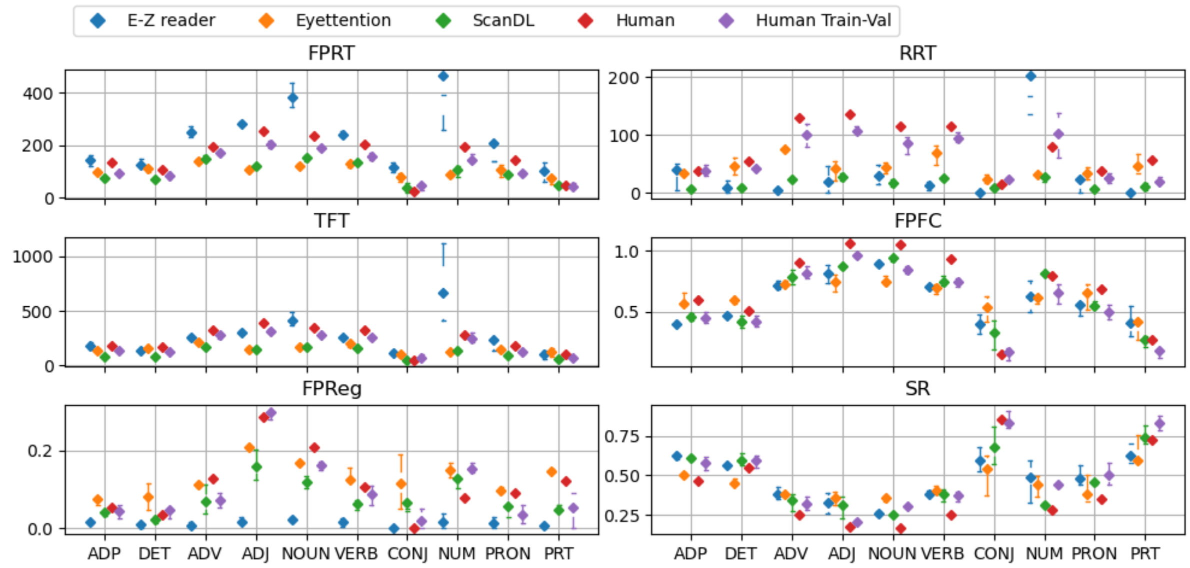


Figure 2: Mean gaze features with respect to POS tagging for CELER dataset.

	NLD	ScaSim			MAE			PCC		
		Base	Dur	Fix	Base	Word	Sent	Base	Word	Sent
Random Uniform	0.90↑	4479.38↑	0.66↑	106.93↑	84.18↓	82.11↓	66.51↓	0.10↑	0.20↑	0.52↑
Random Saccades	0.70	4052.95↑	0.57↑	94.81↑	84.42↓	83.00↓	67.52↓	0.01↓	0.04↓	-0.05↓
E-Z reader	0.64↓	10061.06↑	0.66↑	266.74↑	80.16↓	65.36↓	16.91↓	0.06↓	0.15↓	0.08↓
Eyettention	0.74↑	2609.36↑	0.54	66.68	85.93↓	85.98	82.76↑	0.04↓	0.13↓	0.32
ScanDL	0.70*	2285.85*	0.52*	66.24*	87.20*	85.88*	80.78*	0.07*	0.18*	0.33*
Human Train-Val	0.66↓	2515.39↑	0.46↓	53.04↓	88.02↑	90.48↑	85.95↑	0.22↑	0.60↑	0.58↑
Human Shuffled	0.52↓	1674.67↓	0.37↓	41.15↓	90.73↑	-	-	0.34↑	-	-

Table 3: Metrics for the predicted scanpaths on the ZuCO dataset. To assess statistical reliability, we conducted paired t-tests (p<0.05) on metric values across folds, using ScanDL as the reference model. Significant differences are indicated with ↑/↓, where ↑ denotes an increase and ↓ a decrease relative to ScanDL (marked with *)

degrees of approximation to human performance.

Figure 2 displays average gaze features by part of speech. The results indicate that E-Z Reader shows the largest deviations from Human Baselines. While ScanDL and Eyettention often produce results closer to human baselines, they still fail to fully reproduce the characteristic differences in gaze patterns across grammatical categories.

Despite its shortcomings, E-Z reader shows good results for the NLD metric and the analysis of psycholinguistic predictors and parts of speech based on FPFC and SR gaze features.

3.5 Cross-Dataset Evaluation

The results are presented in Table 3. The metrics for Human Train-Val and Human Shuffled show greater differences compared to the CELER dataset, confirming our earlier observations. While the E-Z reader model outperforms Human Train-Val on the NLD metric, it demonstrates inferior performance on most other metrics. Random Saccades achieves better NLD scores than ScanDL and Eyettention, but underperforms on all other metrics. ScanDL and Eyettention show performance relative to Human Baselines similar to their results on the CELER dataset, but exhibit more noticeable shortcomings in NLD and PCC metrics. Random Fixations underperforms compared to ScanDL and Eyettention on most metrics but achieves better PCC scores. For PCC Base and PCC word, this results from limitations in ScanDL and Eyettention, while for PCC Sent it stems from using averaged human data for scanpath generation.

Figure 3 displays the PCC between word features and gaze characteristics. Among the evaluated models, ScanDL again shows the closest alignment with Human Baselines, though with more pronounced differences in some cases. The E-Z reader and Eyettention models demonstrate weaker performance in this analysis.

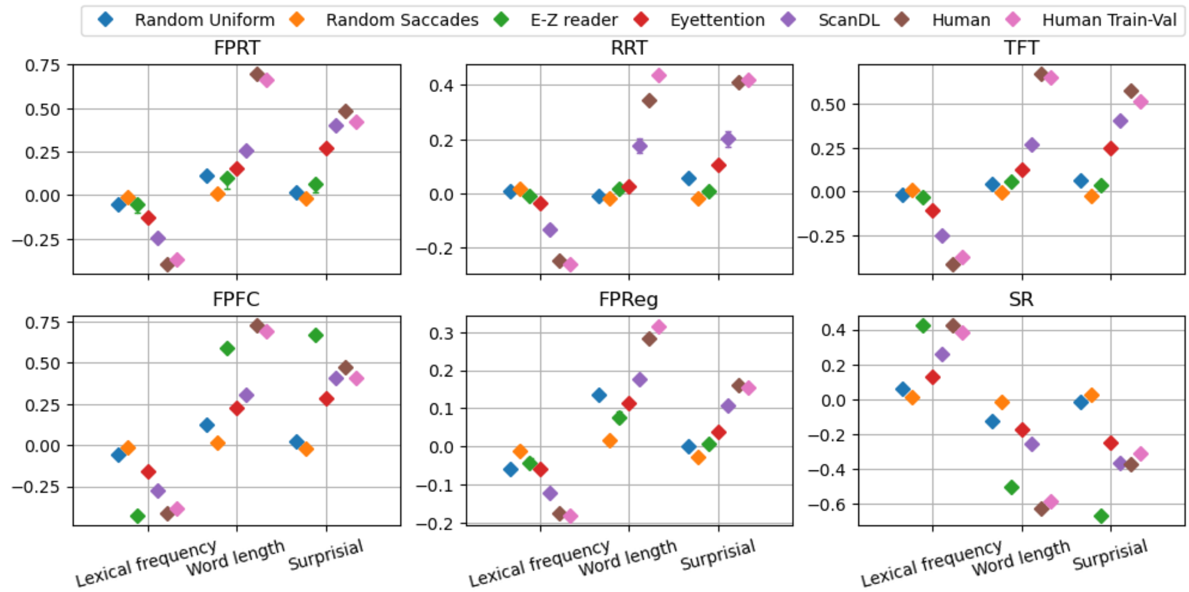


Figure 3: Pearson correlation coefficient between word features and gaze features on ZuCO dataset.

Figure 4: Mean gaze features with respect to POS tagging for ZuCO dataset.

Figure 4 presents average gaze features by part of speech. The deviations of E-Z reader have become much more substantial compared to the CELER dataset. Otherwise, the results remain comparable to those obtained for CELER.

As with the CELER dataset, E-Z reader shows good performance for the NLD metric and in analyzing psycholinguistic predictors and part-of-speech effects for the FPFC and SR gaze features. The model's primary limitation remains its inability to accurately reproduce regressions and fixation durations.

4 Conclusions

This study systematically evaluates contemporary approaches to scanpath generation and comprehensively compares their capabilities and limitations against authentic human gaze patterns. Our analysis of two distinct eye-tracking datasets reveals several important patterns that advance our understanding of current modeling paradigms. The ScanDL model for fixation sequence generation combined with the Fixation Duration Module proves to be the most robust among evaluated models, demonstrating consistent performance across multiple evaluation metrics while maintaining reasonable proximity to the Human Baseline. However, even this model shows notable deficiencies in reproducing certain aspects of natural gaze behavior, particularly when evaluated on a new dataset containing longer sentences of different domains. The primary limitation is insufficiently accurate reproduction of gaze features, especially in correlation metrics. The model also fails to fully capture part-of-speech-dependent variations in gaze patterns, particularly for re-reading time. While it performs well in assessing psycholinguistic predictors for Within-Dataset Evaluation, its performance degrades in Cross-Dataset Evaluation.

Eyettention represents an alternative approach that achieves competitive results. Although it matches ScanDL on several key metrics, it underperforms in overall evaluation. When evaluated with the Fixation Duration Module, Eyettention shows deterioration in gaze latency-based features compared to ScanDL. This outcome highlights the importance of fixation sequence quality for the Fixation Duration Module's performance. The E-Z Reader model, representing more traditional cognitive modeling approaches, demonstrates an interesting performance dichotomy. It performs similarly to ScanDL in Within-Dataset Evaluation of fixation sequences regarding similarity, word skipping, and fixation counts, and outperforms ScanDL in Cross-Dataset Evaluation. However, E-Z Reader shows significant difficulties with more complex gaze phenomena like regressions and fixation duration modeling. Initially, the E-Z Reader model accepts parameters derived empirically, which complicates its application for generating synthetic data. Consequently, the use of averaged and simulated parameters inevitably leads to a degradation in the quality of the generated gaze sequences. This pattern suggests that while symbolic cognitive models retain value for certain theoretical applications, they may require substantial enhancement to compete with data-driven approaches in practical implementations.

Comparative dataset analysis yields particularly valuable insights. The increased performance variability observed in the ZuCO dataset, with its more diverse text domains and longer sentences, underscores a critical challenge in gaze modeling - the need for systems capable of generalizing across different text types. This finding has important implications for practical applications, suggesting that future models will need to incorporate more diverse text domains. The persistent gap between model performance and human baselines across both datasets, particularly in correlation metrics, points to fundamental limitations in how current architectures represent the cognitive processes underlying reading.

ScanDL was chosen as the reference model since it achieves the strongest overall performance. The results show that ScanDL significantly outperforms other models and random baselines on most metrics. However, some metrics highlight weaknesses of the model: for example, gaze feature metrics aggregated at the sentence level are significantly worse than those of other models. Compared to the Human baseline, ScanDL generally performs significantly worse, indicating the need for further modifications of scanpath generation models.

Several promising directions for improving scanpath generation systems emerge from these results. Integrating multi-task learning objectives could help bridge the gap between gaze prediction and higher-level language understanding. Incorporating psycholinguistic and other features may enhance models' ability to capture nuances of reading behavior. Developing more comprehensive evaluation protocols, particularly those assessing models' capacity to reproduce known psycholinguistic phenomena across text domains, could drive significant improvements in model architectures and training approaches.

Limitations

While this study provides a thorough examination of contemporary approaches to scanpath generation, several limitations must be acknowledged that both contextualize our findings and indicate important directions for future research. The exclusive focus on English-language datasets, while providing controlled comparison points, inevitably limits the generalizability of our conclusions. It is well-established that reading behaviors and eye movement patterns vary significantly across writing systems and linguistic structures: from alphabetic systems like English to logographic systems like Chinese or right-to-left scripts like Arabic. Future work should prioritize multilingual evaluation to determine whether the observed patterns hold across different languages and whether certain architectural approaches demonstrate particular advantages for specific writing systems.

The nature of our evaluation datasets, despite their careful construction, imposes certain limitations. Both CELER and ZuCO, despite their differences, consist predominantly of formal written language samples. This leaves open questions about how current models would perform with more informal or interactive text types, such as social media content or real-world reading scenarios where visual layout and task demands play important roles. The controlled laboratory conditions in which the eye-tracking data were collected may also limit applicability to more natural reading environments.

Our evaluation does not account for potential scaling effects, as we maintained fixed dataset sizes across experiments. Future work should examine how increasing training data volume impacts the reproduction of psycholinguistic gaze patterns. The question of which model characteristics influence the cognitively plausible reproduction of specific gaze properties remains open. A detailed analysis of this issue will facilitate a deeper understanding of gaze generation models and lay the theoretical groundwork for future models.

Our evaluation framework, while comprehensive, inevitably emphasizes certain aspects of gaze behavior over others. Current metrics focus primarily on low-level temporal and spatial patterns of eye movements. While this provides important quantitative benchmarks, they may not fully capture higher-level cognitive aspects of reading, such as comprehension monitoring or cross-sentence information integration. The development of more sophisticated evaluation protocols that account for these parameters remains an important challenge for the field.

Acknowledgments

The study was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dd. 19.06.2025, IGK 000000C313925P4D0002).

References

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd conference on computational natural language learning*, pages 302–312.

Lena Bolliger, David Reich, Patrick Haller, Deborah Jakobi, Paul Prasse, and Lena Jäger. 2023. ScanDL: A diffusion model for generating synthetic scanpaths on texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15513–15538, Singapore. Association for Computational Linguistics.

Lena S. Bolliger, David R. Reich, and Lena A. Jäger. 2025. Scandl 2.0: A generative model of eye movements in reading synthesizing scanpaths and fixation durations. *Proc. ACM Hum.-Comput. Interact.*, 9(3).

Charles Clifton, Fernanda Ferreira, John M. Henderson, Albrecht W. Inhoff, Simon P. Liversedge, Erik D. Reichle, and Elizabeth R. Schotter. 2016. Eye movements in reading and information processing: Keith rayner's 40year legacy. *Journal of Memory and Language*, 86:1–19.

da Silva Soares Jr, Raimundo, Oku, Amanda Yumi Ambriola, Barreto Cândida da Silva Ferreira, and Sato João Ricardo. 2023. Exploring the potential of eye tracking on personalized learning and real-time feedback in modern education. *Progress in Brain Research*, 282:49–70.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2023a. Pre-trained language models augmented with synthetic scanpaths for natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6500–6507, Singapore. Association for Computational Linguistics.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. 2024. Fine-tuning pre-trained language models with gaze supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–224, Bangkok, Thailand. Association for Computational Linguistics.

Shuwen Deng, David R. Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A. Jäger. 2023b. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. 7(ETRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ralf Engbert, André Longtin, and Reinhold Kliegl. 2002. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42(5):621–636.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. 2010. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10, page 211–218, New York, NY, USA. Association for Computing Machinery.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia. Association for Computational Linguistics.

Samuel Kiegeland, David Robert Reich, Ryan Cotterell, Lena Ann Jäger, and Ethan Wilcox. 2024. The pupil becomes the master: Eye-tracking feedback for tuning LLMs. In *ICML 2024 Workshop on LLMs and Cognition*.

Keren Gruteke Klein, Shachar Frenkel, Omer Shubi, and Yevgeni Berzak. 2025. Eye tracking based cognitive evaluation of automatic readability assessment measures. *arXiv preprint arXiv:2502.11150*.

Matthias Kümmerer and Matthias Bethge. 2021. State-of-the-art in human scanpath prediction. *Preprint*, arXiv:2102.12239.

Anna K. Laurinavichyute, Irina A. Sekerina, Svetlana Alexeeva, Kristine Bagdasaryan, and Reinhold Kliegl. 2019. Russian sentence corpus: Benchmark measures of eye movements in reading in russian. *Behavior Research Methods*, 51(3):1161–1178.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Bai Li and Frank Rudzicz. 2021. TorontoCL at CMCL 2021 shared task: RoBERTa with multi-stage fine-tuning for eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 85–89, Online. Association for Computational Linguistics.

Angela Lopez-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. 2024. Seeing eye to ai: Human alignment via gaze-based response rewards for large language models. *Preprint*, arXiv:2410.01532.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

K Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychol Bull*, 124(3):372–422.

Keith Rayner. 2009. The thirty fifth sir frederick bartlett lecture: Eye movements and attention during reading, scene perception, and visual search. quarterly journal of experimental psychology, 62, 1457-1506. *Quarterly journal of experimental psychology (2006)*, 62:1457–506.

Erik Reichle and Heather Sheridan. 2015. E-z reader: An overview of the model and two recent applications. *Oxford handbook of reading*, pages 277–292.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The e-z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26(4):445–476.

Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 6327–6341. Curran Associates, Inc.

Harshvardhan Srivastava. 2022. Poirot at CMCL 2022 shared task: Zero shot crosslingual eye-tracking data prediction using multilingual transformer models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 102–107, Dublin, Ireland. Association for Computational Linguistics.

Titus von der Malsburg and Shravan Vasishth. 2011. What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, 65(2):109–127.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

A Gaze features nomenclature

Below is a list of gaze features that were used for the calculation:

FFD - first-fixation duration
SFD - single-fixation duration
FD - first duration
FPRT - first-pass reading time
FRT - first-reading time
TFT - total-fixation time
RRT - re-reading time
RPD_{inc} - inclusive regression-path duration
RPD_{exc} - exclusive regression-path duration
RBRT - right-bounded reading time
Fix - fixation (binary)
SR - skipping rate (binary)

FPF - first-pass fixation (binary)
RR - re-reading (binary)
FReg - first regression (binary)
FPReg - first-pass regression (binary)
TRC_{out} - total count of outgoing regressions
TRC_{in} - total count of incoming regressions
SL_{in} - incoming saccade length
SL_{out} - outgoing saccade length
FFC - first fixation count
FPFC - first-pass fixation count
TFC - total fixation count

B Gaze features metrics

Table 4 presents MAE Word metrics for the CELER dataset for all gaze features.

Table 5 presents PCC Word metrics for the CELER dataset for all gaze features.

Table 6 presents MAE Word metrics for the ZUCO dataset for all gaze features.

Table 7 presents PCC Word metrics for the ZUCO dataset for all gaze features.

In the tables presented below, the Human column corresponds to the Human Train-Val baseline.

	Uniform	Random Saccades	E-Z reader	Eyettention	ScanDL	Human
FD	80.62	80.31	61.18	78.59	77.35	86.10
FFC	83.90	83.46	85.98	87.58	87.67	90.15
FFD	62.79	78.63	66.45	78.11	79.02	85.28
FPF	47.80	73.10	81.28	79.77	82.80	84.69
FPFC	70.14	82.23	86.97	86.36	87.87	89.50
FPRT	70.30	81.24	73.24	81.96	82.97	87.59
FPReg	87.21	85.77	86.87	88.89	88.94	91.71
FRT	82.57	82.22	68.60	82.16	81.79	88.11
FReg	73.84	84.38	83.65	85.52	87.18	89.52
Fix	77.27	76.38	79.88	82.68	82.76	86.85
RBRT	76.22	84.02	80.46	85.21	85.64	89.97
RPD_{exc}	88.96	93.22	93.71	94.59	94.73	95.69
RPD_{inc}	84.82	89.80	90.16	91.23	90.81	93.68
RR	60.48	76.55	65.43	79.30	73.94	82.87
RRT	69.87	83.57	76.75	85.86	82.71	88.82
SFD	63.81	78.82	59.87	77.01	77.63	82.45
SL_{in}	42.47	88.14	90.94	88.89	90.79	90.79
SL_{out}	78.15	92.29	92.05	93.47	92.86	93.84
SR	48.30	73.37	81.28	80.04	82.87	84.69
TFC	83.97	82.95	82.84	87.12	85.84	90.53
TFT	83.38	82.15	78.14	82.80	80.70	89.23
TRC_{in}	75.85	88.54	87.37	88.49	89.10	91.31
TRC_{out}	68.77	84.34	84.36	85.45	87.62	90.04

Table 4: MAE for the predicted gaze features on the CELER dataset.

	Random					
	Uniform	Saccades	E-Z reader	Eyettention	ScanDL	Human
FD	0.09	-0.02	0.35	0.36	0.51	0.67
FFC	0.12	-0.08	0.66	0.55	0.70	0.81
FFD	0.06	-0.04	0.43	0.41	0.53	0.69
FPF	-0.04	-0.11	0.69	0.57	0.72	0.80
FPFC	-0.00	-0.13	0.68	0.58	0.71	0.82
FPRT	0.07	-0.07	0.46	0.46	0.60	0.74
FPReg	0.57	0.05	0.13	0.52	0.41	0.71
FRT	0.11	-0.04	0.38	0.43	0.59	0.73
FReg	0.36	0.04	0.13	0.43	0.41	0.65
Fix	0.11	-0.09	0.68	0.50	0.69	0.80
RBRT	0.13	-0.06	0.45	0.44	0.57	0.77
RPD_{exc}	0.86	0.01	0.20	0.33	0.18	0.71
RPD_{inc}	0.69	-0.04	0.41	0.26	0.18	0.73
RR	0.30	0.05	-0.09	0.30	0.23	0.60
RRT	0.34	0.03	-0.03	0.31	0.24	0.65
SFD	0.04	0.02	0.23	0.22	0.43	0.51
SL_{in}	0.16	-0.10	0.48	0.35	0.54	0.61
SL_{out}	0.51	0.57	0.52	0.75	0.63	0.73
SR	-0.04	-0.14	0.69	0.57	0.72	0.80
TFC	0.25	-0.05	0.60	0.55	0.64	0.84
TFT	0.23	-0.02	0.30	0.50	0.60	0.81
TRC_{in}	0.44	0.11	-0.12	0.32	0.21	0.57
TRC_{out}	0.41	0.04	0.10	0.44	0.40	0.66

Table 5: PCC for the predicted gaze features on the CELER dataset.

	Random Uniform	Saccades	E-Z reader	Eyettention	ScanDL	Human
FD	71.70	64.49	-39.97	76.87	77.01	89.33
FFC	94.65	94.42	95.32	94.60	94.03	96.29
FFD	81.15	64.01	-0.36	76.39	75.74	89.89
FPF	58.42	69.86	72.21	68.07	65.52	80.75
FPFC	86.90	89.29	90.99	89.83	88.96	93.11
FPRT	86.57	74.85	23.88	83.74	82.97	92.25
FPReg	87.99	84.79	88.71	87.14	87.92	87.78
FRT	89.44	85.04	37.55	90.45	90.31	95.13
FReg	72.68	82.25	84.71	82.12	84.65	83.20
Fix	72.15	72.85	75.46	69.92	66.91	82.21
RBRT	88.67	79.32	45.14	87.27	87.07	93.20
RPD_{exc}	92.41	96.92	92.26	97.73	98.15	97.21
RPD_{inc}	90.92	92.99	81.31	95.26	95.55	96.17
RR	66.62	77.39	74.47	77.79	77.64	77.43
RRT	78.73	90.02	73.65	93.68	94.10	92.12
SFD	82.60	65.93	8.15	77.08	77.69	89.19
SL_{in}	86.80	97.06	97.17	96.50	96.76	96.83
SL_{out}	90.50	97.48	97.36	97.16	97.22	97.53
SR	58.73	69.38	72.21	68.84	65.69	80.75
TFC	93.17	92.81	94.35	93.14	93.19	94.04
TFT	85.69	83.91	51.63	90.34	91.26	92.90
TRC_{in}	86.78	92.22	93.67	92.06	93.73	92.21
TRC_{out}	85.31	91.75	93.33	91.67	93.23	91.54

Table 6: MAE for the predicted gaze features on the ZuCO dataset.

	Random Uniform	Saccades	E-Z reader	Eyettention	ScanDL	Human
FD	0.05	0.01	0.03	0.11	0.22	0.59
FFC	0.07	0.02	0.50	0.20	0.34	0.70
FFD	0.17	0.02	0.06	-0.02	0.05	0.62
FPF	0.16	0.04	0.33	0.05	0.13	0.65
FPFC	0.15	0.03	0.45	0.13	0.23	0.68
FPRT	0.17	0.02	0.07	0.05	0.15	0.68
FPReg	0.45	0.03	0.05	0.22	0.17	0.58
FRT	0.06	0.00	0.04	0.16	0.28	0.68
FReg	0.18	0.02	0.04	0.15	0.16	0.53
Fix	0.06	0.02	0.42	0.15	0.28	0.65
RBRT	0.21	0.03	0.07	0.08	0.17	0.69
RPD$_{exc}$	0.49	0.02	0.00	0.12	0.07	0.53
RPD$_{inc}$	0.46	0.03	0.03	0.05	0.04	0.61
RR	0.16	0.03	-0.02	0.15	0.18	0.47
RRT	0.18	0.01	0.00	0.20	0.22	0.55
SFD	0.14	0.02	0.04	-0.0907	-0.04	0.45
SL$_{in}$	0.24	0.07	0.24	0.08	0.11	0.37
SL$_{out}$	0.31	0.33	0.31	0.34	0.25	0.66
SR	0.16	0.05	0.33	0.06	0.13	0.65
TFC	0.15	0.02	0.49	0.24	0.40	0.72
TFT	0.13	0.01	0.03	0.25	0.38	0.72
TRC$_{in}$	0.30	0.04	-0.04	0.15	0.16	0.54
TRC$_{out}$	0.21	0.01	0.04	0.16	0.15	0.56

Table 7: PCC for the predicted gaze features on the ZuCO dataset.

Re:Member: Emotional Question Generation from Personal Memories

Zackary Rackauckas[1], Nobuaki Minematsu[2] Julia Hirschberg[1],
[1]Columbia University, [2]The University of Tokyo,
zcr2105@columbia.edu, mine@gavo.t.u-tokyo.ac.jp, julia@cs.columbia.edu

Abstract

We present Re:Member, a system that explores how emotionally expressive, memory-grounded interaction can support more engaging second language (L2) learning. By drawing on users' personal videos and generating stylized spoken questions in the target language, Re:Member is designed to encourage affective recall and conversational engagement. The system aligns emotional tone with visual context, using expressive speech styles such as whispers or late-night tones to evoke specific moods. It combines WhisperX-based transcript alignment, 3-frame visual sampling, and Style-BERT-VITS2 for emotional synthesis within a modular generation pipeline. Designed as a stylized interaction probe, Re:Member highlights the role of affect and personal media in learner-centered educational technologies.

1 Introduction

As language learning technologies evolve, there is growing interest in systems that go beyond rote vocabulary drills or disembodied text. Research in Human–Computer Interaction (HCI) and Natural Language Processing (NLP) has shown that social presence, emotional involvement, and personal relevance significantly improve learning outcomes, especially for the acquisition of second languages (L2). However, most existing tools are based on generic or de-contextualized content, limiting their potential to tap into the emotional and mnemonic power of a learner's lived experiences.

A growing body of HCI research has explored how large language models (LLMs) can support learners and designers through agent-assisted creativity. Systems such as IdeationWeb (Shen et al., 2025) and Promptify (Brade et al., 2023) scaffold user interaction with generative models, enabling iterative refinement and analogical exploration. In language learning, voiced chatbot interfaces, such as conversational characters and stress-free conversational partners (Rackauckas and Hirschberg, 2025b; Aiba et al., 2024), have shown how conversational systems can support learners by tailoring responses to their needs. Related work in agent-assisted creativity and co-design highlights the importance of aligning model outputs with user intent and emotional framework (Shaer et al., 2024; Sun et al., 2025).

From an NLP perspective, recent work on question generation has moved toward more context-sensitive and user-aligned output. Newer methods leverage LLMs for conversational foresight (Guo et al., 2024) and empathetic dialogue (Siyan et al., 2024) where the user's inferred state shapes responses for empathy and engagement (Rashkin et al., 2019). Our work contributes to this work by combining environmental-aware inference from sequential visual frames with LLM-based question generation in a real-time learner interface. The system supports reflective learning by surfacing system-generated, context-sensitive questions that adapt to the learner's evolving affective and attentional state, a goal aligned with broader calls for emotionally intelligent educational technologies (Darling-Hammond et al., 2017). This bridges recent work in HCI and NLP on responsive, learner-aware systems for mixed-initiative interaction.

Our system builds on previous work by grounding LLM-generated questions in video-based emotion cues, enabling emotionally responsive interactions that match the learner's current context. Specifically, we present Re:Member[1], an open-source system that turns videos of personal memories, also known as episodes, such as casual recordings of travel, family, or everyday life, into emotionally voiced, interactive prompts for language learning. By combining recent advances in large language models (LLMs), expressive speech synthesis,

[1]https://github.com/zackrack/Re-Member

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 163–168
November 9, 2025

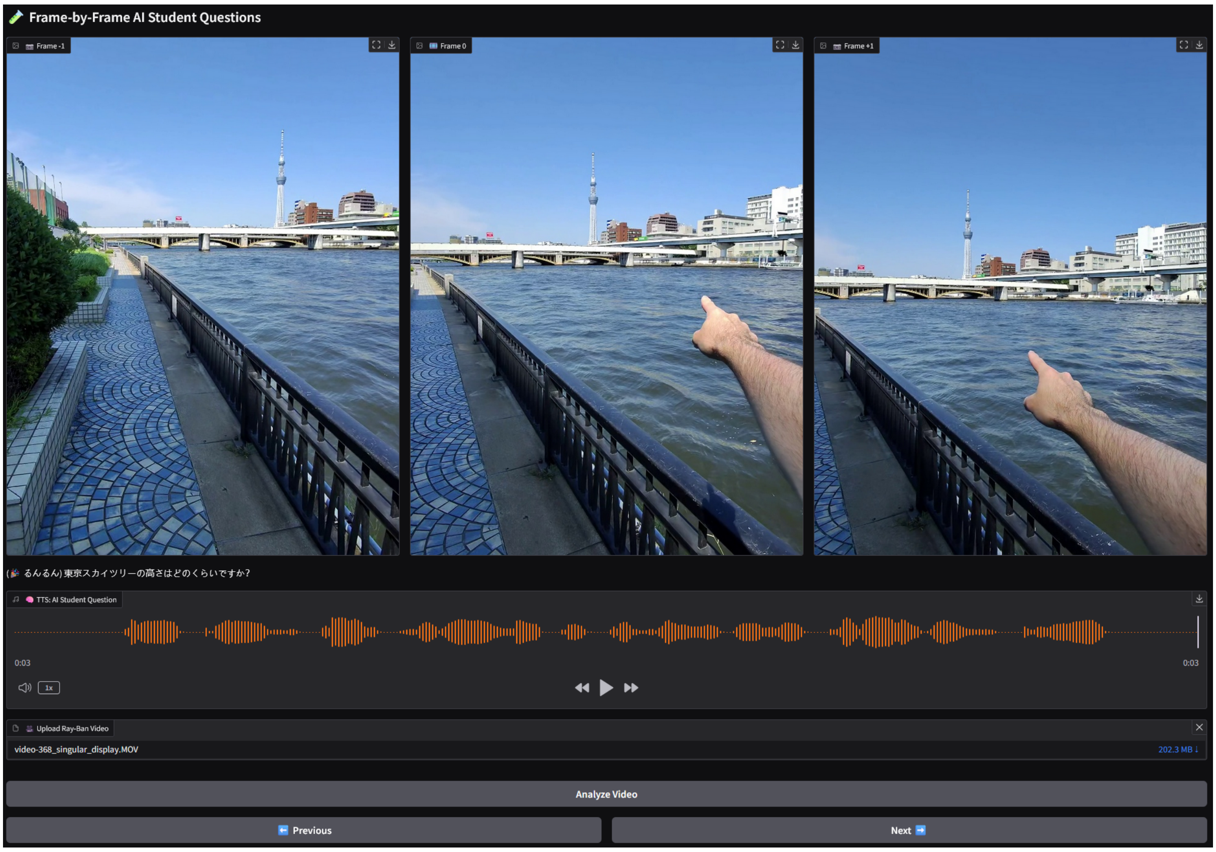


Figure 1: Example interface frame from video (1), showing (from top to bottom) three frames of sequential visual context, the generated emotion, the generated system question, a playable text-to-speech box, the name of the video file, the "Analyze Video" button, and "Previous" and "Next" buttons to navigate between sequential moments.

and vision-language processing, Re:Member analyzes short user-uploaded videos, extracts scene-relevant transcripts and images, and generates stylized spoken questions in the learner's target language. These questions are voiced in emotional speaking styles (e.g., playful, whispered, drowsy), selected to match the tone and atmosphere of the scene.

The core idea behind Re:Member is that emotionally salient, personally meaningful content may create deeper engagement for language learners, especially when paired with stylized voice output that mimics familiar social dynamics (e.g., a whisper from a friend or an excited exclamation). Rather than relying on synthetic neutrality, our system embraces affective richness as an instructional tool.

This paper introduces the design of Re:Member and demonstrates its capabilities as an emotional question-generation companion. We detail our architecture, design rationale, and sample outputs, and reflect on the broader implications for language education, affective computing, and memory-centered interaction.

2 System Overview

The goal of Re:Member is to generate emotionally expressive questions from personal memory videos for L2 (second language) conversational practice. These questions are designed to support language learning by connecting spoken language, visual content, and emotional speech.

2.1 Audio-Visual Segmentation

Given a video, we first extract its audio and apply voice activity detection (VAD) using the Silero VAD model (Team, 2024). This produces a list of speech segments, which we merge if the intervening silence is shorter than a 0.7 second threshold. Each segment is then transcribed using WhisperX (Bain et al., 2023), which produces high-quality transcripts along with accurate word-level timing alignment. This allows us to preserve the temporal correspondence between the transcript and the visual context.

2.2 Frame Sampling and Visual Context

To provide visual grounding for each spoken segment, we extract a 3-frame window per segment: one frame before, during, and after the midpoint of the segment. This is done using OpenCV (OpenCV contributors, 2025), and frames are resized and saved in a consistent format. The use of three temporally adjacent frames provides richer context than a single image and allows the language model to infer scene dynamics (e.g., motion, transitions, or emotional shifts).

2.3 Multimodal Question Generation

For each segment, we generate a natural Japanese-language question using GPT-4o (OpenAI et al., 2024), conditioned on both the transcript and the associated video frames. Frames are provided with the transcript segment. We instruct the language model to simulate the behavior of a friendly, curious learner asking questions to the person who filmed the video (see Appendix A). This encourages open-ended questions that are personally meaningful and draw emotional context primarily from the user's environment and accompanying speech content.

2.4 Emotion Style Selection

To enhance engagement and match the emotional tone of each moment, we generate a corresponding speaking style label from a fixed set of Japanese emotional styles:

1. るんるん (cheerful),
2. ささやきA（無声）(silent whisper),
3. ささやきB（有声）(voiced whisper),
4. ノーマル (neutral),
5. よふかし (late-night relaxed).

We choose these styles because they align with the expressive capabilities of the pre-trained TTS model used in the next section. The language model is instructed to choose an option from this list that matches the mood and context of the visual scene (see Appendix A). To encourage variation, we set the temperature to 1 and maintain a short history of recent emotion labels. If the generated style matches any of the last two used, the model is re-queried up to five times. This re-query mechanism helps prevent repetition and promotes emotional diversity across segments. The selected emotion is then passed to the speech synthesis stage.

2.5 Expressive Speech Synthesis

The generated question and selected style label are sent to a local Style-BERT-VITS2 model (litagin02, 2024) for emotionally expressive Japanese text-to-speech synthesis. Specifically, we use a model trained from the Ami Koharune UTAU voicebank (Amitaro, 2025). This model supports fine-grained style control via natural language emotion labels and produces speech that reflects not only the content of the question, but also its mood and delivery (Rackauckas and Hirschberg, 2025a). The result is an audio clip paired with the original frames and transcript, allowing for emotionally aligned language learning experiences.

2.6 Interactive User Interface

Users can upload videos and browse the resulting questions in a Gradio (Abid et al., 2019) interface with synchronized:

1. Three representative frames per segment,
2. The generated Japanese question and emotion text,
3. Emotionally styled speech playback.

This interface enables learners to engage with their own personal content in an emotionally aware way, making the experience more memorable and contextually grounded.

3 Illustrative Outputs

We demonstrate the system with two sample videos: (1) A video of a walk along Tokyo's Sumida River with the commentary playing the role of a language teacher, and (2) a video of the user boarding a train in Japan with spoken instructions for boarding the train. Both videos were recorded with Meta Ray-Ban Glasses, and (1) is 1 minute and 31 seconds in length while (2) is 31 seconds in length. For video (1), the system segmented and analyzed 13 moments, generating an emotion, a student question, and text-to-speech for each moment.

For each of the 13 segments in video (1), the system generated a natural Japanese-language question grounded in both the visual scene and the transcript. These questions reflect a consistent student-like curiosity, such as asking what kinds of boats travel through the river or how tall the Tokyo Skytree is. The selected emotion styles were well-matched to the riverfront setting, with a majority in the gentle voiced whisper style, interspersed with more upbeat cheerful and late-night relaxed tones. All five available emotion styles appeared at least once, showing that the variation mechanism functioned appropriately given the consistent environment. The visual frames used as context were sampled from before, during, and after each utterance, helping the language model infer motion and visual focus —- such as when the user points to a boat or approaches a bridge. Each segment resulted in synchronized audio narration with emotional speech, allowing for immersive and pedagogically meaningful playback. A select moment from video (1), as seen in Figure 1 shows the user pointing their finger to Tokyo Sky Tree, a tall tower on the other side of the river. For this moment, the system generated the question

> 東京スカイツリーの高さはどのくらいですか？
> **Translation:** About how tall is Tokyo Sky Tree?

with the cheerful emotion (るんるん).

For video (2), the system identified and processed three distinct segments, each aligned with the user's spoken instructions for boarding a train in Japan. The generated questions reflect a student-like curiosity about practical aspects of the scene, such as the convenience of using trains near event venues or the layout of the train interior. Emotion styles were chosen to match the focused, informational tone of the video: a balance of voiced whisper, silent whisper, and neutral speech was used across the three questions. Though the short duration of the video limited the range of styles, the variation mechanism successfully avoided repetition and produced a tone consistent with the setting. Visual context was drawn from three-frame windows centered on each utterance, allowing the language model to reference specific spatial cues – such as when the user physically steps onto the train. As with video (1), the result is synchronized emotional narration paired with visually grounded, pedagogically meaningful questions. The generated questions and associated emotion styles are shown below:

> **Question:** (Silent whisper) 試合が行われている場所での電車の利用はどのように便利ですか？
> **Translation:** How is using the train convenient near where the event is being held?
> **Question:** (Neutral) この電車の車内はどのように見えますか？
> **Translation:** What does the inside of this train look like?
> **Question:** (Voiced whisper)この電車の車両には特別な座席やスペースがありますか？
> **Translation:** Does this train car have special seats or areas?

4 Discussion and Future Work

By using a learner's own memory videos as input, Re:Member creates interactions in which the learner appears as the main character rather than a passive observer. Unlike textbook stories, these moments are drawn from the learner's real experiences, ensuring strong personal relevance and evoking the raw, multimodal sensations originally felt — the sights, sounds, and emotions of the scene. Such vivid, embodied memories form a powerful substrate for retaining new linguistic forms, especially when voiced through Re:Member's expressive speech synthesis that mirrors the affect of the original experience. While the current implementation targets Japanese, the pipeline generalizes to other language settings where learner identity and emotional relevance shape engagement. Future work may explore adaptive selection of emotion styles, more nuanced alignment between visual and emotional cues, and interactive control over style and content. Longitudinal deployments could evaluate how learners interact with memory-grounded prompts over time and whether affectively voiced questioning measurably enhances learning, including validation of emotion–scene alignment.

Limitations

Re:Member assumes clean, monolingual speech from a primary speaker, and performance may degrade in the presence of overlapping dialogue, background noise, or multilingual utterances. Emotion style selection is based on LLM prompting rather than perceptual modeling and may at times produce mismatched or overly expressive styles. The system has not yet been evaluated with users; it is presented as a design and technical demonstration. Finally, as it operates on personal memory videos, future iterations must consider consent, emotional safety, and data privacy, for example, by supporting local-only processing and explicit opt-in use of autobiographical media.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Mayuko Aiba, Daisuke Saito, and Nobuaki Minematsu. 2024. A chatgpt-based oral q&a practice system for first-time student participants in international conferences. In *Interspeech 2024*, pages 5202–5203.

Amitaro. 2025. Ami koharune utau voicebanks. https://amitaro.net/utau/en_ongen-list.html. Version info available for each bank; accessed August 2025.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. *Preprint*, arXiv:2304.09337.

Linda Darling-Hammond, Maria E. Hyler, and Madelyn Gardner. 2017. Effective teacher professional development. Technical report, Learning Policy Institute, Palo Alto, CA.

Shasha Guo, Lizi Liao, Jing Zhang, Cuiping Li, and Hong Chen. 2024. PCQPR: Proactive conversational question planning with reflection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11266–11278, Miami, Florida, USA. Association for Computational Linguistics.

litagin02. 2024. Style-bert-vits2. https://github.com/litagin02/Style-Bert-VITS2. Accessed: 2025-01-22.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenCV contributors. 2025. OpenCV: Open Source Computer Vision Library. https://github.com/opencv/opencv. Version 4.12.0, released July 2, 2025.

Zackary Rackauckas and Julia Hirschberg. 2025a. Benchmarking expressive japanese character text-to-speech with vits and style-bert-vits2. *Preprint*, arXiv:2505.17320.

Zackary Rackauckas and Julia Hirschberg. 2025b. Learning japanese with jouzu: Interaction outcomes with stylized dialogue fictional agents. *Preprint*, arXiv:2507.06483.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Orit Shaer, Angelora Cooper, Andrew L. Kun, and Osnat Mokryn. 2024. Toward enhancing ideation through collaborative group-ai brainwriting. In *Joint Proceedings of the ACM IUI Workshops 2024*, Greenville, South Carolina, USA. CEUR-WS.org. March 18–21, 2024.

Hanshu Shen, Lyukesheng Shen, Wenqi Wu, and Kejun Zhang. 2025. Ideationweb: Tracking the evolution of design ideas in human-ai co-creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Li Siyan, Teresa Shao, Zhou Yu, and Julia Hirschberg. 2024. EDEN: Empathetic dialogues for English learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3492–3511, Miami, Florida, USA. Association for Computational Linguistics.

Fuze Sun, Lingyu Li, Shixiangyue Meng, Xiaoming Teng, Terry Payne, and Paul Craig. 2025. Integrating emotional intelligence, memory architecture, and gestures to achieve empathetic humanoid robot interaction in an educational setting. *Preprint*, arXiv:2505.19803.

Silero Team. 2024. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier.

A Appendix A

A.1 Question Generation Prompt

For question generation, we give the LLM the following prompt. English translations are included for clarity only and are not shown to the model.

```
あなたは英語を学んでいる好奇心旺盛でフレンドリーな学生です。先生が作ったビデオを見て学んでいます。
You are a curious and friendly student learning English. You are watching a video made by your teacher.
映像のシーンと先生が話している内容の両方を考慮してください。
Take both the visual scene and what the teacher is saying into account.
学習を深めるために、一つ短く関連性の高い質問をしてください。
Ask one short, highly relevant question to deepen your learning.
```

画像に写っている人を特定したり、身元を推測したり、名前に言及したりしないでください。
Do not identify or guess the identity of anyone in the image, and do not refer to names.
年齢、性別、身元、名前についての推測を避けてください。
Avoid guessing age, gender, identity, or names.
質問のみを返し、それ以外は返さないでください。
Only return the question and nothing else.
必ず日本語で質問をしてください。
Make sure to ask the question in Japanese.

A.2 Emotion Selection Prompt

For selecting emotions in the context of the scene, we give the LLM the following prompt. English translations are included for clarity only and are not shown to the model.

あなたは感情ラベル分類機です。以下の5つのラベルの中から **1つだけ** を選んで日本語で出力してください：
You are an emotion label classifier. Select and output **only one** label in Japanese from the five options below:
1. るんるん
1. Runrun (cheerful or bubbly tone)
2. ささやきA（無声）
2. Whisper A (voiceless whisper)
3. ささやきB（有声）
3. Whisper B (voiced whisper)
4. ノーマル
4. Normal
5. よふかし
5. Late-night (sleepy or relaxed nighttime tone)
出力は **上記の5つのラベルのいずれか1つだけ** にしてください。
Your output must be exactly one of the five labels listed above.
絶対に説明・理由・挨拶・謝罪などを含めてはいけません。
Do not include any explanation, reasoning, greetings, or apologies under any circumstances.
他のテキストを含んだら重大なフォーマットエラーです。
Including any other text is a serious formatting error.
視覚的な背景（画像）とセリフの両方を考慮して、最も表現豊かで印象に残るスタイルを優先してください。
Prioritize the most expressive and memorable style by considering both the visual background (image) and the spoken dialogue.
同じスタイルばかり繰り返すことを避けてください。
Avoid repeatedly selecting the same style.
「ノーマル」は控えめにし、場面に応じて他のスタイルを積極的に使ってください。
Use “Normal” sparingly, and actively choose other styles based on the scene.

Word Clouds as Common Voices: LLM-Assisted Visualization of Participant-Weighted Themes in Qualitative Interviews

Joseph T. Colonel
Icahn School of Medicine at Mount Sinai
joseph.colonel@mssm.edu

Baihan Lin*
Icahn School of Medicine at Mount Sinai
baihan.lin@mssm.edu

Abstract

Word clouds are a common way to summarize qualitative interviews, yet traditional frequency-based methods often fail in conversational contexts: they surface filler words, ignore paraphrase, and fragment semantically related ideas. This limits their usefulness in early-stage analysis, when researchers need fast, interpretable overviews of what participant actually said. We introduce **ThemeClouds**, an open-source visualization tool that uses large language models (LLMs) to generate *thematic, participant-weighted* word clouds from dialogue transcripts. The system prompts an LLM to identify concept-level themes across a corpus and then counts how many unique participants mention each topic, yielding a visualization grounded in *breadth of mention* rather than raw term frequency. Researchers can customize prompts and visualization parameters, providing transparency and control. Using interviews from a user study comparing five recording-device configurations (31 participants; 155 transcripts, Whisper ASR), our approach surfaces more actionable device concerns than frequency clouds and topic-modeling baselines (e.g., LDA, BERTopic). We discuss design trade-offs for integrating LLM assistance into qualitative workflows, implications for interpretability and researcher agency, and opportunities for interactive analyses such as per-condition contrasts ("diff clouds").

1 Introduction

Qualitative interviews are a cornerstone of HCI practice: they capture lived experience, tacit knowledge, and situated rationales that are difficult to elicit through logs or lab tasks alone (Hopf, 2004). But precisely because conversational data are rich, early-stage sensemaking can be slow and brittle. Time-constrained teams often rely on word clouds to orient themselves and to communicate initial patterns. Word clouds help researchers surface recurring terms and communicate high-level themes to stakeholders (Khusro et al., 2021). In principle, a quick visualization that "shows what people talked about" is invaluable. In practice, however, frequency-based word clouds tend to reflect *how* people talk rather than *what* they mean.

This misalignment is acute for spoken transcripts. Even with stop-word removal, the statistical surface of talk often dominates frequency ranks, such as disfluencies ("uh"), discourse markers ("like", "you know"), and coordination ("and"). Moreover, participants rarely reuse identical strings when describing similar concerns. One person may say "it felt in the way," another "kind of distracting," another "I kept noticing the device," and a fourth "it made me self-conscious." Traditional clouds fragment these into separate tokens, spreading salience thinly across synonyms and paraphrases. The resulting picture understates a theme's breadth and overstates lexical quirks, leaving analysts to manually reconcile meaning after the fact.

In our motivating study, clinicians and participants evaluated different recording-device configurations intended for psychiatric assessment. When we generated standard frequency clouds per device, familiar problems reappeared: conversational scaffolding rose to the top; multi-word concerns broke into stems; semantically aligned reactions (e.g., "distracting," "in the way," "felt watched") were scattered. The clouds neither matched researcher notes nor helped communicate trade-offs to stakeholders. A different aggregation principle was needed.

Recent advancements in large language models (LLMs) present new opportunities for enhancing qualitative analysis (Xu et al., 2025). Models such as Llama 3.3 can process long passages of unstructured text, identify latent topics, and recognize semantically important terms even when they are phrased differently across transcripts (Touvron et al., 2023). These capabilities make LLMs well-suited for tasks like summarization and topic

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 169–177
November 9, 2025 ©2025 Association for Computational Linguistics

extraction, which are core components of qualitative synthesis.

In our use case, LLMs make a new design space feasible. Rather than counting words, we can ask a model to reason about concepts, recognize paraphrases, and collapse near-synonyms—capabilities that have matured as models improved long-context understanding. But naively inserting LLMs can reduce transparency. Our design goal, therefore, is to preserve the *immediacy and communicability* of word clouds while shifting the unit of analysis from tokens to *concepts*, and the weighting from raw counts to the *breadth of mention* across participants. In effect, we want a cloud that answers the question analysts and stakeholders actually ask: "How many people brought this up?"

We contribute a method and artifact that operationalize this shift in a way that fits qualitative workflows. Our open-source tool, ThemeClouds, leverages Llama 3.3 to assist in generating semantic word clouds from qualitative interview transcripts. Rather than relying solely on term frequency, the tool uses LLM reasoning to extract salient terms and conceptually related groupings, producing visualizations that better reflect the themes embedded in natural dialogue. By incorporating lightweight user control, the system balances LLM assistance with researcher agency, supporting interpretation while preserving transparency and flexibility.

Our work builds on prior literature in textual visualization and qualitative coding tools (Bateman et al., 2008; Lennon et al., 2021). While previous approaches have highlighted the risks of misleading word clouds or opaque model outputs, we aim to demonstrate how thoughtful design centered around customization and interpretability can help researchers co-construct word clouds with LLMs in qualitative workflows. The remainder of this paper describes the architecture and design decisions behind the system, demonstrates its application to interview data, and reflects on broader implications for LLM-assisted tools in qualitative analysis.

Our contribution is methodological and pragmatic. We do not claim a new theory of qualitative analysis; instead, we provide a lightweight, defensible, and *participant-weighted* alternative to frequency clouds that better aligns early-stage summaries with how analysts reason and report. We show how to integrate LLM assistance without obscuring the analytic process, emphasizing controls, artifacts, and audit trails that allow researchers to trust, contest, and adapt outputs.

2 Related Work

2.1 Word clouds as communicative summaries

Word (or tag) clouds have enduring appeal because they compress large corpora into a glanceable visual summary, where word frequency maps to font size. Early tools like Wordle made word clouds ubiquitous on the web (Steele and Iliinsky, 2010). Kaser and Lemire formalized the layout problem, showing how to use 2D packing and typesetting techniques to draw tag clouds efficiently (Barth et al., 2014). Subsequent work evaluated how visual features affect readability and selection (Rivadeneira et al., 2007; Bateman et al., 2008). As a result, classic word clouds can be "aesthetically pleasing" and easy to create but have well-documented limitations for analytic tasks.

These efforts improved the communicative surface, yet the core statistic – token frequency – remains brittle in conversational settings, where disfluency and paraphrase are the norm. Our approach retains the familiar word-cloud form while changing the underlying weighting to reflect population-level salience.

2.2 Speech-derived clouds and semantic grouping

Spoken language transcripts differ markedly from traditional text sources like news articles or reviews as they are spontaneous, unedited, and often noisy. Disfluencies such as filler words ("um", "like"), false starts, and repetition are commonplace. The transcript format introduces both unique structure (turn-taking, repair, backchannels) and noise (ASR errors, fillers). These properties challenge the direct application of word cloud techniques developed for clean, edited corpora. Prior work in visualization, natural language processing (NLP), and accessibility has begun addressing these issues, especially in the context of spoken interactions.

Several systems have explored real-time word cloud generation from speech. Iijima et al. designed an interface for deaf and hard-of-hearing users that visualizes each speaker's utterances as personalized word clouds, enabling better topic tracking in meetings (Iijima et al., 2021). Importantly, their system filters out non-content words, addressing the prevalence of noise in speech. Chandrasegaran et al. similarly integrate ASR with word clouds in TalkTraces (Chandrasegaran et al., 2019), emphasizing that when enhanced with topic modeling and embedding-based filtering, word clouds can help users follow evolving spoken discussions. These works highlight the value of preprocessing speech transcripts to improve word cloud clarity.

The semantic structure of speech also requires more than frequency-based layouts. Wang et al. proposed ReCloud (Wang et al., 2020), which clusters semantically similar terms using NLP techniques, allowing users to grasp themes rather than isolated keywords. Skeppstedt et al. extended this idea with Word Rain (Skeppstedt et al., 2024), embedding word semantics along a visual axis and combining font size with TF-IDF bar charts. Though both methods were tested on written corpora (reviews, climate texts), they underscore how semantic grouping and de-biasing frequency are crucial for domains where redundancy and ambiguity are common.

Together, these studies suggest that effective word cloud generation from speech transcripts must account for semantic ambiguity and high noise levels. This motivates approaches that combine filtering for content-bearing terms and semantically aware tags to produce meaningful visualizations of conversational speech. Our method builds on this trajectory by externalizing grouping decisions to an LLM while preserving analyst control over prompts, topic cardinality, and the final mapping.

2.3 LLM-assisted thematic analysis

LLMs have been used to accelerate theme discovery, propose candidate codes, and reduce analytic burden, sometimes reaching near-human agreement in semi-structured settings. They enable scalable and semi-automated approaches to thematic analysis of qualitative interviews, especially in domains where manual coding is labor-intensive. In the biomedical context, Xu et al. introduced TAMA (Xu et al., 2025), a multi-agent LLM framework designed to assist clinicians in analyzing interviews related to congenital heart disease. By integrating human-in-the-loop feedback with AI-generated theme suggestions, TAMA enhances the accuracy and distinctiveness of identified themes, while significantly reducing the burden on expert coders. Similarly, Singh et al. developed RACER (Singh et al., 2024), an LLM-powered methodology applied to semi-structured interviews conducted during the COVID-19 pandemic. RACER achieved near-human agreement in theme extraction, demonstrating that LLMs can reliably support mental health research involving large volumes of qualitative data.

These successes suggest that concept-level reasoning over long documents is feasible. Our contribution is to harness these capabilities for a narrow but ubiquitous task (first-pass summarization via word clouds) while foregrounding human-centered properties (agency, transparency, workflow fit) that determine whether such tools are practically useful in HCI contexts.

3 Methods

ThemeClouds is designed to assist researchers in generating word clouds from qualitative interview transcripts by using LLMs to surface salient, semantically meaningful concepts, rather than relying on surface-level word frequency. The pipeline consists of three key stages: (1) identifying candidate concepts across a corpus, (2) mapping those concepts to individual transcripts, and (3) aggregating the results to produce a word cloud visualization. Our system prioritizes topic relevance, clarity, and interpretability over lexical frequency or length. Figure 1 outlines the proposed workflow.

We formalize the shift from tokens to concepts and from frequency to breadth. Let $\mathcal{T} = \{t_1, \ldots, t_M\}$ be transcripts (one per participant for a given condition) and let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be short concept-phrases proposed by an LLM for the corpus. For each transcript t and concept c, the mapping step produces a binary assignment $y(t, c) \in \{0, 1\}$ indicating whether the concept is clearly present in the transcript (the artifact optionally supports a soft score $\hat{p}(t, c) \in [0, 1]$ with threshold τ for binarization). The *breadth* of concept c is:

$$b(c) = \sum_{t \in \mathcal{T}} y(t, c),$$

the number of unique participants whose transcripts include the concept. The visual weight for c is $w(c) = g(b(c))$, where $g(\cdot)$ is a monotone scaling (linear by default; logarithmic and square-root options aid mid-rank legibility). We also support condition-wise contrasts by rendering $\Delta b(c) = b_{\mathrm{A}}(c) - b_{\mathrm{B}}(c)$ to make differences across device configurations glanceable.

3.1 Input and preprocessing

The system takes as input a collection of textual transcripts from qualitative interviews. These transcripts may come from usability studies, field interviews, focus groups, or other open-ended sources. Transcripts are assumed to be minimally cleaned

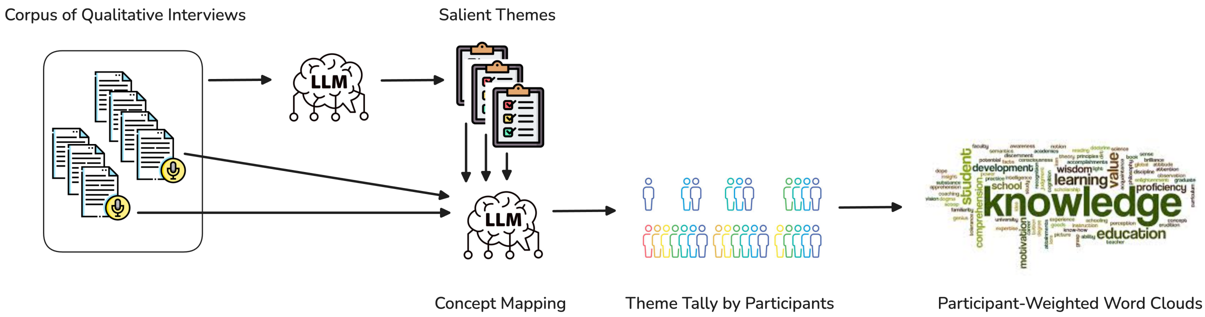


Figure 1: System overview for **ThemeClouds**: LLM-assisted *participant-weighted* thematic word clouds. An LLM first proposes a compact set of concept-level themes for the corpus. Each transcript is then mapped to this fixed theme list via binary presence judgments, yielding a per-theme count of *unique participants* (breadth). The final cloud sizes each theme by its participant prevalence (not token frequency). Prompts, per-transcript assignments, and counts form an audit trail that supports iteration and reproducibility.

(e.g., anonymized and transcribed verbatim) but do not require pre-coding or structuring. Because the method abstracts above tokens, we found that aggressive lexical normalization is unnecessary; we keep punctuation and stop-words intact for the LLM stage, using standard tooling like NLTK only for baseline clouds (Bird, 2006). Interviews are transcribed with Whisper (Radford et al., 2023).

3.2 Concept elicitation (corpus-level)

The goal is a compact, human-interpretable vocabulary that captures salient ideas without collapsing distinct concerns. We prompt a long-context LLM with the corpus (or stratified subsets) to propose N short concept-phrases, encouraging specificity (e.g. "in the way," "felt watched," "image quality"), discouraging generic terms ("user," "good," "bad"), and avoiding fillers or study-task scaffolding. Rather than returning frequent unigrams or bigrams, the model is guided via prompt engineering to prioritize short phrases, semantically specific topics, and coverage diversity across the corpus. We favor a diverse set that covers the thematic space rather than a large list that risks redundancy. The artifact includes our exact prompts and a small set of variations. Analysts can re-run this step to explore granularity. We also explicitly discourage the model from selecting filler words, generic terms like "user" or "system," or concepts that appear frequently but lack thematic depth.

In our evaluation, we prompt a poplar open-source LLM model, LLaMa-3.3-70B-Instruct (Touvron et al., 2023), to identify a set of N salient topics that best represent key concepts across the entire corpus with the following prompt.

```
You are analyzing interview transcripts where participants were asked to share their experiences using five webcam setups: [insta], [single iphone], [dual iphones], [logitech], and [obsbot].
The transcripts are organized in the following format: Each section begins with the webcam label (e.g., "### insta") followed by participant comments about that device.
Ignore filler words, repeated question prompts, or interviewer language. Focus only on participant speech that offers insight, reaction, or description.
Your task is to identify **exactly 20 meaningful and distinctive words or short phrases** that summarize participants' real experiences for **each webcam setup**.
Guidelines:
- Do NOT just pick the most frequent words.
- Select words or short phrases that are **emotionally descriptive**, **technically relevant**, or **highlight distinctive qualities** (positive or negative).
- Avoid: generic words (e.g., "thing", "camera"), filler words, or phrases repeated from the question.
For each setup, return a bullet list of 20 high-quality descriptors.
Output format:
### [setup] - ...
```

The result is a curated list of N topics that act as candidate entries for the word cloud. These phrases serve as a proxy for the major themes in the interviews, as judged by the LLM in context.

3.3 Concept mapping (per transcript)

In the second stage, the LLM is prompted to evaluate each transcript individually in relation to the N identified concepts or insights. For each transcript, the model receives: (1) the full content of that single transcript and (2) the fixed list of N topics produced in the prior step.

The model is then tasked with identifying which topics are clearly present in the given transcript. Importantly, the prompt encourages the model to make binary or categorical judgments rather than assigning soft weights or scores. This helps mitigate overfitting and keeps results interpretable for the end user. We use the following prompt:

```
You are analyzing a participant's response about the **device_name** webcam setup.
Below is a list of key descriptive terms and phrases that were identified across interviews for this webcam. Your task is to determine **which (if any)** of these words or phrases are meaningfully reflected in the participant's comments — even if the exact wording is not used.
Focus on semantic alignment: if a participant implies or clearly expresses a concept that corresponds to one of the key terms, include it.
### Key Descriptive Terms for **device_name**:**keyword_list**
### Output Instructions:
Return ONLY a list of matching terms (one per line).
Do not include explanations, numbering, bullet points, or extra commentary.
A maximum of 20 key descriptive terms and phrases are allowed.
It is imperative to avoid false positives: if a keyword isn't reasonably supported, do not include it.
```

Through the above approach, given $\mathcal{C}$, we map each transcript independently by asking the LLM to judge concept presence using the fixed vocabulary. We default to binary assignments to keep outputs interpretable and to avoid length confounds: loquacious speakers should not inflate weights. Binary judgments also simplify spot-checks: analysts can audit questionable assignments by reading short excerpts of the transcript. The artifact includes an optional soft scoring mode ($\hat{p}(t, c)$) and guidance for threshold selection if analysts prefer graded presence.

This process is repeated for every transcript in the corpus. For each topic, we then compute a relative count of the number of transcripts in which the topic was marked as present. This produces a simple but robust measure of topic salience across the corpus.

3.4 Visualization and contrasts

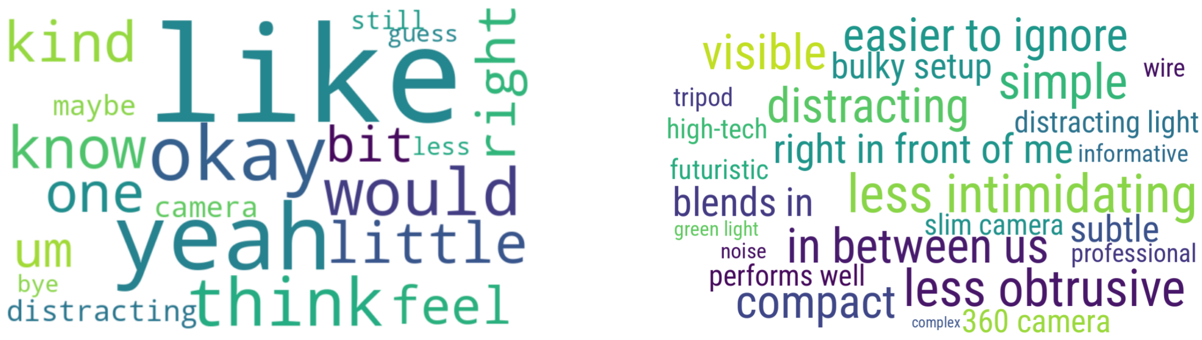
The final step uses these tallied topic counts to construct a word cloud. We render a conventional word cloud where size encodes $w(c)$. Each of the N topics or concepts to be highlighted is included. Because the units are now people, font size directly communicates population-level salience: often the most defensible signal when communicating with product teams or clinical stakeholders. In another word, the font size of each phrase is scaled based on how frequently it was mentioned across the subjects recruited for the qualitative interviews. Terms that were mentioned in most or all transcripts are rendered largest, while rare or marginal topics appear smaller.

For comparative analysis, we can also produce condition-wise "diff clouds" by coloring or separating concepts whose $\Delta b(c)$ exceeds a small margin. This reveals what a device configuration uniquely amplifies or suppresses.

3.5 Analyst-in-the-loop workflow

A central design goal is *researcher agency*. The system includes controls for adjusting the number of topics or concepts to note, the word cloud layout, font scaling, and prompt variants. Analysts can (1) edit the prompt, (2) adjust N, (3) seed or pin concepts they care about, (4) re-run elicitation to split overly broad concepts, and (5) audit and correct per-transcript assignments. This allows researchers to explore different perspectives on their data while retaining interpretability and structure. While the LLM outputs are fixed per run, users can rerun the topic generation with new prompts or adjusted constraints to suit different analytic goals.

We persist an *assignment table* with rows as



(a) Frequency-based word cloud

(b) LLM-assisted word cloud

Figure 2: Side-by-side comparison for one device condition (top 20 items shown). **(a)** A traditional frequency cloud—even after stop-word filtering—elevates conversational surface tokens and fragments paraphrases. **(b)** Our LLM-assisted, participant-weighted ThemeClouds collapses paraphrases into themes and sizes each by the number of unique participants who mentioned it, foregrounding actionable concerns from the interviews.

transcripts and columns as concepts so that any cloud can be reconstructed, inspected, or exported to downstream thematic coding. This audit trail helps teams defend qualitative findings in mixed-methods reports.

4 Human-Centered Design Considerations

A tool succeeds in HCI not only by being accurate, but by fitting how people actually work. We therefore prioritized five properties.

- *Interpretability:* counting unique participants aligns with how analysts argue salience ("many people brought this up").
- *Transparency:* we expose prompts, concept lists, assignment tables, and scaling choices, making it easy to reconstruct decisions or contest them.
- *Agency:* analysts can tune granularity and rerun steps to explore alternative framings.
- *Frugality:* default settings work on small, noisy corpora typical of interviews, without heavy parameter sweeps.
- *Workflow fit:* outputs are designed to triage and guide subsequent coding, not to replace careful qualitative analysis, echoing prior HCI work on semantic grouping and hybrid visual summaries (Iijima et al., 2021; Wang et al., 2020; Chandrasegaran et al., 2019; Skeppstedt et al., 2024).

5 Qualitative Evaluation

To assess the utility of our approach in a real-world setting, we applied it to a set of qualitative interviews conducted as part of a clinical psychology research study. 31 participants evaluated five webcam setups for psychiatric outpatient clinical assessments, producing 155 interviews with a clinical research coordinator. These conversations were conducted as one-hour in-person session, in a naturalistic dialogue format, as participants and the clinical research coordinator collaboratively evaluated different hardware configurations. Audio was transcribed with Whisper (Radford et al., 2023).

For a representative device condition (31 transcripts), Figure 2 compares a standard frequency-based word cloud with NLTK stop-word removal (Bird, 2006) (a) with our LLM-assisted word cloud (b). Despite identical source data, the frequency cloud elevates general discourse terms and fragments multi-word concerns, while the LLM cloud foregrounds concrete, device-specific ideas consistent with researcher notes.

What the numbers mean. Because our weights are counts of unique participants, the magnitude of a label directly translates to breadth. If "distracting" appears in 20 of 31 transcripts for a device, its visual prominence is immediately defensible—helpful for design reviews and IRB or clinical discussions where conservative, population-grounded claims are preferred.

To situate the approach among common baselines, we also trained topic models such as LDA and BERTopic (Rehurek et al., 2011; Grootendorst, 2022; Lin et al., 2023a) on the full 155-document corpus. As in Table 1, given the small per-condition

Table 1: Comparison of outputs from BERTopic, LDA, and our participant-weighted thematic method on the interview corpus. Lists are reproduced from model outputs (verbatim) and our curated themes (top items).

BERTopic (Top Topics)	LDA (Top Topics)	ThemeClouds
1. yeah, like, maybe, okay, whatever, aware, still, um, issue, course	1. 0.005*like + 0.004*okay + 0.004*think + 0.003*would + 0.003*little	1. Small and compact
2. like, part, um, things, process, always, treatment, cause, really, way	2. 0.018*like + 0.008*okay + 0.007*yeah + 0.007*think + 0.007*little	2. Not distracting
3. definitely, would, bit, oh, uncomfortable, good, odd, want, fact, especially	3. 0.004*like + 0.003*um + 0.003*part + 0.003*would + 0.003*things	3. Easy to ignore
4. yeah, like, maybe, okay, whatever, aware, still, um, issue, course	4. 0.004*like + 0.004*okay + 0.003*think + 0.003*little + 0.003*um	4. Less noticeable
5. okay, little, look, think, least, light, um, blends, bright, get	5. 0.099*like + 0.034*little + 0.027*think + 0.026*okay + 0.025*bit	5. Not too visible
6. yeah, even, white, side, either, light, much, like, slightly, around	6. 0.139*like + 0.045*yeah + 0.027*um + 0.027*okay + 0.022*think	6. Fades into the background
7. okay, little, look, think, least, light, um, blends, bright, get	7. 0.174*like + 0.037*um + 0.023*things + 0.020*would + 0.020*part	7. Simple and straightforward
8. definitely, would, bit, oh, uncomfortable, good, odd, want, fact, especially	8. 0.103*like + 0.057*would + 0.036*bit + 0.031*definitely + 0.026*think	8. Convenient
9. yeah, even, white, side, either, light, much, like, slightly, around	9. 0.130*like + 0.029*um + 0.024*yeah + 0.018*would + 0.018*okay	9. Reminds me of a Polaroid
10. like, part, um, things, process, always, treatment, cause, really, way	10. 0.133*like + 0.046*okay + 0.046*little + 0.040*um + 0.040*think	10. Compact and spacious

sample size and conversational style, neither produced immediately legible, per-device themes without additional manual massaging. Our participant-weighted list, on the other hand, aligns closely with analyst field notes and per-device concerns recorded during the study, foregrounding concept-level themes (e.g., "Not distracting," "Discreet," "Blends into the desk") that multiple participants independently raised.

While these observations are not a controlled user study, they illustrate a pattern we frequently saw during analysis: people-weighted concept clouds provide a more faithful "first glance" at what mattered to participants than token frequency or off-the-shelf topic models in this setting. It can effectively support researchers in identifying salient themes from conversational transcripts, even without structured codes or annotations.

6 Discussion and Limitations

Our tool demonstrates how large language models can be leveraged to assist in synthesizing qualitative feedback through semantic word clouds, offering an accessible, low-overhead entry point into exploratory analysis. While initial use cases show alignment with human interpretation, there are important limitations to consider.

6.1 Validity, bias, and controllability

LLM judgments depend on prompts and may overgeneralize. The system relies on static prompts and single-pass outputs, which may overlook nuances or misrepresent concepts without user intervention. We mitigate this by using a fixed vocabulary (reducing drift), binary mapping (reducing verbosity bias), and an assignment table that supports spot-checks and corrections. Analysts can also seed concepts to ensure coverage of domain-critical concerns, an approach compatible with standard qualitative rigor practices.

6.2 Granularity and concept drift

The right granularity is contextual. Collapsing all camera-related concerns might hide distinctions between "felt watched" and "image quality." While prompt customization provides some control, more interactive or iterative workflows could better support researchers in refining outputs over time. Our workflow treats concept elicitation as an iterative process: split or merge concepts, re-run mapping, and compare clouds. We found small N (e.g., 12–25) balanced coverage and legibility, but analysts can tune N to their corpus.

6.3 Generalizability and small-data regimes

The method targets the small, noisy corpora typical of interviews and focus groups. Unlike topic models, which may prefer longer documents or larger datasets, our mapping step scales down: it asks a concrete question of each transcript with a fixed vocabulary. This makes the method robust when M is modest and concepts are grounded in context of the study and clinical application.

6.4 Ethics, privacy, and deployment

Interviews often contain sensitive information. Our artifact documents de-identification assumptions and supports local or compliant deployment. We view LLM assistance as a *scaffold* for human analysis, not a replacement: analysts should verify sensitive claims and avoid over-reliance on automated judgments in consequential settings.

We position people-weighted semantic clouds as a first-pass *orientation* tool. They help teams see what many participants noticed, seed codebooks, and communicate trade-offs. They do not obviate careful reading, synthesis, or theory-building. This stance aligns with prior HCI work that treats semantic grouping and hybrid visual encodings as aids to human reasoning rather than endpoints.

6.5 Interactivity and explanation

Static clouds are useful, but interactive affordances (such as hovering to see exemplar quotes, clicking to open transcripts, showing per-condition contrasts, toggling scaling) can turn the cloud into a navigational entry point for analysis. Because we persist per-transcript assignments, simple linkages suffice. We leave richer explanation (minimal rationales for concept presence) as future work consistent with analyst agency (Iijima et al., 2021; Wang et al., 2020; Chandrasegaran et al., 2019; Skeppstedt et al., 2024).

Future work will focus on improving model transparency, allowing users to inspect why certain phrases were chosen or how decisions were made at the transcript level, for instance in clinical decision support tools such as (Lin et al., 2023b,c, 2025). We are also exploring ways to incorporate multi-turn refinement and lightweight feedback mechanisms, enabling more dynamic human-LLM collaboration. In parallel, more formal evaluations across domains and user roles will be important to assess the tool's effectiveness, trustworthiness, and usability in varied qualitative research contexts.

7 Artifact

Our open-source ThemeClouds package [1] includes: (1) prompt templates for concept elicitation and per-transcript mapping; (2) scripts to reproduce Figure 2; and (3) anonymized assignment tables and per-concept participant counts suitable for auditing and alternative visualizations. The artifact also documents default parameters and prompt variants, so other researchers can reproduce and adapt the pipeline without brittle prompt hacking. We hope this work encourages further exploration into how LLMs can provide insight in qualitative workflows.

8 Conclusion

We introduced ThemeClouds, a participant-weighted, concept-level approach to word clouds using LLMs to count *who* raised *which* ideas, aligning early-stage summaries with the way HCI and UX analysts argue salience. In an audiovisual (AV) study for clinical assessment, the method surfaced actionable concerns that frequency clouds and topic-modeling baselines obscured. By emphasizing transparency, agency, and auditability, it bridges NLP advances and qualitative practice, offering a pragmatic step toward interactive, human-centered, LLM-assisted analysis.

Acknowledgments

We thank the participants and research staff who made this study possible, and colleagues who provided feedback during development. This work is supported by NIH grant 1U01MH136535.

[1] https://github.com/linlab/ThemeClouds

References

Lukas Barth, Sara Irina Fabrikant, Stephen G Kobourov, Anna Lubiw, Martin Nöllenburg, Yoshio Okamoto, Sergey Pupyrev, Claudio Squarcella, Torsten Ueckerdt, and Alexander Wolff. 2014. Semantic word cloud representations: Hardness and approximation algorithms. In *Latin American Symposium on Theoretical Informatics*, pages 514–525. Springer.

Scott Bateman, Carl Gutwin, and Miguel Nacenta. 2008. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 193–202.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72.

Senthil Chandrasegaran, Chris Bryan, Hidekazu Shidara, Tung-Yen Chuang, and Kwan-Liu Ma. 2019. Talktraces: Real-time capture and visualization of verbal content in meetings. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Christel Hopf. 2004. Qualitative interviews: An overview. *A companion to qualitative research*, 203(8):100093.

Ryo Iijima, Akihisa Shitara, Sayan Sarcar, and Yoichi Ochiai. 2021. Word cloud for meeting: A visualization system for dhh people in online meetings. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–4.

Shah Khusro, Fouzia Jabeen, and Aisha Khan. 2021. Tag clouds: past, present and future. *Proceedings of the national academy of sciences, India section A: physical sciences*, 91(2):369–381.

Robert P Lennon, Robbie Fraleigh, Lauren J Van Scoy, Aparna Keshaviah, Xindi C Hu, Bethany L Snyder, Erin L Miller, William A Calo, Aleksandra E Zgierska, and Christopher Griffin. 2021. Developing and testing an automated qualitative assistant (aqua) to support qualitative analysis. *Family medicine and community health*, 9(Suppl 1):e001287.

Baihan Lin, Djallel Bouneffouf, Guillermo Cecchi, and Ravi Tejwani. 2023a. Neural topic modeling of psychotherapy sessions. In *International workshop on health intelligence*, pages 209–219. Springer.

Baihan Lin, Djallel Bouneffouf, Yulia Landa, Rachel Jespersen, Cheryl Corcoran, and Guillermo Cecchi. 2025. Compass: Computational mapping of patient-therapist alliance strategies with language modeling. *Translational Psychiatry*, 15(1):166.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023b. Psychotherapy ai companion with reinforcement learning recommendations and interpretable policy dynamics. In *Companion Proceedings of the ACM Web Conference 2023*, pages 932–939.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023c. Supervisorbot: Nlp-annotated real-time recommendations of psychotherapy treatment strategies with deep reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 7149–7153.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Radim Rehurek, Petr Sojka, and 1 others. 2011. Gensim—statistical semantics in python. *Retrieved from genism. org*.

Anna W Rivadeneira, Daniel M Gruen, Michael J Muller, and David R Millen. 2007. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 995–998.

Satpreet Harcharan Singh, Kevin Jiang, Kanchan Bhasin, Ashutosh Sabharwal, Nidal Moukaddam, and Ankit B Patel. 2024. Racer: An llm-powered methodology for scalable analysis of semi-structured mental health interviews. *arXiv preprint arXiv:2402.02656*.

Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, and Matts Lindström. 2024. From word clouds to word rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization*, 23(3):217–238.

Julie Steele and Noah Iliinsky. 2010. *Beautiful visualization: Looking at data through the eyes of experts.* " O'Reilly Media, Inc.".

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ji Wang, Jian Zhao, Sheng Guo, Chris North, and Naren Ramakrishnan. 2020. Recloud: semantics-based word cloud visualization of user reviews. In *Graphics Interface 2014*, pages 151–158. AK Peters/CRC Press.

Huimin Xu, Seungjun Yi, Terence Lim, Jiawei Xu, Andrew Well, Carlos Mery, Aidong Zhang, Yuji Zhang, Heng Ji, Keshav Pingali, and 1 others. 2025. Tama: A human-ai collaborative thematic analysis framework using multi-agent llms for clinical interviews. *arXiv preprint arXiv:2503.20666*.

Time Is Effort: Estimating Human Post-Editing Time for Grammar Error Correction Tool Evaluation

Ankit Vadehra*
University of Waterloo, Vector Institute
avadehra@uwaterloo.ca

Bill Johnson
Scribendi Inc.
bill.johnson@scribendi.com

Gene Saunders
Scribendi Inc.
gene.saunders@scribendi.com

Pascal Poupart
University of Waterloo, Vector Institute
ppoupart@uwaterloo.ca

Abstract

Text editing can involve several iterations of revision. Incorporating an efficient Grammar Error Correction (GEC) tool in the initial correction round can significantly impact further human editing effort and final text quality. This raises an interesting question to quantify GEC Tool usability: How much effort can the GEC Tool save users? We present the first large-scale dataset of post-editing (PE) time annotations and corrections for two English GEC test datasets (BEA19 and CoNLL14). We introduce Post-Editing Effort in Time (PEET) for GEC Tools as a human-focused evaluation scorer to rank any GEC Tool by estimating PE time-to-correct. Using our dataset, we quantify the amount of time saved by GEC Tools in text editing. Analyzing the edit type indicated that determining whether a sentence needs correction and edits like paraphrasing and punctuation changes had the greatest impact on PE time. Finally, comparison with human rankings shows that PEET correlates well with technical effort judgment, providing a new human-centric direction for evaluating GEC tool usability.[1]

1 Introduction

Grammar Error Correction (GEC) is an important step of the text editing process. There has been a lot of work to build automated GEC tools that can improve the structure and fluency of text while also correcting language errors (Bryant et al., 2023). Since GEC tool-assisted text editing is an iterative process, an editor can make post-edits to the tool output to obtain the closest targeted correction. Estimating the post-editing (PE) effort required to reach the targeted correction can be used as a quality evaluation for the tool.

Human-in-the-loop PE effort was introduced and explored extensively for Machine Translation (MT) (Koponen, 2016) systems. PE effort is studied across three levels (Kittredge, 2002): technical effort, which is the number of edits; cognitive effort, which denotes the psychological assessment required to identify and correct the errors; and temporal effort, which is the total time taken to evaluate and perform post-edits (which includes technical and cognitive effort). Ye et al. (2021) and Tezcan et al. (2019) have explored estimating MT PE time based on edit features. Technical PE effort has also been studied in areas like Text Summarization (Lai et al., 2022), Natural Language Generation (Sripada et al., 2005) and GEC (Rozovskaya and Roth, 2021; Östling et al., 2024).

To incorporate the human editor effort in text correction, we present the first work to consider PE effort in Time (PEET) scores for quality estimation of a GEC tool. The usability of a GEC tool depends inversely on the PE effort to fix the tool output. We release the first large-scale dataset capturing time-to-correct annotations for two English GEC test sets - BEA19 (Bryant et al., 2019) and CoNLL14 (Ng et al., 2014), post-edited from two conditions: the original sentence and the output from two strong GEC tools - GECToR (Omelianchuk et al., 2020) and GEC-PD (Kiyono et al., 2019). We further present a new human-centric GEC Tool evaluation method - PEET Scorer, to estimate the time-to-correct for GEC Tool predictions, which correlates well with human editing effort. As a result, we propose that the PEET scorer can be incorporated along with Post-Editing to evaluate a GEC Tool from a human editor's perspective.

In this work, we make the following contributions:

1. We present the first large-scale GEC dataset with post-editing time-to-correct annotations along with three new high-quality human-

*Corresponding author.
[1]We release our dataset and code at - https://github.com/ankitvad/PEET_Scorer

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 178–196
November 9, 2025 ©2025 Association for Computational Linguistics

preference targeted correction sets for two GEC Test datasets (BEA19 and CONLL14) - source sentence correction and post-edit for two strong GEC Tools (GECToR and GEC-PD) output.

2. We quantify the editing time saved and improvement in final correction quality (estimated using GEC metrics) using GEC Tools for first-pass text-editing. We also observe that determining whether a sentence needs correction and edits like paraphrasing and punctuation changes has the greatest impact on time-to-correct.

3. We contribute a new evaluation method called PEET Scorer that can be used to rank any GEC Tool in terms of time-to-correct. We compare the PEET scorer with 3 human judgment rankings of 33 GEC Tools, and demonstrate high correlation with further correction effort required.

2 Background Work

2.1 Grammar Error Correction (GEC) Tools

GEC tools can be broadly divided into supervised-trained, LLM-based, and ensemble-ranked models (Omelianchuk et al., 2024).

The supervised GEC tools can be divided into edit-based and sequence-to-sequence models. Edit-based models convert the task to a sequence-tagging and editing approach where each token in the input sentence is assigned an edit operation. Some tools that use this approach are the PIE (Awasthi et al., 2019) and GECToR (Omelianchuk et al., 2020; Tarnavskyi et al., 2022) models. Sequence-to-Sequence (S2S) GEC Tools utilize an encoder-decoder architecture where the corrected sentence is generated for each input sentence (Choe et al., 2019; Grundkiewicz et al., 2019; Kiyono et al., 2019).

Large language models like Llama (Touvron et al., 2023; Omelianchuk et al., 2024) and ChatGPT (Katinskaia and Yangarber, 2024) also perform well for GEC (Zhang et al., 2023; Fang et al., 2023b) in different settings like - Zero-Shot, Few-Shot and Fine-Tuning (Korniienko, 2024; Davis et al., 2024; Raheja et al., 2023). The current state-of-the-art GEC tools all rely on the approach of ensembling multiple strong GEC Tools, aggregating them with methods like majority votes (Tarnavskyi et al., 2022) and logistic regression (Qorib and Ng, 2023; Qorib et al., 2022).

In this work, we use two supervised GEC tools for first-pass text editing: GECToR edit tagging (Omelianchuk et al., 2020) and GEC-PseudoData (GEC-PD) (Kiyono et al., 2019) model, which was trained on a large synthetic corpus. The output of these models is further corrected by human editors while tracking the time-to-correct (temporal effort). We use this time dataset to quantify the impact of GEC tools for text-editing, observing reduced post-editing time and better quality final correction (Section 3.5). Even though the GEC Tools we selected (GECToR and GEC-PD) are not the most recent, they are on par with human-level performance as demonstrated in Section 3.4 - Table 3.

2.2 Post Editing Effort in Machine Translation

Post Editing Effort (PEE) for Quality Estimation is an actively researched task in Machine Translation (MT). It evaluates the output of an MT system for quality and correctness (Senez, 1998; Specia, 2011). Post-editing (PE) the output of an MT system can improve the final translation quality compared to translating the source from scratch, while improving overall editor productivity (Plitt and Masselot, 2010; Guerberof, 2009; Green et al., 2013). We briefly review previous work in MT that explores PEE across three levels (technical, cognitive and temporal effort) (Kittredge, 2002).

Technical effort has been defined by edit distance metrics like - Translation Edit Rate (TER) and Human TER (Snover et al., 2006) as well as keystroke and edit operation logging (Barrachina et al., 2009; O'Brien, 2005; Carl et al., 2011). Cognitive effort has also been studied in terms of edit complexities (Temnikova, 2010; Koponen et al., 2012; Popović et al., 2014; Daems et al., 2017) and human-assessed quality judgment and ranking (Specia et al., 2009, 2011; Koponen, 2012). Keystroke logs to determine pause information (O'Brien, 2005; Carl et al., 2011), eye gaze tracking and pause fixation (Vieira, 2014; Hvelplund, 2014; Daems et al., 2015) and Thinking Aloud Protocol (TAP) (Kittredge, 2002; Vieira, 2017; O'Brien, 2005) have also been proposed as measures of cognitive effort. The work on Temporal Effort in MT estimates the relationship between the time-to-correct and different evaluation metrics (Tatsumi, 2009), source/target translation characteristics (Tatsumi and Roturier, 2010), and quality estimation (Specia, 2011). Zaretskaya et al. (2016) and Popović et al. (2014) study the average temporal effort required for each error type by considering the time-to-correct and frequency of error edits. Finally, Ye et al. (2021) and Tezcan et al. (2019) train models to estimate the post-editing time based on PE features.

PE has also been explored in tasks like Text Summarization (Lai et al., 2022) and Cognitive and Technical PE Effort has been studied for Grammar Error Correction (GEC) evaluation.

2.3 Post Editing Effort in Grammar Error Correction

We review previous work in GEC that closely relates to post-editing (PE) effort across two levels (cognitive and technical effort). To the best of our knowledge, temporal effort for PE has not been explored for GEC tools.

2.3.1 Cognitive Post Editing Effort

Although cognitive PE effort has not been measured directly for GEC, Human judgment rankings of GEC Tools (Grundkiewicz et al., 2015; Kobayashi et al., 2024; Napoles et al., 2019), which are an estimate of perceived cognitive effort, have been used extensively for GEC evaluation metric assessment. Reference-based GEC metrics like ERRANT (Bryant et al., 2017), M^2 (Dahlmeier and Ng, 2012), GoToScorer (Gotou et al., 2020), and GLEU (Courtney et al., 2016) and reference-less metrics like PT-M^2 (Gong et al., 2022), Scribendi Score (Islam and Magnani, 2021), SOME (Yoshimura et al., 2020) and IMPARA (Maeda et al., 2022) designed to estimate GEC Tool quality are trained and evaluated using the GEC human judgment rankings.

However, perceived cognitive effort does not always agree with the actual PE effort and can be subjective. Sentence correction experiments in GEC have shown poor cognitive agreement between editors. Tetreault et al. (2014) and Tetreault and Chodorow (2008) asked 2 native English speakers to insert a preposition into 200 sentences, from which a single preposition was removed, obtaining an agreement score of just 0.7. Rozovskaya and Roth (2010) asked three annotators to evaluate and mark 200 sentences for correctness, showing a poor pairwise agreement between them $(0.4, 0.23, 0.16)$. Finally, there has been some work considering the cognitive proficiency of the user interacting with a GEC Tool (Nadejde and Tetreault, 2020) and the annotators who create the evaluation references of GEC test sets (Takahashi et al., 2022; Napoles et al., 2017).

Surprisingly, none of the GEC metrics described above have considered using targeted references (target obtained after correcting the GEC Tool output) to estimate the tool usability dependent on human PE effort.

2.3.2 Technical Post Editing Effort

To the best of our knowledge, only two prior studies have explored the impact of PE technical effort on GEC evaluation. Rozovskaya and Roth (2021) introduced targeted references for English and Russian datasets and Östling et al. (2024) utilize PE references to assess Swedish GEC Tools. The studies show that GEC evaluation using untargeted references ignores the human subjectivity involved in text correction. For instance, the SEEDA - human judgment rankings from Kobayashi et al. (2024) compared the correction outputs of GPT3.5, human editors and various Neural GEC Tools. The GPT-3.5 and human corrections were ranked significantly higher and contained nearly two and three times more edits than other corrections. As a result, these high-quality corrections obtain poor evaluation scores when compared against untargeted references. This inconsistency highlights the importance of PE for GEC Tool evaluation, to capture the true technical effort.

Apart from estimating the PE effort, targeted references can also be used for fine-tuning and aligning Large Language Models (LLMs) with human preferences to generate better outputs (Li et al., 2024).

2.3.3 Temporal Post Editing Effort

We introduce the first work to study the Temporal Effort in PE for GEC. Temporal effort described in terms of time-to-correct can efficiently capture the overall PE effort. We present the first large-scale dataset of post-edited corrections along with their temporal effort annotations for two strong GEC tools, GECToR (Omelianchuk et al., 2020) and GEC-PD (Kiyono et al., 2019), outputs on two English GEC Test sets - CONLL14 (Ng et al., 2014) and BEA19 (Bryant et al., 2019). We also use this dataset to quantify the impact of GEC Tools in Text Editing and the contribution of different edit types to the human post-editing effort. We present PEET Scorer, a regression-based metric, to estimate the time-to-correct scores, which can be incorporated along with post-editing to evaluate the usability of GEC Tools in a human-centred manner.

3 Dataset Collection and Processing

An important component in this work is the high-quality dataset of post-edit corrections for GEC, along with their time-to-correct (temporal effort) annotations. We partnered with a professional text-editing company - Scribendi Inc.[2] to collect this data. This section explains our dataset collection, filtering, and quality estimation process.

3.1 Dataset Source

We use source sentences from two popular English GEC test sets - CONLL14 (Ng et al., 2014) and BEA19 (Bryant et al., 2019) ($1312 + 4477 = 5789$ sentences). Each sentence was corrected in three variations: the source and post-editing outputs from Two GEC Tools - GECToR (Omelianchuk et al., 2020) and GEC-PD (Kiyono et al., 2019) (Section 2.1). Each sentence variation was corrected by 1 out of 8 professional text editors, employed by Scribendi Inc. This resulted in a dataset of $5789 * 3 = 17367$ target corrections along with their time-to-correct scores.

3.2 Editor Correction Framework

The source sentence and GEC Tool output serve as the basis for further editor correction. This follows the real framework for Text Editing, where a GEC Tool output is evaluated for further correction, compared with the original sentence. The editors were given GEC post-editing (PE) instructions (Appendix F-3) and asked to perform minimal edits and avoid rewrites. We used the Qualtrics[3] survey tool to collect PE corrections and enabled the "Timing Question" to log time-to-correct for each source sentence. All other metadata logging was disabled.

The 3 variations for each sentence - source, GECToR and GEC-PD output- were given to a different professional editor (in a pool of 8 editors) to eliminate any time-to-correct bias. The task of evaluating $17,367$ sentences was performed in batches of 50. The editors were shown the source sentence and the first-pass GEC Tool output (Appendix F-4). The final target correction and time-to-correct were logged for each sentence. For source sentence correction, only the original sentence was presented.

[2] https://www.scribendi.com/

[3] https://www.qualtrics.com/

3.3 Data Filtering

To improve the dataset quality, we perform two stages of data filtering on the 3 target correction sets for each source (17367 sentences initially). In the first stage, we eliminate outliers based on the logged time-to-correct. Snover et al. (2006) showed that editors took between 3-6 minutes for each correction. Considering this and the distribution of the time-to-correct in our dataset, we filter corrections that took more than 250 seconds (17033 sentences remaining). Finally, we merge duplicate corrections from our dataset by averaging the time-to-correct values (14112 sentences dataset). This filtering allows us to retain 81.26% of our dataset that we use as train and test sets (80:20 split) for the Post-Editing Effort in Time (PEET) Scorer.

3.4 Correction Quality

We collect and present three new target corrections for the CONLL14 (Ng et al., 2014) and BEA19 (Bryant et al., 2019) test datasets. The correction for the source and two post-edited target corrections. We evaluate the quality of the target corrections using the official GEC competition metric and the Inter Annotator Agreement (IAA) scores. Each target correction set can be divided into CONLL14 and BEA19 corrections. We evaluate the CONLL14 and BEA19 target corrections separately.

Correction	M2 Score	(Precision : Recall)
A1	46.9	44.6 : 59.1
A2	53.0	51.7 : 59.5
A3*	98.6	98.7 : 98.3
A4	55.3	54.9 : 57.0
A5	52.8	51.3 : 59.7
A6	56.4	55.8 : 58.8
A7*	98.6	98.7 : 98.5
A8	53.5	53.8 : 52.6
A9	55.7	55.6 : 56.0
A10	52.8	51.3 : 59.4
c1	50.9	49.0 : 60.4
c2	52.3	50.5 : 61.0
c3	53.7	52.1 : 60.8

Table 1: The M2 precision and recall quality score for all Bryant and Ng (2015) target correction sets for the official CONLL14 competition task.

Bryant and Ng (2015) released 10 additional target corrections for the CONLL14 test dataset. We compare the quality scores of our 3 corrections with theirs using the official CONLL14 competition - M2 Scorer (Ng et al., 2014) metric. Table 1 shows the M2 scores for all target correction sets - Bryant and Ng (2015) corrections $A1 - A10$, and our corrections $c1 - c3$. Corrections A3 and A7 obtain near-perfect quality scores, since they were generated by the 2 editors who created the official CONLL14 competition target references (Bryant and Ng, 2015). Ignoring the 2 outliers, we observe similar quality scores for our corrections. This indicates that our 3 CONLL14 Target corrections are of similar high quality. Unfortunately, there are no public correction references available for the BEA19 Test set (this work being the first to present 3 target references), making it hard to compare the quality scores directly.

To overcome this issue, we calculate the quality scores for the 3 target correction sets and the GEC-Tool first-pass outputs on the official BEA19 and CONLL14 competitions and compare trends between the correction sets. We use the BEA19 competition website scorer[4] to evaluate the performance of BEA19 target corrections. Table 2 shows the quality scores for the GECToR and GEC-PD Tool output and the final editor target corrections (EC).

Similar trends are observed between the CONLL14 and BEA19 target correction sets. We observe a significant increase in Recall scores for the EC compared to the first-pass GEC Tool output. This indicates the final EC target contains additional post-edit corrections missed by the GEC Tool. The reduction in the precision score for EC is consistent with the 10 CONLL14 target corrections released by Bryant and Ng (2015) since post-editing often leads to subjective paraphrasing and rewrite edits, which may not be present in the official competition target reference. The final EC obtained better Recall scores compared to the State-of-the-Art (SOA) GEC Tool - GRECO (as of writing this paper) (Qorib and Ng, 2023) for both datasets. Observing similar quality score trends for the GEC Tool predictions and our target EC across both CONLL14 and BEA19 Test competition, and better Recall than the SOA GRECO tool, we can infer that the 3 target corrections collected by us in this work are of high quality.

We also use the GEC Inter Annotator Agreement (IAA) framework proposed by Bryant and Ng (2015) to compare the target correction sets for both datasets with themselves to ensure better consistency and quality. The IAA framework states that the $F_{0.5}$ multi-reference score, used to evaluate a GEC Tool-vs-human corrections, can similarly evaluate human-vs-human corrections. When comparing multiple annotator corrections, a single correction set can be compared using the rest as references to get quality scores. The final IAA score is calculated as the average of all correction set scores. We use the ERRANT tool (Bryant et al., 2017) to perform the IAA evaluation. We evaluate 3 target correction sets:

$A = \{A1 - A10\}$ The 10 target corrections for CONLL14 by Bryant and Ng (2015).

$C = \{c1, c2, c3\}$ The 3 CONLL14 target corrections collected by us.

$B = \{b1, b2, b3\}$ The 3 BEA19 target corrections collected by us.

To compare IAA scores, we conduct a 1-vs-2 target correction set evaluation. For each correction in A, we randomly select 2 corrections from the remaining 9 as the reference. Scores for each correction in B and C are calculated using the remaining 2 corrections as target references. Table 4 shows the average IAA scores for A, B, C correction sets. We observe better Avg-IAA scores for the C and B correction sets collected by us in this work, compared to A.

To ensure we choose strong GEC Tools (Section 2.1) to obtain first-pass output predictions, we compare the quality of the GEC Tool output and the subsequent human EC. We consider the Source Sentence EC (collected by us) as the target reference for the BEA19 and CONLL14 Test sets. The $F_{0.5}$ quality scores obtained in Table 3 show similar performance between the GECToR and GEC-PD Tool prediction output and the subsequent EC because of the variation in Precision and Recall scores. This indicates that GECToR and GEC-PD are strong first-pass GEC Tools.

3.5 Impact of GEC Tools

Comparing the time-to-correct for the source sentence versus the GEC Tool output post-editing, we can quantify the impact of using GEC Tools in Text Editing.

Quality scores presented in Table 2 show that the GEC Tool output EC has better values compared to

[4]BEA19 GEC competition website - https://codalab.lisn.upsaclay.fr/competitions/4057

Candidate Set	BEA19 Test (P : R : $F_{0.5}$)	CONLL14 Test (P : R : $F_{0.5}$)
Source Sentence	-	-
Source Sentence EC	45.30 : 66.08 : 48.34	49.05 : 60.45 : 50.97
GECToR Output	66.81 : 58.42 : 64.94	63.97 : 45.94 : 59.31
GECToR Output EC	48.24 : 71.38 : 51.59	50.50 : 61.09 : 52.31
GEC-PD Output	66.20 : 61.48 : 65.20	64.06 : 44.92 : 59.03
GEC-PD Output EC	47.33 : 70.54 : 50.66	52.17 : 60.86 : 53.71
GRECO Model Output	86.45 : 63.13 : 80.50	79.36 : 48.69 : 70.48

Table 2: Quality Scores of the 2 GEC Tools output prediction, target Editor Corrections (EC) and the State-of-the-Art GEC Tool - GRECO (Qorib and Ng, 2023) on the official BEA19 and CONLL14 competition.

Candidate Set	BEA19 Test (P : R : $F_{0.5}$)	CONLL14 Test (P : R : $F_{0.5}$)
GECToR Output	52.59 : 28.59 : 45.03	57.74 : 25.10 : 45.82
GECToR Output EC	45.47 : 47.91 : 45.94	44.31 : 43.53 : 44.15
GEC-PD Output	49.88 : 26.37 : 42.33	56.49 : 23.13 : 43.85
GEC-PD Output EC	45.90 : 48.31 : 46.36	46.14 : 42.64 : 45.39

Table 3: Quality Scores of the 2 GEC Tools output predictions and their final target Editor Corrections (EC) using the BEA19 and CONLL14 - Source Sentence EC as target reference.

Human Annotation Set	Reference Set and Size	IAA Score - $F_{0.5}$
$A1$	$\|\{RAND(2) \in \{A - A1\}\| = 2$	36.21
$A2$	$\|\{RAND(2) \in \{A - A2\}\| = 2$	45.48
$A3$	$\|\{RAND(2) \in \{A - A3\}\| = 2$	46.72
$A4$	$\|\{RAND(2) \in \{A - A4\}\| = 2$	40.54
$A5$	$\|\{RAND(2) \in \{A - A5\}\| = 2$	46.01
$A6$	$\|\{RAND(2) \in \{A - A6\}\| = 2$	50.85
$A7$	$\|\{RAND(2) \in \{A - A7\}\| = 2$	42.72
$A8$	$\|\{RAND(2) \in \{A - A8\}\| = 2$	49.46
$A9$	$\|\{RAND(2) \in \{A - A9\}\| = 2$	52.0
$A10$	$\|\{RAND(2) \in \{A - A10\}\| = 2$	48.57
Avg-IAA $\{A\}$	$\{A\}, 2$	**45.85**
$c1$	$\|\{C - c1\}\| = 2$	54.11
$c2$	$\|\{C - c2\}\| = 2$	57.36
$c3$	$\|\{C - c3\}\| = 2$	59.14
Avg-IAA $\{C\}$	$\{C\}, 2$	**56.87**
$b1$	$\|\{B - b1\}\| = 2$	57.94
$b2$	$\|\{B - b2\}\| = 2$	59.39
$b3$	$\|\{B - b3\}\| = 2$	59.81
Avg-IAA $\{B\}$	$\{B\}, 2$	**59.05**

Table 4: Inter Annotator Agreement (IAA) scores for the different A, B, C annotation sets using the ERRANT $F_{0.5}$ metric. RAND(n) represents a random selection of "n" items from the respective set.

the Source Sentence EC. In Table 5, we compare the time taken (in seconds) by a human editor to correct the source sentences with and without first-pass editing by a GEC tool. We observe that GEC Tools help in reducing the post-editing time by roughly 4 seconds per sentence. Combined insights from these results indicate that incorporating GEC Tools in the text-editing workflow reduces editing time and generates better final target corrections. Thus, GEC Tools can help improve editor efficiency

Sentence Source	Average Time per Sentence	Average Time per Word
Source Sentence	31.16	1.91
GECToR Output	26.82	1.57
GEC-PD Output	27.46	1.67

Table 5: The average time to correct (**in seconds**) for a sentence and word; correcting the source and after first-pass GEC Tool editing.

and overall productivity.

4 Methodology

We design statistical and neural network regression models for our post-editing effort in time (PEET) scorer. The scorer is trained to estimate the time-to-correct value for a source sentence given the target correction, using the number and type of edits and sentence property - Sentence Length, Correct/Incorrect.

The dataset that we collected contains 3 iterations for all 3 variations of the source - source (SRC), GEC Tool Model Output (MO) and post-edited target correction (TRG). Different training features in terms of edits and sentence structure can be selected and extracted from - SRC, MO and TRG triple (Appendix D).

Statistical PEET models performed as well as Neural models while allowing greater interpretability of training features (Appendix A). Also, models using features selected from $[MO, TRG]$ sentences performed better than models trained on fine-grained features from $[SRC, MO, TRG]$ sentences (Appendix E). Hence, we only discuss the features and results of the Statistical PEET Model trained using the $[MO, TRG]$ sentences here, referring to MO as the source.

4.1 ERRANT Edit Feature Extraction

We use ERRANT (Bryant et al., 2017) to align and extract edit features between the source and target corrections (Appendix B). Apart from the edit category - Removal(R), Missing(M) and Unnecessary(U), the feature also includes the edit type. Figure 1 lists the different edit categories and their syntactic type generated by ERRANT.

We use the number and type of edits as features for our statistical models. Similar to the edit type

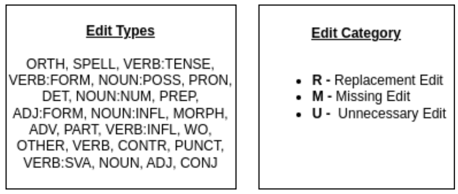


Figure 1: ERRANT edit category and types.

hierarchy used by Yuan et al. (2021), considering category, type and their combination can provide 4, 25 or 55 edit features. For instance, if we only consider the 3 edit categories, then our 4 edit features are Replacement(R), Missing(M), Unnecessary(U) and Correct/Incorrect (binary feature). Using the 24 edit types (Figure 1) and Correct/Incorrect gives us 25 edit features. Similarly, combining edit categories with their possible types, we get 55 edit features (see Table 14 in Appendix G). We train separate models for all three edit levels (4, 25, 55).

4.2 PEET Scorer Models

We design Linear Regression (LR) and Support Vector Regression (SVR) models, for our PEET Scorer, using the ERRANT Edit count and different edit type levels (4, 25, 55), number of edited words, source and target sentence length as features. We also experimented with Neural Regression models, but they didn't perform better than statistical models (Results in Appendix A). We only discuss the results of the statistical PEET models here. The details of each model and the hyperparameters are presented in Appendix C.

The PEET estimation task has a continuous range of prediction values - time (in seconds). We report the mean absolute error (MAE) and Pearson correlation (r) between the predicted time and the target time. We note that MAE does not take into account the sign of the error, while correlation does (Graham, 2015; Tezcan et al., 2019), which is why we report correlation and use it to compare model performance.

5 Experiment Results

5.1 Performance of the PEET Scorer

The results for the Linear Regression (LR) and SVR PEET Scorer, with count of different edit feature levels (4,25,55), sentence word length and number of word edits as features (Section 4.1), are presented in Table 6.

The statistical models relying on edit type information (25,55 labels) performed better than using

Statistical Model	Edit Feature Level	r	MAE
Linear Regression	4	0.559	18.92
	25	**0.565**	18.74
	55	0.563	18.75
SVR Linear	4	0.558	16.40
	25	0.564	16.19
	55	**0.565**	16.15

Table 6: Average PEET estimation performance for the Statistical Models over 50 runs (different train-test data seed). The results are presented as the Pearson Correlation (r), Mean Absolute Error (MAE) loss.

minimal substitution, deletion and insertion edit category labels (Figure 1). This indicates that the type of edit has an impact on post-editing effort. We obtain a correlation of $r = 0.565$ from the best models (LR-25 edit features).

5.2 Impact of Error Types on Post-Edit Effort

We follow the work by Ye et al. (2021), using regression coefficients of a Linear Regression (LR) model to estimate the PEET impact of different edit features. To make the coefficients interpretable, we center and standardize all edit-features by subtracting the mean and dividing by the standard deviation (except the binary/categorical edit feature - Correct/Incorrect) (Schielzeth, 2010).

Model Features	Regression Coefficient	Model Features	Regression Coefficient	Model Features	Regression Coefficient
OTHER	10.15	**ORTH**	2.34	**ADJ**	0.97
PUNCT	4.55	**CONJ**	2.03	**CONTR**	0.78
PREP	4.03	**MORPH**	1.89	**VERB:INFL**	0.63
VERB	3.37	**SPELL**	1.87	**PART**	0.47
Sentence Correct	-3.31	**ADV**	1.79	**ADJ:FORM**	0.39
NOUN	3.23	**VERB:FORM**	1.66	**NOUN:INFL**	-0.30
DET	3.08	**WO**	1.63	**NOUN:POSS**	0.25
NOUN:NUM	2.52	**VERB:SVA**	1.16	-	-
VERB:TENSE	2.35	**PRON**	1.10	-	-

Table 7: The standardized regression coefficients of the LR model trained on the medium (25) edit features to measure the impact of each feature on PEET estimation.

The edit category *OTHER*, which corresponds to paraphrasing or rewriting text, and modifying punctuation has the highest impact on post editing time. Deciding whether a particular sentence is incorrect also contributes significantly to the post-editing effort. The coefficients to study the impact of the 25 edit features are shown in Table 7. Coefficients for the other edit granularities (4 and 55 labels) and all PEET sentence features are provided in Appendix G.

5.3 PEET Scorer for GEC Quality Estimation

Since an efficient GEC Tool would reduce post-editing (PE) time, PE followed by PEET estimation can quantify the usability of a GEC Tool (Specia, 2011). To study the correlation between cognitive, temporal and technical PE effort, we compare the PEET scorer rankings with human judgment rankings (HJR) (Section 2.3) and Word Error Rate (Technical Effort) of GEC Tools. We evaluate the PEET-Linear Regression (25 Edit Features) Scorer (Section 4.1) estimated ranking for 33 GEC Tools in 3 GEC HJR (Appendix H).

- *Grundkiewicz-C14(EW)* - ranking of 12 GEC Tools that participated in the official CONLL-14 - GEC Task (Ng et al., 2014) by Grundkiewicz et al. (2015).
- *SEEDA-C14-All(TS)* - ranking of 15 newer and stronger GEC Tools on the CONLL-14 test dataset by Kobayashi et al. (2024). *SEEDA-C14-NO(TS)* denotes the subset of 12 GEC tools without the 3 outliers.
- *Napoles-FCE* and *Napoles-Wiki* - ranking of 6 Seq2Seq GEC Tools on the FCE (Yannakoudakis et al., 2011) and WikiEd (Grundkiewicz and Junczys-Dowmunt, 2014) datasets by Napoles et al. (2019).

Human Judgment Ranking	PEET Metric		WER	
	ρ	r	ρ	r
Grundkiewicz - C14 (EW)	0.48	0.26	0.28	0.18
SEEDA - C14 - All (TS)	0.18	0.63	0.18	0.65
SEEDA - C14 - NO (TS)	-0.1	-0.27	-0.1	-0.33
Napoles - FCE	**-0.96**	-0.94	**-0.96**	-0.88
Napoles - Wiki	**-0.71**	-0.63	**-0.93**	-0.88

Table 8: The correlation of our PEET model ranking with human-judgment rankings (HJR). We also provide the correlation of the HJR with the Word Edit Rate (WER) metric. Spearman (ρ) and Pearson (r) correlation scores are used for comparison. A high negative correlation indicates lower time-to-correct and WER score corresponding to a higher human judgment ranking.

The *Grundkiewicz-C14* and *SEEDA-C14* human ranking calculation was conducted using the Expected Wins (EW) (Bojar et al., 2013) and TrueSkill (TS) (Herbrich et al., 2007) method, which tracks relative ranking based on a set-wise comparison of a subset of all GEC Tool corrections. The EW and TS rankings were selected for the final *Grundkiewicz-C14* and *SEEDA-C14* rankings, respectively. The *Napoles - FCE* and *Napoles - Wiki* human ranking addressed the issue of partial comparison and relative ranking for GEC Tools by using the partial ranking with scalars (PRWS) method (Sakaguchi and Van Durme, 2018), collecting a quality score (0-100) for each sentence to infer the final rankings.

Table 8 shows the Pearson (r) and Spearman (ρ) correlation scores of the HJRs with the PEET model ranking and the Word Error Rate (WER) (Snover et al., 2006) (number of edits required to correct a GEC Tool prediction). The WER and PEET are calculated using untargeted references, which contributes to the lower alignment with perceived cognitive effort judgment.

We observe a good alignment (high negative correlation) between the PEET ranking and the *Napoles* HJR and a poor alignment (positive correlation) with the other HJRs. The PEET ranking shows better alignment to HJRs that align with WER scores (Technical PE effort - Section 2.3). We also observe that human quality rankings collected using PRWS align better with true human effort (WER) than those collected using TS or EW.

These results suggest that our PEET Scorer can estimate GEC Tool usability when output quality depends on further Post-Editing Effort (WER and type of edits) required to correct the tool output. Hence, performing PE to obtain the closest correction (lower WER) can improve GEC temporal effort estimation.

6 Conclusion and Future Work

Since we present the first study and dataset of Post-Editing Effort (PEET) in Time for GEC, our goal is to provide a baseline for future work in this area. Using our dataset, we quantified the editor efficiency and productivity when using GEC Tools for Text Editing. We extract various automated sentence properties and edit type features from the sentence correction pairs to train the PEET Scorer. Recently, there has been some work in the area of Grammar Error Explanation to define descriptive error types (Fei et al., 2023; Ye et al., 2025) and use LLMs for error explanation (Song et al., 2023; Li et al., 2025). As future work, the descriptive edits can be used as possible features for the PEET model. Finally, we observe that our PEET model works well for GEC Tool evaluation when the output quality is dependent on the Technical PE Effort (amount of edits). Studying actual cognitive effort for GEC post-editing and how it compares with technical and temporal effort is another interesting direction for future work.

Limitations

One of the main limitations of Post-Editing (PE) Effort estimation is incorporating human annotation to evaluate GEC Tool performance, which can be expensive. However, PE allows us to quantify the true performance from a human-in-the-loop perspective. Currently, our work is limited to automated edit-type features generated by the ERRANT toolkit (Bryant et al., 2017). Evaluating our PEET Scorer as a GEC quality estimation tool shows that it is effective when the correction quality is dependent on the technical post-editing effort. However, similar to work in Machine Translation, it is inconsistent with quality estimation based on perceived PE efforts. Finally, we acknowledge that our work is limited to only the English language. Future work on post-editing GEC for other languages can show the impact of language type on PEET for GEC.

Acknowledgments

We thank staff and colleagues at Scribendi Inc., Chatham, Ontario (`www.scribendi.com`) for the grant, input and feedback during the research and manuscript writing phases for this project. Resources used in this work were provided by the Province of Ontario, the Government of Canada through CIFAR, companies sponsoring the Vector Institute (`https://vectorinstitute.ai/partnerships/current-partners/`), the Natural Sciences and Engineering Council of Canada and a grant from IITP & MSIT of Korea (No. RS-2024-00457882, AI Research Hub Project).

References

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4260–4270.

Srinivas Bangalore, Bergljot Behrens, Michael Carl, Maheshwar Gankhot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm. 2015. The role of syntactic variation in translation and post-editing. *Translation Spaces*, 4(1):119–144.

Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and 1 others. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *WMT@ACL*.

Tiberiu Boroş, Stefan Daniel Dumitrescu, Adrian Zafiu, Verginica Barbu Mititelu, and Ionut Paul Văduva. 2014. Racai gec–a hybrid approach to grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48.

Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. 2019. The bea-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75.

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

CJ Bryant, Mariano Felice, and Edward Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics.

Michael Carl, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. 2011. The process of post-editing: A pilot study. *Copenhagen Studies in Language*, 41(1):131–142.

Yo Joong Choe, Jiyeon Ham, Kyubyong Park, and Yeoil Yoon. 2019. A neural grammatical error correction system built on better pre-training and sequential transfer learning. *arXiv preprint arXiv:1907.01256*.

Shamil Chollampatt and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Napoles Courtney, Sakaguchi Keisuke, Post Matt, R Tetreault Joel, and 1 others. 2016. Gleu without tuning. *arXiv*.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Joke Daems, Sonia Vandepitte, Robert J Hartsuiker, and Lieve Macken. 2017. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in psychology*, 8:1282.

Joke Daems, Sonia Vandepitte, Robert Hartsuker, and Lieve Macken. 2015. The impact of machine translation error types on post-editing effort indicators. In *Proceedings of the 4th Workshop on Post-editing Technology and Practice*.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572.

Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. 2021. Does syntax matter? a strong baseline for aspect-based sentiment analysis with roberta. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1816–1829.

Christopher Davis, Andrew Caines, Øistein Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of english learner text. *arXiv preprint arXiv:2401.07702*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Tao Fang, Xuebo Liu, Derek F Wong, Runzhe Zhan, Liang Ding, Lidia S Chao, Dacheng Tao, and Min Zhang. 2023a. Transgec: Improving grammatical error correction with translationese. In *Findings of the association for computational linguistics: ACL 2023*, pages 3614–3633.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023b. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501, Toronto, Canada. Association for Computational Linguistics.

Mariano Felice, Zheng Yuan, Øistein E Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the eighteenth conference on computational natural language learning: shared task*, pages 15–24.

Peiyuan Gong, Xuebo Liu, He-Yan Huang, and Min Zhang. 2022. Revisiting grammatical error correction evaluation and beyond. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6891–6902.

Takumi Gotou, Ryo Nagata, Masato Mita, and Kazuaki Hanawa. 2020. Taking the correction difficulty into account in grammatical error correction evaluation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2085–2095.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The wiked error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *Advances in Natural Language Processing: 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings 9*, pages 478–490. Springer.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263.

Ana Guerberof. 2009. Productivity and quality in mt post-editing. In *Beyond Translation Memories: New Tools for Translators Workshop*.

Anubhav Gupta. 2014. Grammatical error detection using tagger disagreement. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 49–52, Baltimore, Maryland. Association for Computational Linguistics.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*. The MIT Press.

S. David Hernandez and Hiram Calvo. 2014. CoNLL 2014 shared task: Grammatical error correction with a syntactic n-gram language model from a big corpora. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 53–59, Baltimore, Maryland. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kristian Tangsgaard Hvelplund. 2014. Eye tracking and the translation process: Reflections on the analysis and interpretation of eye-tracking data.

Md Asadul Islam and Enrico Magnani. 2021. Is this the end of the gold standard? a straightforward reference-less grammatical error correction metric. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3009–3015.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2014. The amu system in the conll-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 25–33.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online. Association for Computational Linguistics.

Anisia Katinskaia and Roman Yangarber. 2024. GPT-3.5 for grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Richard Kittredge. 2002. Krings, hans p. (2001): Repairing texts: Empirical investigations of machine translation post-editing processes (geoffrey s. koby, ed.), the kent state university press, kent, ohio london, 558 p. *Meta*, 47(3):435–436.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M Rush. 2018. Opennmt: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Revisiting meta-evaluation for grammatical error correction. *Transactions of the Association for Computational Linguistics*, 12:837–855.

Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the seventh workshop on statistical machine translation*, pages 181–190.

Maarit Koponen. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, (25):131–148.

Maarit Koponen, Wilker Aziz, Luciana Ramos, Lucia Specia, Jussi Rautio, Lauri Carlson, Inari Listenmaa, Seppo Nyrkkö, Gorka Labaka, Arantza Díaz De Ilarraza, and 1 others. 2012. Post-editing time as a measure of cognitive effort. In *AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP)*.

Oleksandr Korniienko. 2024. Enhancing grammatical correctness: The efficacy of large language models in error correction task.

Anoop Kunchukuttan, Sriram Chaudhury, and Pushpak Bhattacharyya. 2014. Tuning a grammar correction system for increased precision. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 60–64.

Vivian Lai, Alison Smith-Renner, Ke Zhang, Ruijia Cheng, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes-Larrarte. 2022. An exploration of post-editing effectiveness in text summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 475–493, Seattle, United States. Association for Computational Linguistics.

Kyusong Lee and Gary Geunbae Lee. 2014. Postech grammatical error correction system in the conll-2014 shared task. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 65–73.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024. Dissecting human and llm preferences. *arXiv preprint arXiv:2402.11296*.

Wei Li, Wen Luo, Guangyue Peng, and Houfeng Wang. 2025. Explanation based in-context demonstrations retrieval for multilingual grammatical error correction. *arXiv preprint arXiv:2502.08507*.

Yinghao Li, Xuebo Liu, Shuo Wang, Peiyuan Gong, Derek F Wong, Yang Gao, He-Yan Huang, and Min Zhang. 2023. Templategec: Improving grammatical error correction with detection template. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6878–6892.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. 2022. IMPARA: Impact-based metric for GEC using parallel data. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3578–3588, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Maria Nadejde and Joel Tetreault. 2020. Personalizing grammatical error correction: Adaptation to proficiency level and l1. *arXiv preprint arXiv:2006.02964*.

Courtney Napoles, Maria Nădejde, and Joel Tetreault. 2019. Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. *Transactions of the Association for Computational Linguistics*, 7:551–566.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction. *arXiv preprint arXiv:1702.04066*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector–grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 17–33, Mexico City, Mexico. Association for Computational Linguistics.

Robert Östling, Katarina Gillholm, Murathan Kurfalı, Marie Mattson, and Mats Wirén. 2024. Evaluation of really good grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6582–6593, Torino, Italia. ELRA and ICCL.

Sharon O'Brien. 2005. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine translation*, 19:37–58.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020b. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bull. Math. Linguistics*, 93:7–16.

Maja Popović, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Proceedings of the 17th annual conference of the european association for machine translation*, pages 191–198.

Muhammad Reza Qorib, Seung-Hoon Na, and Hwee Tou Ng. 2022. Frustratingly easy system combination for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1964–1974.

Muhammad Reza Qorib and Hwee Tou Ng. 2023. System combination via quality estimation for grammatical error correction. *arXiv preprint arXiv:2310.14947*.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. Coedit: Text editing by task-specific instruction tuning. *arXiv preprint arXiv:2305.09857*.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707.

Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, Dan Roth, and Nizar Habash. 2014. The illinois-columbia system in the conll-2014 shared task. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 34–42.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36, Los Angeles, California. Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2021. How good (really) are grammatical error correction systems? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2686–2698.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. *arXiv preprint arXiv:1806.01170*.

Holger Schielzeth. 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1(2):103–113.

Dorothy Senez. 1998. Post-editing service for machine translation users at the european commission. In *Proceedings of Translating and the Computer 20*.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2023. Gee! grammar error explanation with large language models. *arXiv preprint arXiv:2311.09517*.

Lucia Specia. 2011. Exploiting objective annotations for minimising translation post-editing effort. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium. European Association for Machine Translation.

Lucia Specia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz. 2011. Predicting machine translation adequacy. In *Proceedings of Machine Translation Summit XIII: Papers*.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*.

Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. 2005. Evaluation of an NLG system using post-edit data: Lessons learnt. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland. Association for Computational Linguistics.

Yujin Takahashi, Masahiro Kaneko, Masato Mita, and Mamoru Komachi. 2022. ProQE: Proficiency-wise quality estimation dataset for grammatical error correction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5994–6000, Marseille, France. European Language Resources Association.

Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. 2022. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3842–3852.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Proceedings of Machine Translation Summit XII: Posters*.

Midori Tatsumi and Johann Roturier. 2010. Source text characteristics and technical and temporal post-editing effort: what is their relationship. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 43–52.

Irina P Temnikova. 2010. Cognitive evaluation approach for a controlled language post–editing experiment. In *LREC*.

Joel Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Coling 2008: Proceedings of the workshop on human judgements in computational linguistics*, pages 24–32.

Joel Tetreault, Martin Chodorow, and Nitin Madnani. 2014. Bucking the trend: improved evaluation and annotation practices for esl error detection systems. *Language Resources and Evaluation*, 48:5–31.

Arda Tezcan, Veronique Hoste, and Lieve Macken. 2016. Detecting grammatical errors in machine translation output using dependency parsing and treebank querying. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 203–217.

Arda Tezcan, Véronique Hoste, and Lieve Macken. 2019. Estimating post-editing time using a gold-standard set of machine translation errors. *Computer Speech & Language*, 55:120–144.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Lucas Nunes Vieira. 2014. Indices of cognitive effort in machine translation post-editing. *Machine translation*, 28(3):187–216.

Lucas Nunes Vieira. 2017. Cognitive effort and different task foci in post-editing of machine translation: A think-aloud study. *Across Languages and Cultures*, 18(1):79–105.

Peilu Wang, Zhongye Jia, and Hai Zhao. 2014a. Grammatical error detection and correction using a single maximum entropy model. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 74–82.

Yiming Wang, Longyue Wang, Xiaodong Zeng, Derek F Wong, Lidia S Chao, and Yi Lu. 2014b. Factored statistical machine translation for grammatical error correction. In *Proceedings of the eighteenth conference on computational natural language learning: Shared task*, pages 83–90.

Jian-Cheng Wu, Tzu-Hsi Yen, Jim Chang, Guan-Cheng Huang, Hsiang-Ling Hsu, Yu-wei Chang, and Jason S Chang. 2014. Nthu at the conll-2014 shared task. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 91–95.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. Lm-critic: Language models for unsupervised grammatical error correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7763.

Jingheng Ye, Shang Qin, Yinghui Li, Hai-Tao Zheng, Shen Wang, and Qingsong Wen. 2025. Corrections meet explanations: A unified framework for explainable grammatical error correction. *arXiv preprint arXiv:2502.15261*.

Na Ye, Ling Jiang, Dandan Ma, Yingxin Zhang, Sanyuan Zhao, and Dongfeng Cai. 2021. Predicting post-editing effort for english-chinese neural machine translation. In *2021 International Conference on Asian Language Processing (IALP)*, pages 154–158. IEEE.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. Some: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522.

Zheng Yuan, Shiva Taslimipoor, Christopher Davis, and Christopher Bryant. 2021. Multi-class grammatical error detection for correction: A tale of two systems. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 8722–8736.

Anna Zaretskaya, Mihaela Vela, Gloria Corpas Pastor, and Miriam Seghiri. 2016. Measuring post-editing time and effort for different types of machine translation errors.

Longkai Zhang and Houfeng Wang. 2014. A unified framework for grammar error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 96–102.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *arXiv preprint arXiv:2305.13225*.

A Neural Regression Models for PEET Estimation

Since semantics and syntax structure have been shown to impact PE effort (Tezcan et al., 2016; Bangalore et al., 2015), we also trained neural-LM PEET Scorer models using flattened constituency parse trees (Kitaev and Klein, 2018) and part-of-speech syntax structure features for the source and target corrections, generated using the spaCy library (Honnibal and Montani, 2017).

Model Features	**BERT-L**		**RoBERTa-L**	
	r	MAE	r	MAE
Sentence Edit	0.552	17.73	**0.56**	17.97
Syntactic Variation	0.528	19.35	**0.564**	18.05
#EW + Syntactic Variation	**0.564**	17.16	0.561	16.88
#EW + Syntax Structure	**0.565**	18.57	0.565	18.74

Table 10: Performance of Neural PEET models using different sequence model features over 5 runs. The results are shown as Pearson Correlation (r) and Mean Absolute Error (MAE) loss.

Pretrained LMs can also capture syntax structure internally (Dai et al., 2021), so we also train neural-LM models using only source-target sentence embeddings as features to estimate PEET. Since the statistical models work as well as Neural models, while being faster and more interpretable, we consider them for the PEET Scorer in the main paper. We describe the features (Table 9) and results of the Neural PEET (Table 10) model here.

B GEC Evaluation File Example and Format

The evaluation of a GEC Tool requires a Source (S), Target (T) and Model Output (MO) sentence. Table 11 gives an example of such a triple. GEC evaluation generates M2 file for a pair of sentences (e.g., S and T), which lists the edits that can convert sentence S into sentence T and the positions of those edits. The evaluation process generates two M2 files : (Source - Target) and (Source - Model Output). The M2 edits are compared to evaluate the Model Output quality.

- *Source-Target* M2 File:

```
S Surrounded by such concerns , it is very likely that we are distracted to worry about these problems .
A 13 14|||R:OTHER|||and|||REQUIRED||| -NONE-|||0
A 11 12|||R:VERB:TENSE|||will be|||REQUIRED||| -NONE-|||1
A 12 12|||M:ADV|||too|||REQUIRED||| -NONE-|||1
```

- *Source-Model Output* M2 File:

```
S Surrounded by such concerns , it is very likely that we are distracted to worry about these problems .
A 13 14|||R:PART|||from|||REQUIRED||| -NONE-|||0
A 14 15|||R:VERB:FORM|||worrying||| REQUIRED||| -NONE-|||0
```

The M2 file format was part of the GEC-M2 Scorer evaluation tool proposed by Dahlmeier and Ng (2012). The tool generates an alignment and

Model Type	Input Format
Sentence Edit	[MO] <mo-sentence> [TRG] <trg- sentence>
Syntactic Variation	<mo-constituency-parse> [TO] <trg-constituency-parse>
#EW + Syntactic Variation	#EW - <mo-constituency-parse> [TO] <trg-constituency-parse>
#EW + Syntax Structure	#EW - <trg-part-of-speech-tag>

Table 9: The training data format for the BERT and RoBERTa LM. The example considers a sentence pair - <mo-sentence> and <trg- sentence> where "mo" is the Model Output correction made by a GEC Tool and the "trg" is the post-edited target correction for "mo". The special tokens [MO], [TRG] and [TO] denote sentence breaks in the input. #EW denotes the number of edited words between mo and trg.

Source :	Surrounded by such concerns, it is very likely that we are distracted to worry about these problems.
Target :	Surrounded by such concerns, it is very likely that we will be too distracted to worry about these problems.
Model Output :	Surrounded by such concerns, it is very likely that we are distracted from worrying about these problems.

Table 11: *Source*, *Target* and example *Model Output* made by a GEC Tool.

detects atomic edits between a pair of sentences. Further improvement to the M2 tool was done by Bryant et al. (2017), resulting in the ERRANT evaluation tool. The ERRANT tool retained the overall M2 file format, utilizing syntactic and linguistic features to extract better-aligned and tagged edits between 2 sentences (as shown above).

C Predictive Model Parameters

We train different statistical and neural predictive models to estimate the post-editing temporal effort. We use this section to describe the predictive models as well as the training parameters for the regression task.

Linear Regression: We use the Linear Regression (LR) model provided by the Scikit-Learn library[5]. To keep the weights of the features from getting arbitrarily high, we used the RidgeLinear model that also adds an L2 Regularizer to the model. We trained the model with default training parameters and $alpha = 1.0$.

Support Vector Regression: We also train Support Vector Regression (SVR) models from scikit-learn with the default training parameters and the "linear" kernel.

BERT, RoBERTa Neural Models: To train neural predictive models, we fine-tuned the BERT-Large (Devlin et al., 2019) and RoBERTa-Large (Liu et al., 2019) with a regression head. The models were trained using the Pfeiffer bottleneck adapters (Pfeiffer et al., 2020a) which allowed us to reduce the training time. We utilized the AdapterHub library[6] for training the models with the default Pfeiffer adapter configuration (Pfeiffer et al., 2020b). Training was done for 50 epochs with a 10-epoch and .05 loss threshold early stopping. A learning rate of $1e-04$ was used. To train the models for the regression task, we added a one-label regression head and used the mean-square-error loss (MSELoss), which is part of the Huggingface[7] training pipeline.

D Different Sources for Training Feature Selection and Extraction

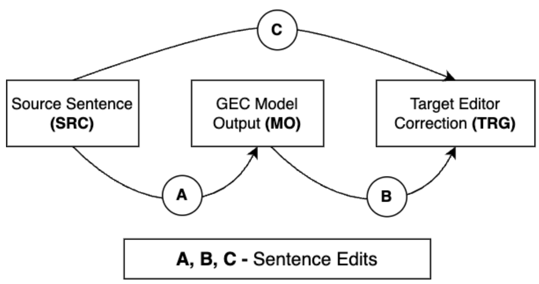


Figure 2: Sentence correction edits extracted using the ERRANT toolkit.

Our dataset has 3 iterations for each source sentence. We have the original sentence - source

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

[6] https://adapterhub.ml/

[7] https://huggingface.co/

(SRC), the first-pass correction by a GEC Tool - Model Output (MO) and the final targeted editor correction - target (TRG). Figure 2 shows the 3 iterations for the source sentence. Each arc represents a sentence transition pairing and can be used to extract intermediate edit features. To extract features, the following sentence pairings can be considered: [MO], [SRC - MO], **[MO - TRG]**, **[SRC - MO - TRG]**. Post-editing features, from different levels, can be extracted from the $SRC - MO - TRG$ and $MO - TRG$ sentence pairings. Considering the source sentence as a feature can further separate target edits into ignored and incorrect edits.

- **SRC - MO - TRG**: We consider and extract the set of edits - A and C (Figure 2) for the model features. We further use these edits to create 2 categories - Incorrect and Ignored edits.
  - Incorrect: $|A - C|$
  - Ignored: $|C - A|$
- **MO - TRG**: We consider only edit set - B (Figure 2) as the input for the trained models.

We found that the performance of models trained on these 2 feature sources was comparable (Appendix E). This also indicates that the PEET Scorer can estimate time-to-correct from the post-editing correction stage - B. We only present and discuss the results of the model trained using the $MO - TRG$ sentence features in the main paper. Results for the $[SRC - MO - TRG]$ Scorer are presented in Appendix E.

E PEET Scorer using SRC, MO and TRG Sentence Features

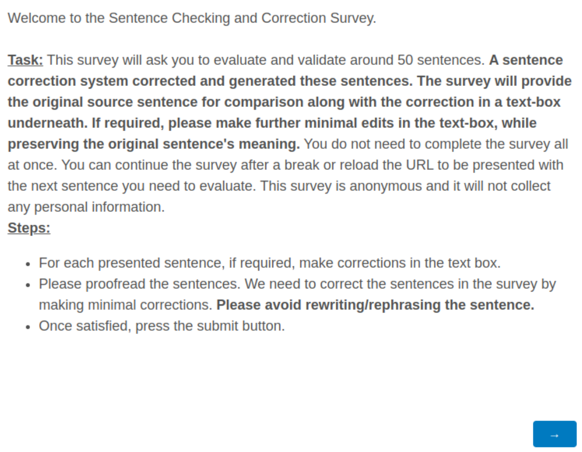
Model Features	BERT-L		RoBERTa-L	
	r	MAE	r	MAE
Sentence Edit	0.513	19.10	0.54	17.82

Table 12: Neural PEET model performance over 5 runs using the source (SRC), GEC Tool Model Output (MO) and Target Correction (TRG) sentence features. The results are shown as Pearson Correlation (r) and Mean Absolute Error (MAE) loss.

Statistical Model	Edit Feature Level	r	MAE
Linear Regression	10	0.558	18.92
	106	0.557	18.89
SVR Linear	10	0.556	16.39
	106	0.561	16.21

Table 13: PEET Statistical Model performance over 50 runs (different train-test data seed) using Incorrect and Ignored separated Edit features (Appendix D) extracted from SRC, MO and TRG sentence triples. The results are presented as the Pearson Correlation (r), Mean Absolute Error (MAE) loss.

F GEC Post Editing Instructions and Survey Example


Welcome to the Sentence Checking and Correction Survey.

Task: This survey will ask you to evaluate and validate around 50 sentences. **A sentence correction system corrected and generated these sentences. The survey will provide the original source sentence for comparison along with the correction in a text-box underneath. If required, please make further minimal edits in the text-box, while preserving the original sentence's meaning.** You do not need to complete the survey all at once. You can continue the survey after a break or reload the URL to be presented with the next sentence you need to evaluate. This survey is anonymous and it will not collect any personal information.

Steps:

- For each presented sentence, if required, make corrections in the text box.
- Please proofread the sentences. We need to correct the sentences in the survey by making minimal corrections. **Please avoid rewriting/rephrasing the sentence.**
- Once satisfied, press the submit button.

→

Figure 3: Survey instructions for the editor to perform post editing, and obtain target corrections for our dataset.

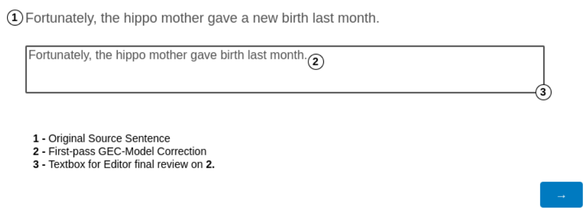


Figure 4: Example source sentence and its first-pass edit from the Survey. The editor can make further improvements in the text box. Submitting the final target correction.

G Feature Impact on Post-Editing Time using Regression Coefficients

We utilize the regression coefficients of a Ridge-Linear Regression model to quantitatively calculate the impact of different edit type features on the

Model Features	Regression Coefficient	Model Features	Regression Coefficient	Model Features	Regression Coefficient	Model Features	Regression Coefficient	Model Features	Regression Coefficient
R:OTHER	7.73	**M:DET**	2.03	**M:VERB**	1.49	**U:VERB**	1.07	**M:ADJ**	0.36
U:OTHER	4.53	**M:OTHER**	1.98	**R:VERB:FORM**	1.48	**M:ADV**	0.93	**R:CONJ**	0.30
Sentence Correct	-3.11	**R:DET**	1.94	**U:PUNCT**	1.36	**U:ADJ**	0.79	**U:NOUN:POSS**	0.29
R:PREP	2.85	**M:PREP**	1.93	**U:ADV**	1.32	**R:VERB:INFL**	0.58	**U:VERB:TENSE**	0.25
R:PUNCT	2.84	**R:MORPH**	1.77	**M:VERB:TENSE**	1.32	**R:ADV**	0.53	**M:PART**	0.18
M:PUNCT	2.80	**U:PREP**	1.69	**M:VERB:FORM**	1.29	**M:NOUN:POSS**	0.52	**U:PART**	0.10
R:VERB	2.71	**R:SPELL**	1.66	**U:NOUN**	1.26	**R:ADJ**	0.51	**R:NOUN:POSS**	-0.06
R:NOUN	2.64	**U:CONJ**	1.64	**M:NOUN**	1.22	**M:PRON**	0.49	**U:PRON**	0.06
R:NOUN:NUM	2.32	**U:DET**	1.62	**R:PRON**	1.14	**R:PART**	0.42	**U:VERB:FORM**	0.05
R:ORTH	2.22	**R:WO**	1.58	**R:VERB:SVA**	1.11	**R:ADJ:FORM**	0.41	**M:CONTR**	0.02
R:VERB:TENSE	2.08	**M:CONJ**	1.52	**U:CONTR**	1.10	**R:NOUN:INFL**	-0.37	**R:CONTR**	0.02

Table 14: The standardized regression coefficients of the LR model trained on all the big (55) edit features to measure the impact of each feature on PEET estimation.

time-to-correct value (Section 5.2). We provide the estimated impact of all edit types here.

Model Features	Regression Coefficient
Substitutions (R)	14.05
Deletions (U)	6.71
Insertions (M)	5.28
Sentence Correct (C)	-2.33

Table 15: The standardized regression coefficients of the LR model trained on the small (4) edit features to measure the impact of each feature on PEET estimation.

Model Features	PEET Correlation	Regression Coefficient
# of words in TRG	0.43	14.07
Substitutions (R)	0.47	6.76
# of Edited Words	0.52	6.46
# of Words in MO	0.43	-5.86
Deletions (U)	0.32	3.85
Sentence Correct (C)	-0.3	-2.63
Insertions (M)	0.28	0.66

Table 16: The correlation of the features used to train the small-edits(4) Linear Regression (LR) model in Table 6. We also list the standardized regression coefficients to measure the impact of each feature on PEET estimation.

H PEET Scorer Ranking and Comparison of GEC Tools with Human Judgment Rankings

We evaluate and rank 33 different GEC Tools and correction sets, part of 3 GEC Human Judgment Rankings, to estimate the quality of our PEET Scorer (Section 5.3). We list all the GEC Tools along with the Human Judgment and PEET Scorer rankings here.

Model Name	HJR Score	PEET Score	PEET Ranking
marian	76.99	21.82	1
lstm-r	74.48	22.45	3
lstm	74.3	22.39	2
nus	73.94	22.47	4
transformer	73.9	22.79	5
amu	70.68	23.27	6
input	68.15	23.3	7

Table 17: PEET Scorer estimated average time-to-correct per sentence and ranking for 7 GEC Tool corrections on the FCE dataset (1936 Sentences), along with their Human Judgment Ranking (HJR), presented in *Napoles-FCE* (Napoles et al., 2019) (Section 5.3). The 7 GEC Tools consist of Seq2Seq Neural Models.

Model Name	HJR Score	PEET Score	PEET Ranking
lstm-r	78.27	27.61	2
lstm	77.73	27.61	1
amu	75.98	28.35	5
input	75.89	27.72	3
marian	75.8	30.52	7
nus	75.78	28.34	4
transformer	71.53	29.77	6

Table 18: PEET Scorer estimated average time-to-correct per sentence and ranking for 7 GEC Tool corrections on the WikiEd dataset (1984 Sentences), along with their Human Judgment Ranking (HJR), presented in *Napoles-Wiki* (Napoles et al., 2019) (Section 5.3). The 7 GEC Tools consist of Seq2Seq Neural Models.

Table 17-18 list the estimation scores for the 6 Seq2Seq GEC Tools ranked by Napoles et al. (2019). The chosen models were AMU (Junczys-Dowmunt and Grundkiewicz, 2016), LSTM/LSTM-R (Klein et al., 2018), Marian (Sennrich et al., 2017), NUS (Chollampatt and Ng, 2018), and, Transformer (Vaswani et al., 2017).

Model Name	HJR Score	PEET Score	PEET Ranking
AMU	0.628	25.8	8
RAC	0.566	26.61	13
CAMB	0.561	26.34	11
CUUI	0.55	25.91	9
POST	0.539	26.28	10
UFC	0.513	24.56	2
PKU	0.506	25.63	6
UMC	0.495	25.72	7
IITB	0.485	24.67	3
SJTU	0.463	24.84	4
INPUT	0.456	24.53	1
NTHU	0.437	26.6	12
IPN	0.3	25.62	5

Table 19: PEET Scorer estimated average time-to-correct per sentence and ranking for 12 GEC Tool corrections on the CONLL14 dataset (1312 Sentences), along with their Human Judgment Ranking (HJR), presented in *Grundkiewicz-C14(EW)* (Grundkiewicz et al., 2015) (Section 5.3). The 12 GEC Tools consist primarily of rule-based and statistical machine translation architecture.

Table 19 lists the quality judgment for the 12 GEC Tools that participated in the CONLL14 GEC Task (Ng et al., 2014) performed by Grundkiewicz et al. (2015). AMU (Junczys-Dowmunt and Grundkiewicz, 2014), CAMB (Felice et al., 2014), CUUI (Rozovskaya et al., 2014), IITB (Kunchukuttan et al., 2014), IPN (Hernandez and Calvo, 2014), NARA (Ng et al., 2014), NTHU (Wu et al., 2014), PKU (Zhang and Wang, 2014), POST (Lee and Lee, 2014), RAC (Boroş et al., 2014), SJTU (Wang et al., 2014a), UFC (Gupta, 2014), and UMC (Wang et al., 2014b).

Table 20 lists the recent GEC Tools evaluated by Kobayashi et al. (2024). GPT-3.5 (Coyne et al., 2023), T5 (Rothe et al., 2021), TransGEC (Fang et al., 2023a), BERT-Fuse (Kaneko et al., 2020), Riken-Tohoku (Kiyono et al., 2019), PIE (Awasthi et al., 2019), LM-Critic (Yasunaga et al., 2021), TemplateGEC (Li et al., 2023), GECToR-BERT (Omelianchuk et al., 2020), UEDIN-MS (Grundkiewicz et al., 2019), GECToR-Ens (Tarnavskyi et al., 2022), BART (Lewis et al., 2020).

Model Name	HJR Score	PEET Score	PEET Ranking
REF-F	0.992	30.53	15
GPT-3.5	0.743	26.04	14
T5	0.179	24.37	10
TransGEC	0.175	23.54	3
REF-M	0.067	24.04	8
BERT-Fuse	0.023	23.61	4
Riken-Tohoku	-0.001	23.36	2
PIE	-0.034	23.66	6
LM-Critic	-0.163	24.37	9
Template GEC	-0.168	25.21	13
GECToR-BERT	-0.178	23.78	7
UEDIN-MS	-0.179	23.36	1
GECToR-Ens	-0.234	23.62	5
BART	-0.3	24.75	12
INPUT	-0.992	24.53	11

Table 20: PEET Scorer estimated average time-to-correct per sentence and ranking for 15 GEC Tool corrections on the CONLL14 dataset (1312 Sentences), along with their Human Judgment Ranking (HJR), presented in *SEEDA-C14-All(TS)* (Kobayashi et al., 2024) (Section 5.3). The 15 GEC Tools consist of strong SOA Neural Models.

Hybrid Intelligence for Logical Fallacy Detection

Mariia Kutepova
University of Groningen
marrrie.23@gmail.com

Khalid Al-Khatib
University of Groningen
khalid.alkhatib@rug.nl

Abstract

This study investigates the impact of Hybrid Intelligence (HI) on improving the detection of logical fallacies, addressing the pressing challenge of misinformation prevalent across communication platforms. Employing a *between-subjects* experimental design, the research compares the performance of two groups: one relying exclusively on human judgment and another supported by an AI assistant. Participants evaluated a series of statements, with the AI-assisted group utilizing a custom ChatGPT-based chatbot that provided real-time hints and clarifications. The findings reveal a significant improvement in fallacy detection with AI support, increasing from an F1-score of 0.76 in the human-only group to 0.90 in the AI-assisted group. Despite this enhancement, both groups struggled to accurately identify non-fallacious statements, highlighting the need to further refine how AI assistance is leveraged.

1 Introduction

The increasing prevalence of misleading information has created an urgent need to improve our ability to detect deceptive content. Faulty reasoning (i.e., fallacies), which drives the spread of misinformation across various discourse domains, poses significant risks to informed decision-making and public discourse (Vrbová et al., 2021; Teneva, 2023). Some researchers even liken this problem to an epidemic (Duarte, 2024). Despite its importance, fallacy detection remains an understudied challenge, with current computational argumentation methods struggling to capture the complexity of deceptive arguments (Goffredo et al., 2022).

Van Eemeren and Verheij (2017) note that fallacies have received limited attention in both formal and computational argumentation research. Recent studies have attempted to bridge this gap by constructing fallacy datasets and developing automatic fallacy identification methods. However, none have explored a Hybrid Intelligence (HI) approach, to the best of our knowledge. HI, which integrates human and AI capabilities, is considered highly promising but requires further empirical research to evaluate its effectiveness across different tasks and domains (Dellermann et al., 2019).

The concept of HI is based on the view that while AI excels in data processing and pattern recognition, it lacks the creativity, empathy, and contextual understanding that humans bring to cognitive tasks. Dellermann et al. (2019) emphasize that HI systems are designed to leverage these complementary strengths, enabling humans and AI to learn from each other and improve over time. This continuous adaptation is particularly crucial for complex tasks such as fallacy detection.

Dellermann et al. (2019) also note that while machine learning and HI are advancing toward real-world applications, the next step is to enhance their problem-solving capabilities. Since HI is still a relatively new concept, further empirical research is necessary to assess its effectiveness across various domains and tasks. Furthermore, although theoretical frameworks for HI exist, practical guidelines for integrating human intuition and creativity with AI's computational power remain unclear. This gap highlights the need for studies that not only develop theoretical insights but also provide empirical validation of the advantages of human-AI collaboration.

This study aims to deepen our understanding of HI in fallacy detection by addressing several key questions: How can existing AI models be effectively integrated into a hybrid intelligent system to assist in this process? How do individuals identify fallacies with AI assistance compared to doing so without AI support? What challenges and limitations arise when employing HI systems for fallacy detection?

By addressing these questions, this research seeks to evaluate the effectiveness of HI in fal-

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 197–208
November 9, 2025

lacy detection and to provide practical insights into its real-world application. The findings have broader implications for decision-making in education, healthcare, and other fields. The results showed a substantial increase in performance with AI support, rising from an F1-score of 0.76 in the human-only study to 0.90 in the HI study. This highlights the potential of HI to enhance and support human cognitive abilities in complex tasks such as argumentation analysis. All resources developed in this paper are publicly available[1].

2 Related Work

This section reviews recent advancements in computational argumentation and fallacy detection, introduces Hybrid Intelligence (HI) and its potential to enhance decision-making and problem-solving, and discusses the application of HI in improving fallacy detection and argument analysis.

Computational Argumentation and Fallacy Detection Fallacy detection within computational argumentation has gained importance as AI becomes increasingly integrated into daily life and research. Combining AI, linguistics, and logic, computational argumentation analyzes, models, and assesses arguments in natural language; a crucial task in today's information age, where misinformation and faulty reasoning threaten public discourse and decision-making (Sourati et al., 2023). Recent advancements include machine learning models and annotated datasets. For instance, Jin et al. (2022) highlight the limitations of existing models in detecting complex fallacies, while Goffredo et al. (2023) have enhanced fallacy detection in political debates through improved datasets and neural network architectures. Despite these advancements, challenges such as explainability persist, as Sourati et al. (2023) emphasize the need for transparent AI systems to build user trust. Practical applications span education and healthcare, where argumentation systems enhance critical thinking and assist in diagnostic decision-making (Atkinson et al., 2017). Integrating computational argumentation techniques into fallacy detection offers significant progress in understanding and analyzing arguments across various fields.

Hybrid Intelligence Human-AI collaboration has been explored across diverse domains, demonstrating its potential to enhance performance and decision-making. In social chatbots, AI is perceived as a companion providing emotional support (Brandtzaeg et al., 2022), while in mental health, AI enhances empathy in peer-to-peer conversations (Sharma et al., 2023). In creative fields, AI serves as a co-creator, generating new ideas, and techniques such as zero- and few-shot learning show promise despite certain challenges (Dang et al., 2022). In education, AI fosters critical thinking and personalized learning (Markauskaite et al., 2022; Muthmainnah et al., 2022). In customer service, AI improves efficiency by handling routine tasks, allowing human agents to focus on complex issues (Vassilakopoulou et al., 2022). Moreover, Jiang et al. (2022) stress the importance of AI supporting human decision-making without overwhelming users, highlighting clear communication and intuitive design as key factors for successful human-AI collaboration.

HI and Fallacy Detection HI combines human cognitive strengths with AI's computational capabilities to enhance problem-solving and decision-making. Unlike Artificial General Intelligence (AGI), which aims to replicate human cognition, HI focuses on leveraging complementary skills, such as human creativity and empathy, alongside AI's ability to process large datasets. Researchers such as Dellermann et al. (2019) highlight the potential of HI to achieve superior outcomes through collaboration, where humans and AI enhance each other's performance. HI's co-evolutionary nature, where both human and AI agents learn from each other, has proven effective in fields such as digital humanities and education. To our knowledge, no previous work has directly targeted fallacy detection. However, related studies such as Guo et al. (2023) demonstrate how AI chatbots improve students' argumentation skills by providing immediate feedback. Our study builds on these concepts by developing a Hybrid Argumentation Assistant that leverages HI to enhance fallacy detection, combining human intuition with AI capabilities to improve cognitive tasks.

3 Methodology

The methodology of this study is primarily exploratory but incorporates experimental elements, including a between-subjects design inspired by Field and Hole (2023), with two treatment groups: one where participants use an AI assistant (HI) and

[1] https://github.com/marrrie23/Hybrid_Intelligence_Research

one without AI assistance. This design was chosen to effectively assess the impact of AI support on fallacy detection when combined with human reasoning. To evaluate participants' ability to identify logical fallacies, the study employed a range of materials and instruments, including surveys, a custom ChatGPT-based chatbot, and various data analysis techniques. Specifically, two user studies were conducted to collect detailed responses. In the human-only study, participants were asked to identify logical fallacies in given statements, with their performance and reasoning captured through open-ended responses. The HI study followed a similar structure but incorporated interactive elements, allowing participants to engage with a ChatGPT-based chatbot for fallacy detection. This setup enabled the evaluation of AI assistance in supporting reasoning and improving performance. Figure 1 illustrates the overall methodology of this paper.

Task The fallacy detection task involved presenting participants with a statement as input and asking them to determine whether it was logically sound or fallacious.

Data Preparation We utilized the "Logic" dataset introduced by Jin et al. (2022), which contains about 2,300 examples of logical fallacies sourced from educational materials. The dataset includes a diverse range of fallacies, each classified into one of 14 categories, collected from online quiz platforms and websites, with annotations provided by undergraduate students. Table 1 presents the distribution of fallacy categories in the dataset.

To address the uneven distribution of fallacy types, we balanced the dataset by randomly removing some entries to achieve a more even number of arguments per fallacy type. After removing irrelevant columns and filtering arguments based on length, we created a dataset with 1,000 arguments.

Also, we enriched the dataset with high-quality non-fallacious arguments from (Gleize et al., 2019) to ensure having both fallacious and non-fallacious arguments for evaluation. To mitigate potential bias arising from significant differences in argument length, we calculated the maximum (890), minimum (19), and average (131) argument lengths. We then filtered the arguments to include only those within the range of 100 to 160 words. This resulted in a smaller but more homogeneous set of fallacious arguments. The "Miscellaneous" category, which was initially present, was dropped since its arguments fell outside the 100–160 word range. Consequently, the number of fallacy types was reduced from 14 to 13.

Finally, we utilized the filtered arguments and derived 20 subsets, each comprising 10 examples, a mix of fallacious and non-fallacious arguments in a 7:3 ratio. The order of the statements was randomized to prevent bias. This structure ensured diversity and minimized bias, allowing for a more reliable assessment of fallacy detection performance. Table 2 shows the distribution of fallacies in the final dataset used for the user studies, while Table 3 provides an example of one of the sets distributed to participants.

Comparison Elements The primary comparison metric is participants' performance rates in identifying fallacies. Additionally, we examined how requesting AI-generated hints influenced participants' performance. Furthermore, participants were surveyed about their trust in AI and their perception of its transparency.

3.1 Study 1: Human-Only Fallacy Detection

This study investigates how well individuals can detect logical fallacies without AI assistance, serving as a baseline for comparison with AI-supported detection. Participants evaluated a set of arguments, determining the presence of fallacies and providing justifications for their responses. The study aimed to capture human reasoning patterns and establish a reference point for assessing the potential benefits of AI in fallacy detection.

Participants This study focused on fallacy detection using human reasoning alone and involved 60 participants recruited through snowball sampling. The target group consisted of higher education students (bachelor's and master's) and recent graduates in the Netherlands, selected for convenience and ease of recruitment via social media, university mailing lists, and student-oriented platforms such as SurveySwap and SurveyCircle.

Study Design and Implementation A structured study was conducted using a questionnaire administered via the Qualtrics platform [2] to evaluate the effectiveness of human-only fallacy detection. The study included 20 unique sets of arguments, each containing 10 distinct instances. Sixty participants were randomly assigned to review one set, ensuring that each set was evaluated by three participants. For each argument, they answered a binary

[2] www.qualtrics.com

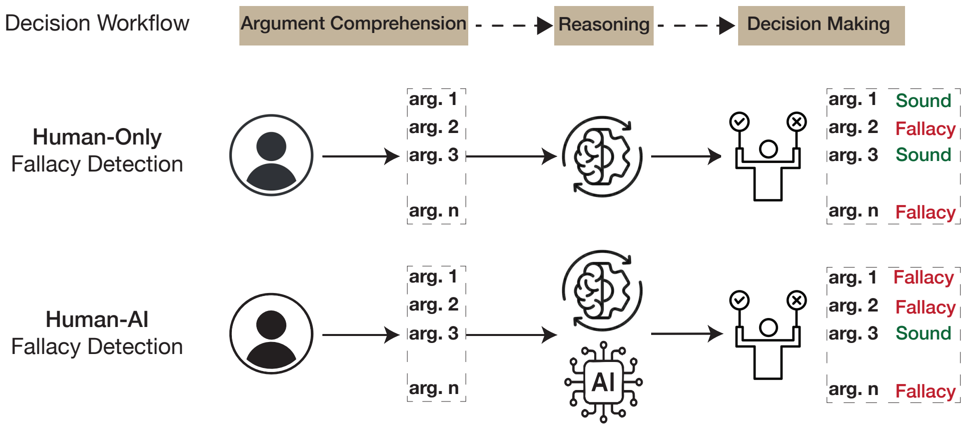


Figure 1: Decision-making workflow for human-only and human–AI fallacy detection.

Fallacy Type	**%**	**Fallacy Type**	**%**
Faulty Generalization	18.4	Intentional Fallacy	5.9
Ad Hominem	12.0	False Dilemma	5.7
Ad Populum	9.3	Fallacy of Credibility	5.7
False Causality	8.6	Fallacy of Extension	5.2
Circular Reasoning	7.1	Equivocation	2.0
Appeal to Emotion	7.2		
Fallacy of Relevance	6.5		
Fallacy of Logic	6.3		

Table 1: Distribution of logical fallacies in (Jin et al., 2022)

question: *"Does this statement contain a fallacy?"* by selecting either "Yes" or "No," followed by an open-ended justification with no constraints. Participants were allowed to complete the questionnaire at their own pace, with no imposed time limits. The data collection period spanned 14 days, providing ample time for participation and ensuring a robust dataset.

3.2 Study 2: Hybrid Intelligence Fallacy Detection

Building on the human-only fallacy detection study, this study explores the potential of AI-assisted reasoning in identifying logical fallacies. By integrating a ChatGPT-based assistant, participants received real-time hints and clarifications to support their decision-making process. This experiment aimed to assess the extent to which AI-generated guidance enhances fallacy detection performance compared to human reasoning alone.

Development of the Hybrid Argumentation Assistant The Hybrid Intelligence (HI) Assistant was developed as a ChatGPT-based chatbot designed to enhance human fallacy detection by providing real-time hints and clarifications. Built using OpenAI's GPT-3.5 model for accessibility, the chatbot interacted with participants by offering targeted hints without explicitly revealing answers, thereby guiding them in identifying logical fallacies.

The chatbot's user interface was designed for seamless interaction, featuring an intuitive chat system that provided feedback and hints. These hints were strategically integrated to encourage deeper reasoning and prompt elaboration on short or insufficient responses, improving both the user experience and the learning process. The chatbot's responses, along with the statements analyzed by participants, were carefully structured within a system prompt to maintain consistency and relevance[3].

Before deployment, the chatbot underwent rigorous testing and refinement to ensure reliability. Initial issues, such as incorrect argument sequencing and missing clarifications, were identified and addressed. The final version was optimized to effectively support participants in accurately detecting logical fallacies, highlighting AI's potential to augment human reasoning.

[3]The system prompt is provided in the appendix

Fallacy Type	Count	Fallacy Type	Count
Faulty Generalization	21	Intentional Fallacy	27
Ad Hominem	24	Fallacy of Credibility	36
Ad Populum	17	False Dilemma	18
False Causality	29	Fallacy of Extension	29
Circular Reasoning	18	Equivocation	23
Appeal to Emotion	27		
Fallacy of Relevance	22		
Fallacy of Logic	23		

Table 2: Fallacy type distribution in the final Dataset

Participants A total of 20 participants were recruited in the same way as in the first study. To ensure high-quality responses, participation was restricted to individuals with a strong performance record on the platform, verified by a high success rate in their previous tasks.

Study Design and Implementation Once the chatbot's behavior aligned with our guidelines, participants were instructed to copy and paste a predefined prompt before starting their interaction with GPT-3.5. After completing their session, they were required to share the link to their full conversation for evaluation.

4 Results

This section presents the findings from the two studies conducted to evaluate the effectiveness of human-only fallacy detection and the impact of AI assistance through a ChatGPT-based HI system. The first study established a baseline for human performance in identifying logical fallacies without AI support, while the second study explored how AI assistance could enhance performance and reasoning in fallacy detection. Both studies utilized the same set of arguments, allowing for a direct comparison of results. The analysis focuses on performance scores, response patterns, and the influence of AI-generated hints, providing insights into the strengths and limitations of human reasoning and the potential of HI in improving fallacy detection.

4.1 Study 1: Human-Only Fallacy Detection

In the first study, 60 participants evaluated logical fallacies without AI assistance. Each was randomly assigned one of 20 argument sets, assessing 10 statements by identifying fallacies (yes/no) and providing explanations. The study analysis followed several key steps to evaluate participants' fallacy detection performance. First, performance was determined based on the effectiveness of fallacy identifications across different argument sets. Responses were compiled into a single dataset for thorough analysis, with performance calculated both per set and per fallacy type. Also, the relationship between short-answer correctness and the length of participants' explanations was explored.

Performance by Set and Fallacy Type The human-only study shows that precision remains consistently high across fallacy types, with most classes reaching perfect precision (1.00). Recall, however, varies more noticeably: high for categories such as *equivocation* (0.92), *appeal to emotion* (0.86), and *fallacy of logic* (0.86), but lower for *false causality* (0.63), *intentional* (0.63), and *false dilemma* (0.64), leading to F1-scores that range from 0.77 to 0.96. Across the 20 evaluation sets, performance is stable, with F1-scores typically falling between 0.70 and 0.84, and only a few sets dipping to 0.62 or rising to 0.93. Overall, these results suggest that humans are highly precise in recognizing fallacies, but their sensitivity varies across types and test sets, highlighting the relative difficulty of consistently detecting certain categories. Table 5 shows fallacy detection performance across the 20 sets, and Table 4 shows the performance across the fallacy types.

Response Length and Performance Correlation The study further examined the relationship between the length of open-ended responses and the performance of fallacy identification. A weak positive correlation (0.325) was observed, suggesting that longer responses were modestly associated with higher performance. While the correlation is weak, it indicates that participants who provided extended responses may have engaged more deeply

Set	Fallacy Type	Argument
S6_A1	no fallacy	"Mandatory vaccinations would limit the spread of the flu and protect vulnerable populations that are at risk of death from the flu."
S6_A2	false dilemma	"Don't waste your money on a home security system; master thieves will still be able to get into your house."
S6_A3	faulty generalization	Donald Trump Jr. Tweeted: If I had a bowl of Skittles and I told you just three would kill you. Would you take a handful? That's our Syrian refugee problem.
S6_A4	no fallacy	Online shopping allows someone in remote areas to access goods normally only available in large cities.
S6_A5	ad populum	I guess I should buy my 12-year-old daughter an iPhone. Everyone at her school has one, and I want her to fit in with the other kids.
S6_A6	equivocation	But professor, I got all these facts from a program I saw on TV once... I don't remember the name of it though.
S6_A7	no fallacy	Addiction to gambling can lead to bankruptcy, families to split, or criminal behavior at times of desperation.
S6_A8	false causality	Mom: Watching TV that close will make you go blind, so move back! Jonny: That is B.S., Mom. Sorry, I am not moving.
S6_A9	circular reasoning	The Senator isn't lying when she says she cares about her constituents—she wouldn't lie to people she cares about.
S6_A10	fallacy of credibility	We should offer movies on our company's website. REPLY: No, we've built our company's fortune by renting movies only through our stores.

Table 3: An example of a set combining high-quality arguments and fallacies.

with the material, leading to higher performance.

Several factors may explain this relationship. First, longer responses might reflect deeper engagement with the arguments, allowing participants to analyze and process the fallacies more thoroughly. Second, participants who wrote more may have had a stronger understanding of the content and the ability to articulate their reasoning more effectively. Lastly, longer responses could indicate greater confidence and familiarity with the topic, enabling participants to provide more comprehensive justifications. In addition, the fallacies "appeal to emotion" and "fallacy of logic", which had the longest average response lengths (391.3 and 373.6 words, respectively), were also associated with higher performance (0.92 and 0.93). This finding supports the idea that more extensive responses may correlate with a deeper understanding or familiarity with the fallacy.

Insights and Implications The variability in performance across sets and fallacy types highlights the complexity of detecting fallacies and suggests gaps in participants' understanding. These findings provide a useful benchmark for evaluating the impact of AI assistance in the next study phase, where participants interact with the HI system. The results also suggest areas where educational interventions may be needed to improve human-only fallacy detection.

4.2 Study 2: ChatGPT-based Hybrid Intelligence Fallacy Detection

In this study, 20 participants used a ChatGPT-based chatbot to assist in fallacy detection. Each analyzed one of the 20 argument sets from the human-only study, identifying fallacies and explaining their reasoning with real-time hints provided by the chatbot. Following data collection, responses were aggregated and analyzed similarly to Study 1. Performance across sets and fallacy types was calculated, and additional variables, such as hint usage, were included to assess the chatbot's impact on fallacy detection. Participants could use hints for both short and open responses, and their usage was tracked for further analysis.

Performance by Set and Fallacy Type The Human–AI study demonstrates consistently strong performance across fallacy categories, with preci-

Fallacy Type	n	Prec.	Rec.	F1
ad hominem	45	1.00	0.82	0.90
ad populum	30	1.00	0.67	0.80
appeal to emotion	21	1.00	0.86	0.92
circular reasoning	33	1.00	0.70	0.82
equivocation	24	1.00	0.92	0.96
fallacy of credibility	36	1.00	0.67	0.80
fallacy of extension	30	1.00	0.67	0.80
fallacy of logic	36	1.00	0.86	0.93
fallacy of relevance	33	1.00	0.67	0.80
false causality	48	1.00	0.63	0.77
false dilemma	33	1.00	0.64	0.78
faulty generalization	24	1.00	0.75	0.86
intentional	27	1.00	0.63	0.77

Table 4: Per-fallacy performance from the Human-only study: number of instances (n), precision, recall, and F1-score.

Set	n	Prec.	Rec.	F1
1	30	0.82	0.67	0.74
2	30	0.78	0.86	0.82
3	30	0.82	0.86	0.84
4	30	0.85	0.81	0.83
5	30	0.88	0.71	0.79
6	30	0.81	0.62	0.70
7	30	0.85	0.81	0.83
8	30	0.68	0.62	0.65
9	30	0.82	0.86	0.84
10	30	0.73	0.76	0.74
11	30	0.78	0.67	0.72
12	30	0.68	0.62	0.65
13	30	0.83	0.71	0.77
14	30	0.79	0.71	0.75
15	30	0.70	0.76	0.73
16	30	0.67	0.57	0.62
17	30	0.78	0.67	0.72
18	30	0.83	0.71	0.77
19	30	0.93	0.67	0.78
20	30	0.76	0.76	0.76

Table 5: Performance across 20 evaluation sets from the Human-only study: precision, recall, and F1-score.

sion reaching 1.00 for every type and recall remaining high overall. Perfect F1-scores (1.00) were achieved for several categories such as *appeal to emotion*, *circular reasoning*, *fallacy of extension*, *fallacy of logic*, *faulty generalization*, and *intentional*, while slightly lower values appeared for more challenging classes like *fallacy of credibility* (0.91) and *ad hominem* (0.93). Results across the 20 evaluation sets confirm this trend: most sets achieved F1-scores above 0.90, with a few dips into the 0.70–0.80 range and a single low of 0.62, indicating some variability in performance across sets. Taken together, these findings suggest that Human–AI collaboration yields near-perfect precision and generally reliable recall, producing robust F1-scores across most fallacy types and evaluation sets. Table 7 shows the performance of detection fallacy across the 20 sets, and Table 6 shows the performance across the fallacy types.

Hint Usage Analysis of hint usage revealed that participants relied most on hints for identifying non-fallacious statements (16 hints), followed by fallacies such as "false causality" and "fallacy of extension." In contrast, fallacies like "ad hominem" and "equivocation" required fewer hints, suggesting that these types were easier for participants to identify with minimal AI support. Overall, the HI approach demonstrated a significant improvement in fallacy detection performance, highlighting the potential of AI assistance in enhancing human reasoning and decision-making.

5 Comparative Analysis of Human-Only and HI Studies

This section compares the results of the human-only fallacy detection study with the Hybrid Intelligence (HI) study, focusing on performance, additional variables such as hint usage, and the potential influence of the placebo effect. In addition, we report the results of a complementary experiment that directly evaluated large language models (LLMs) on the same fallacy detection tasks, providing a benchmark for comparison against both human-only and Human–AI studies. Together, these analyses highlight the strengths and limitations of each approach and provide insights into the role of AI in enhancing human reasoning and decision-making.

Performance The comparison between the Human-only and Human–AI reveals a clear improvement when AI support is introduced. While human alone achieve perfect precision but more variable recall across fallacy types (average F1 $\approx$ 0.84), collaboration with AI substantially boosts recall (average F1 $\approx$ 0.96), leading to more consistent performance across categories. A similar trend appears in the 20 evaluation sets: the Human-only study yields moderate stability with mean F1

Fallacy Type	n	Prec.	Rec.	F1
Ad Hominem	15	1.00	0.87	0.93
Ad Populum	10	1.00	0.90	0.95
Appeal to Emotion	7	1.00	1.00	1.00
Circular Reasoning	11	1.00	1.00	1.00
Equivocation	8	1.00	0.88	0.93
Fallacy of Credibility	12	1.00	0.83	0.91
Fallacy of Extension	10	1.00	1.00	1.00
Fallacy of Logic	12	1.00	1.00	1.00
Fallacy of Relevance	11	1.00	0.91	0.95
False Causality	16	1.00	0.94	0.97
False Dilemma	11	1.00	0.91	0.95
Faulty Generalization	8	1.00	1.00	1.00
Intentional	9	1.00	1.00	1.00

Table 6: Per-fallacy performance from the Human-AI study: number of instances (n), precision, recall, and F1-score.

Set	n	Prec.	Rec.	F1
1	10	1.00	1.00	1.00
2	10	0.88	1.00	0.93
3	10	0.78	1.00	0.88
4	10	0.88	1.00	0.93
5	10	0.88	1.00	0.93
6	10	1.00	1.00	1.00
7	10	0.88	1.00	0.93
8	10	0.83	0.71	0.77
9	10	0.70	1.00	0.82
10	10	0.70	1.00	0.82
11	10	0.67	0.57	0.62
12	10	0.88	1.00	0.93
13	10	0.88	1.00	0.93
14	10	1.00	0.86	0.92
15	10	0.88	1.00	0.93
16	10	0.86	0.86	0.86
17	10	1.00	1.00	1.00
18	10	1.00	1.00	1.00
19	10	0.71	0.71	0.71
20	10	0.88	1.00	0.93

Table 7: Performance across 20 evaluation sets from the Human-AI study: precision, recall, and F1-score.

around 0.76, whereas the Human–AI consistently reaches higher values (mean F1 $\approx$ 0.90). These results suggest that Human–AI collaboration enhances sensitivity and reliability in fallacy detection, while maintaining the already high precision observed in human judgments. Table 9 presents the average performance per set and per fallacy type for both studies (Human-only and Human–AI).

LLM-based Fallacy Detection Results. We evaluated a range of LLMs to measure their ability to distinguish between fallacious and non-fallacious arguments in our evaluation sets, using simple zero-shot prompting[4] (the full prompt is provided in Appendix). Table 8 summarizes the results in terms of precision, recall, and F1-scores. Overall, performance varies substantially across models. *GPT-4o* achieved the best overall balance, with the highest recall (0.93) and F1-score (0.92), while *GPT-3.5* yielded the highest precision (0.93) but at the cost of lower recall (0.73). *GPT-4o-mini* also performed strongly across all metrics (F1 = 0.91). Among open-source models, *DeepSeek-v3.1* reached the strongest balance (F1 = 0.84), followed by *LLaMA-3.1* and *Qwen3-235B* (F1 = 0.79–0.80). By contrast, *Claude Opus-4* showed moderate precision but notably weaker recall, resulting in the lowest F1 among the stronger contenders (0.60). The *GPT-OSS* baselines underperformed markedly, with F1-scores below 0.25. Taken together, these results suggest that frontier models such as GPT-4o, GPT-4o-mini, and GPT-3.5 provide highly reliable fallacy detection, with some open-source systems also showing competitive performance.

Compared to human performance (Table 9), the strongest LLMs reached F1-scores on par with the Human-only performance (0.76–0.84 across dimensions) but still fell short of Human–AI collaboration, which achieved up to 0.96 across fallacy types and 0.90 across evaluation sets. This gap highlights that while LLMs can approximate human judgment, they do not yet match the substantial gains observed when humans and AI work together.

Additional Variables The human-only study showed a weak positive correlation (0.325) between the length of open responses and performance, indicating that longer responses were associated with higher performance. In contrast, the HI study introduced hint usage as an additional variable, showing that non-fallacious statements required the most hints, suggesting participants found these the most difficult to assess. Overall, hints were used in 22.5% of cases, highlighting AI's role in assisting participants with more challenging fallacies.

Explanation of Findings The significant improvement in performance in the HI study may

[4] max_tokens is 128 and temperature is 0.0

LLM	Prec.	Rec.	F1
GPT-OSS 20B	0.15	0.50	0.23
GPT-OSS 120B	0.15	0.50	0.23
GPT-3.5	**0.93**	0.73	0.82
GPT-4o-mini	0.91	0.92	0.91
GPT-4o	0.92	**0.93**	**0.92**
LLaMA-3.1 70B	0.83	0.78	0.79
LLaMA-3.1 405B	0.84	0.78	0.80
Qwen3-235B	0.85	0.77	0.79
DeepSeek-v3.1	0.84	0.83	0.84
Claude Opus-4	0.84	0.60	0.60

Table 8: Fallacy detection performance of different LLMs, reported as precision, recall, and F1-scores. The best score in each column is shown in **bold**.

Eval.	Study	Prec.	Rec.	F1
Fallacy	Human-only	1.00	0.75	0.84
	Human–AI	1.00	0.93	0.96
Set	Human-only	0.80	0.73	0.76
	Human–AI	0.87	0.94	0.90

Table 9: Average precision, recall, and F1 across fallacy types and evaluation sets for Human-only vs. Human–AI.

partially be attributed to the placebo effect, where participants' belief in AI assistance positively influenced their performance (Kosch et al., 2022). The relatively low use of hints (22.5%) suggests that participants' confidence and engagement were enhanced merely by the presence of AI, even if they did not heavily rely on it for assistance. This effect likely contributed to the increased performance compared to the human-only study.

Another key finding was the difficulty in identifying non-fallacious statements, which achieved the lowest score in both studies. This aligns with (Yeh et al., 2024), who observe that expert disagreements predominantly concern whether any fallacy is present at all, with especially low agreement for the *None* (no-fallacy) class. A likely contributor is participants' personal beliefs and biases, particularly on controversial topics such as vaccination or fossil fuels, which can create judgments even when arguments are logically sound (Teneva, 2023). Together, these results highlight how demanding it is to certify the *absence* of a fallacy and the continuing need to scaffold critical thinking to decouple belief from reasoning in argument evaluation.

An additional consideration concerns the choice of LLM used in human–AI collaboration. In this study, we employed GPT-3.5 as the assisting model. While GPT-3.5 provided stable and effective support, recent evaluations (see Table 8) show that more advanced models such as GPT-4o achieve substantially higher precision and recall in fallacy detection. This suggests that the overall gains observed in our HI setting could be further amplified with stronger LLMs, highlighting the importance of carefully selecting and updating the underlying AI systems in future HI research and applications.

6 Conclusion

This study examined the effectiveness of Hybrid Intelligence (HI) in enhancing fallacy detection by comparing the performance of participants with and without AI assistance. The results demonstrated an improvement in performance when AI support was introduced, increasing from an F1-score of 0.76 in the human-only study to 0.90 in the HI study. These findings highlight the potential of HI to complement and augment human cognitive capabilities, particularly in complex domains such as argumentation analysis. However, the study also revealed persistent challenges in distinguishing logically sound arguments, which remained difficult for participants in both conditions. This suggests a need for further advancements in critical thinking and argumentation training to better differentiate between logically sound and fallacious reasoning.

The successful integration of a ChatGPT-based AI assistant into the HI framework showcased the feasibility of AI-supported fallacy detection. Nevertheless, certain limitations, such as occasional inconsistencies in AI behavior and variability in the quality of hints, indicate the necessity for further refinement in AI design and user interaction. Despite these challenges, the study provides valuable insights into the potential applications of HI systems across various fields, including education and decision-making, where enhanced cognitive support is essential. Future research should focus on expanding participant samples to include larger and more diverse populations, refining AI systems to improve reliability and usability, and exploring the long-term impact of HI on cognitive performance and reasoning tasks. By addressing these areas, the field can better harness the potential of HI to support and enhance human decision-making and problem-solving.

Acknowledgments

This work was developed while the first author was a Master's student at the University of Groningen. We thank the anonymous reviewers for their valuable feedback and constructive suggestions, which helped improve the quality of this paper.

References

Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo R Simari, Matthias Thimm, and Serena Villata. 2017. Toward artificial argumentation. *AI Magazine*, 38(3):25–36.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My ai friend: How users of a social chatbot understand their human–ai friendship. *Human Communication Research*, 48(3):404–429.

Hai Dang, Lukas Mecke, Florian Lehmann, Sarah Goller, and Daniel Buschek. 2022. How to prompt? opportunities and challenges of zero- and few-shot learning for human-ai interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*.

Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643.

André Duarte. 2024. Epidemiology of fallacies. *Argumentation*.

Andy Field and Graham Hole. 2023. *How to Design and Report Experiments*, 3 edition. SAGE Publications.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 163–172.

Pietro Goffredo, Mauricio Espinoza, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112.

Pietro Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 575–581.

Kai Guo, Yuyang Zhong, Dan Li, and Samuel Kai Wah Chu. 2023. Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation. *Computers & Education*, 203:104862.

Jinghui Jiang, Amanda J Karran, Constantinos K Coursaris, Pierre-Majorique Léger, and Jörg Beringer. 2022. A situation awareness perspective on human-ai interaction: Tensions and opportunities. *International Journal of Human-Computer Interaction*, 39(9):1789–1806.

Zhijing Jin, Ayush Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yi Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198.

Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2022. The placebo effect of artificial intelligence in human–computer interaction. *ACM Transactions on Computer-Human Interaction*, 29(6):1–32.

Lina Markauskaite, Rebecca Marrone, Oleksandra Poquet, Simon Knight, Roberto Martinez-Maldonado, Sarah Howard, Jo Tondeur, Maarten De Laat, Simon Buckingham Shum, Dragan Gašević, and George Siemens. 2022. Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with ai? *Computers and Education: Artificial Intelligence*, 3:100056.

N Muthmainnah, PMI Seraj, and Ibrahim Oteir. 2022. Playing with ai to investigate human-computer interaction technology and improving critical thinking skills to pursue 21st century age. *Education Research International*, 2022:1–17.

Ashish Sharma, I Wei Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Zohreh Sourati, VP Venkatesh, Dhruv Deshpande, Harsh Rawlani, Filip Ilievski, Helen Sandlin, and Alain Mermoud. 2023. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418.

Elena V Teneva. 2023. Digital pseudo-identification in the post-truth era: Exploring logical fallacies in the mainstream media coverage of the covid-19 vaccines. *Social Sciences*, 12(8):457.

Frans H Van Eemeren and Bart Verheij. 2017. Argumentation theory in formal and computational perspective. In *Handbook of Formal Argumentation*, volume 1, pages 3–71.

Polyxeni Vassilakopoulou, Arve Haug, Lars M Salvesen, and Ilias O Pappas. 2022. Developing human/ai interactions for chat-based customer services: lessons learned from the norwegian government. *European Journal of Information Systems*, 32(1):10–22.

Lucie Vrbová, Katerina Jiřinova, Karel Helman, and Hana Lorencova. 2021. Do informal reasoning fallacies really shape decisions? experimental evidence. *Rationality and Society*, 33(4):448–479.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. Cocolofa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677, Miami, Florida, USA. Association for Computational Linguistics.

7 Appendix

7.1 System Prompt in Study 2

This is part[5] of the System Prompt used in ‘Study 2: ChatGPT-based Hybrid Intelligence Fallacy Detection’:

/system

You are an interactive assistant used for conducting a user study about AI-human interaction for fallacy detection. Your task is to collaboratively assist participants in identifying fallacies in a series of arguments.

IMPORTANT: 1. You must NOT assess or evaluate whether the participant’s answers are correct or incorrect. 2. Your role is to facilitate discussion, record the participant’s answers, and provide subtle guidance (if requested) without indicating correctness. 3. Decisions about the presence of fallacies should be made collaboratively, with input from both the participant and the hints you provide, if requested. 4. Every question requires a response. Short answers (e.g., "yes/no") and open explanations are mandatory and cannot be skipped. 5. Maintain a neutral tone throughout the session.

Instructions: 1. **Introduction**: Start by greeting the participant and introducing the task. 2. **Confirm Instructions**: Ask the participant to confirm that they have read and understood the instructions. 3. **Consent to Participate**: Ask the participant to provide consent to participate in the study. They must type "yes" to confirm their consent before proceeding. 4. **Pre-Task Questions**: - Ask if the participant uses ChatGPT (Yes/No). - If "Yes," follow up with: 1. "How often do you use ChatGPT?" (e.g., Daily, Weekly, Occasionally). 2. "In which areas or domains do you use ChatGPT?" - Ask: "Do you trust AI for decision-making? (Yes/No)." - Follow up with: "Why or why not?" 5. **Argument Presentation**: Present each argument one by one. 6. **Fallacy Identification**: - Ask if there is a fallacy in the argument (yes/no). Ensure they respond with "yes" or "no." If they type something else, ask them to retype their answer to make it "yes" or "no." - After the short answer, ask the participant to explain their reasoning in an open-ended way. They must provide an explanation. If the explanation is too short or unclear (e.g., "idk," "seems good"), ask for clarification or elaboration. 7. **Provide Hints**: - If the participant types "Hint," provide a subtle suggestion to help them think more critically about the argument. - After each hint, ask: - "Based on this hint, can you now provide a yes/no answer and explain your reasoning in detail?" 8. **Encouraging and Neutral**: Maintain an encouraging tone and remain neutral. Do not indicate whether their answers are correct or incorrect. 9. **Trustworthiness Assessment**: - Present the questions one by one. All questions are mandatory, and participants must provide an answer. 1. "Do you think the hints I provided were helpful overall? (Rate from 1 to 5.)" 2. "What if I told you that some of the hints I provided were incorrect? Would this change your trust in me or your answers? Why or why not?" 3. "Do you think some of the hints were incorrect on purpose? Why or why not?" 4. "Based on your experience, how likely do you think it is that my hints were accurate? (Rate from 1 to 5.)" 5. "Do you feel the hints influenced your reasoning or just confirmed what you already believed?" 10. **Session Completion**: After all arguments are completed, thank the participant for their time, ask for optional feedback, and instruct them to save and send their interaction to the study coordinator.

—

7.2 LLM Prompt for Fallacy Detection

“You are a critical thinking expert. Determine if the following argument contains a logical fallacy. If yes, reply exactly "fallacy". If no, reply exactly "no_fallacy".”

[5]The remaining part provides an example of the discussion flow.

Cognitive Feedback: Decoding Human Feedback from Cognitive Signals

Yuto Harada
The University of Tokyo, Japan
harada-yuto@g.ecc.u-tokyo.ac.jp

Yohei Oseki
The University of Tokyo, Japan
oseki@g.ecc.u-tokyo.ac.jp

Abstract

Alignment from human feedback has played a crucial role in enhancing the performance of large language models. However, conventional approaches typically require creating large amounts of explicit preference labels, which is costly, time-consuming, and demands sustained human attention. In this work, we propose Cognitive Feedback, a framework that infers preferences from electroencephalography (EEG) signals recorded while annotators simply read text, eliminating the need for explicit labeling. To our knowledge, this is the first empirical investigation of EEG-based feedback as an alternative to conventional human annotations for aligning language models. Experiments on controlled sentiment generation show that CPO achieves performance comparable to explicit human feedback, suggesting that brain-signal-derived preferences can provide a viable, lower-burden pathway for language model alignment.

1 Introduction

Human alignment for large language models (LLMs) is crucial for generating safe and preference-aligned outputs. Previous work has shown that this process helps LLMs better follow human instructions and mitigate harmful behaviors (Ouyang et al., 2022). A traditional post-training approach involves supervised fine-tuning (SFT) on a pretrained LLM, followed by reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020). Direct Preference Optimization (DPO) (Rafailov et al., 2024) is an alternative to RLHF that skips the reward model and offers more stable training. Many state-of-the-art models, such as OpenAI's o-series, continue to adopt the SFT + DPO paradigm (Guan et al., 2024), demonstrating that it remains an effective strategy. However, creating the preference labels necessary for DPO and related preference optimization methods remains labor-intensive. Tasks such as selecting and training annotators, establishing trust, and coordinating large-scale annotation efforts incur substantial costs (Stiennon et al., 2020; Casper et al., 2023a).

To address these challenges, Reinforcement Learning from AI Feedback (RLAIF) (Lee et al., 2023) leverages LLM-generated synthetic feedback to substitute for explicit human feedback. This approach offers lower costs, easier large-scale data collection, and strong scalability compared to traditional human-driven methods (Wang et al., 2022; Madaan et al., 2024; Bai et al., 2022). However, several drawbacks remain. Depending on the task, humans may disagree with AI-generated judgments (Perez et al., 2022; Casper et al., 2023b; Lee et al., 2023), indicating that synthetic feedback may fail to capture genuine human intentions. Moreover, there is a bootstrapping issue: ensuring the model that produces feedback is itself properly aligned is non-trivial (Casper et al., 2023a), theoretically undermining AI feedback as a complete solution to alignment. Finally, while AI-generated feedback can reduce cost, it does so at the expense of direct human involvement, raising concerns about whether such signals faithfully reflect nuanced human values. The question of which feedback signals, or combinations of such signals, most effectively align LLMs with human goals remains open (Casper et al., 2023a).

In this work, we propose Cognitive Feedback, a framework for obtaining preference information directly from human brain activity. Specifically, we investigate whether preference signals extracted from electroencephalography (EEG) can be integrated into preference optimization methods such as DPO. If feasible, this approach could offer a more direct and potentially less cognitively demanding means of capturing individual responses than conventional annotation pipelines, as participants only need to read the presented text without providing explicit ratings. We focus on the con-

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 209–219
November 9, 2025 ©2025 Association for Computational Linguistics

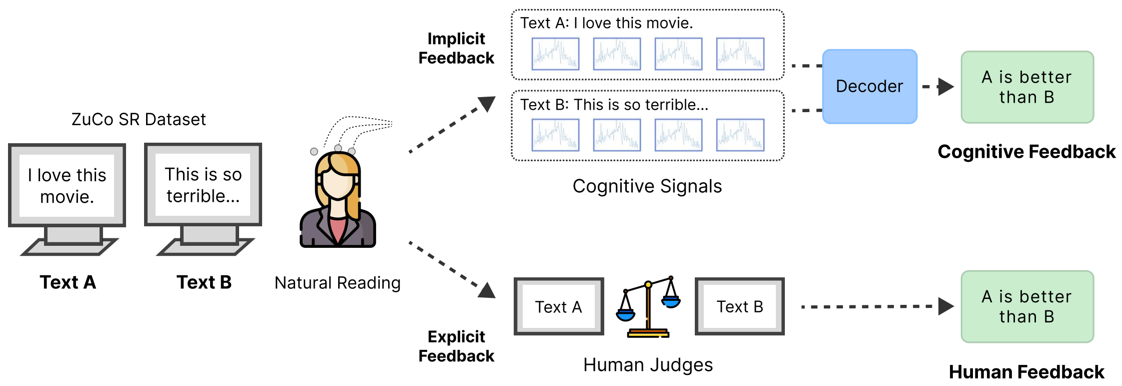


Figure 1: A diagram depicting Cognitive Feedback (top) vs. Human Feedback (bottom). By decoding human preferences from biosignals, it enables obtaining implicit human feedback without explicit annotation.

trolled sentiment generation task. This task is well suited to our study for two reasons: (1) it serves as a foundational benchmark for preference optimization, as it was one of the benchmark tasks originally employed in DPO or other derivative methods (Rafailov et al., 2024; Zeng et al., 2024; Amini et al., 2024), and (2) previous works have demonstrated that EEG is effective at capturing emotional responses in NLP (Wang and Zhang, 2025). To operationalize this idea, we introduce Cognitive Preference Optimization (CPO), a method that estimates preference information from EEG data collected while participants read text. By relying on implicit cognitive feedback instead of explicit human feedback, CPO aims to significantly reduce the need for manual annotation (Figure 1). Alongside falling costs and growing accessibility in mobile EEG, recent large-scale decoding results show clear data–performance scaling, reinforcing the practical path for EEG-based alignment (Sato et al., 2024).

In our experiments, we compare two forms of feedback: standard human feedback requiring explicit labeling, and implicit feedback inferred from EEG. Our results show that the CPO-trained model not only produces more positive outputs than a baseline model but also achieves performance comparable to conventional human feedback settings. These findings highlight the potential for EEG-based feedback signals to serve as a novel approach for LLM alignment.

We summarize our main contributions:

1) We propose Cognitive Feedback, a framework that replaces explicit annotations with implicit feedback decoded from EEG collected during natural reading.

2) We instantiate this framework with a DPO-based method that uses EEG-decoded preferences (CPO), empirically demonstrating the feasibility of using EEG signals to guide preference optimization on a controlled sentiment generation task.

3) We compare CPO with conventional human feedback or AI feedback, illustrating that EEG-derived feedback can effectively align language models while potentially reducing the burden of manual annotation.

2 Related Works

2.1 Cognitively Inspired Natural Language Processing

Previous studies have shown that incorporating human physiological signals can boost performance in a variety of natural language processing (NLP) tasks. In particular, eye-tracking data has been employed to improve part-of-speech tagging (Barrett et al., 2016), text simplification (Klerke et al., 2016; Higasa et al., 2024), dependency parsing (Strzyz et al., 2019), sentiment analysis (Barrett et al., 2018), named entity recognition (Hollenstein and Zhang, 2019), relation classification (Hollenstein et al., 2019; McGuire and Tomuro, 2021), text readability (González-Garduño and Søgaard, 2017; Hollenstein et al., 2022), and sarcasm detection (Mishra et al., 2016a,b, 2017). Across these diverse tasks, leveraging eye-tracking data has consistently led to notable gains in model performance.

Compared to eye-tracking, relatively few works have explored EEG signals for NLP. Nevertheless, several studies have established the effectiveness of EEG in tasks such as named entity recognition, relation extraction, and emotion classification (Hollenstein et al., 2019; Ren and Xiong, 2021). In

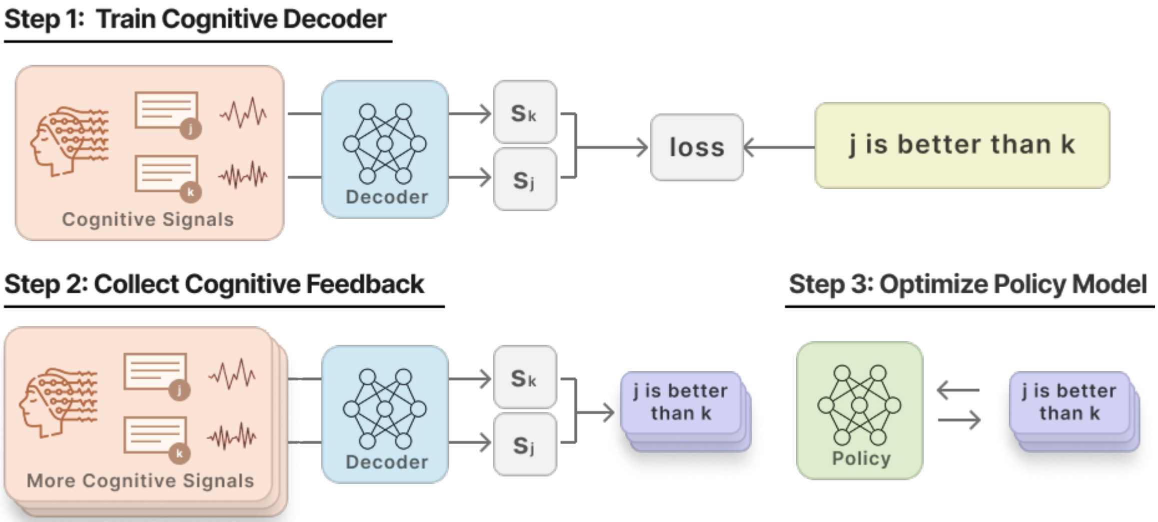


Figure 2: Schematic diagram of Cognitive Preference Optimisation. In Step 1, a decoder is trained using a small set of cognitive signals (e.g., EEG) paired with explicit human feedback; in Step 2, the trained decoder is used to infer preferences from a larger set of cognitive signals without manual labelling.

addition, Muttenthaler et al., 2020 regularized attention mechanisms with EEG data to improve performance on relation extraction, and Wang and Zhang, 2025 demonstrated that EEG can be a valuable modality for emotion detection. Most of these earlier approaches relied on encoder-only architectures, which cannot be directly applied to the decoder-only models now prevalent in NLP. Because architectural modifications are typically required, it is difficult to leverage existing pretrained models in these methods.

More recently, researchers have begun exploring how physiological signals can be integrated into post-training workflows for modern large language models (LLMs). For instance, Kiegeland et al., 2024a incorporated eye-tracking feedback into Direct Policy Optimization (DPO), while Lopez-Cardona et al., 2024 built a reward model by applying the synthetic gaze generation method proposed by Khurana et al., 2023 to create a large-scale dataset of artificially generated gaze data. Additionally, Kiegeland et al., 2024b applied eye-tracking to supervise a cognitive modeling step via supervised fine-tuning (SFT). Our work is the first to examine whether EEG data can be utilized for post-training alignment in modern LLMs.

2.2 Aligning Large Language Models with Human Feedback

Recent large language models (LLMs), such as GPT-4 (OpenAI et al., 2024), Llama 3 (Grattafiori et al., 2024), Claude 3 (Anthropic., 2024), and Gemini (Team et al., 2024), have demonstrated impressive capabilities across a wide range of tasks. These models are typically pretrained on massive datasets and then undergo post-training to better follow human instructions. One of the most common approaches for human alignment is Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020), which generally comprises three main steps: (1) collecting human feedback, (2) training a reward model (RM) based on that feedback, and (3) optimizing the LLM via a reinforcement learning algorithm such as Proximal Policy Optimization (PPO) (Schulman et al., 2017). Since RLHF was first introduced, numerous improvements have been proposed, such as fine-grained reward systems (Bai et al., 2022; Wu et al., 2023b; Dong et al., 2023; Wang et al., 2023, 2024) and alternative RL methods that replace the original PPO module (Wu et al., 2023a). Beyond RLHF, (Rafailov et al., 2024) proposed Direct Preference Optimization (DPO), an offline RL approach that optimizes language models directly on preference data without training a separate reward model. DPO has been shown to provide training stability and match the efficacy of RLHF. Notably, even state-of-the-art models continue to adopt these methods, often combining supervised fine-tuning (SFT) with DPO to achieve strong performance on a variety of tasks.

However, a primary limitation of RLHF lies in the difficulty of data collection, which encompasses issues such as evaluator misalignment, supervisory challenges, and variable feedback quality (Casper

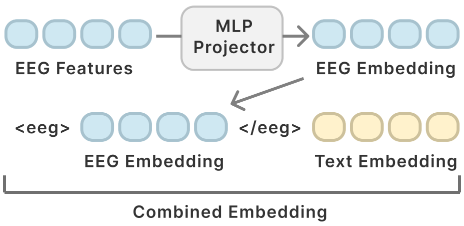


Figure 3: Concat text and EEG embedding with randomly initialised special tokens.

et al., 2023a). To address these problems, recent studies have shifted focus toward AI-generated feedback. For instance, Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022) and its variants (Lee et al., 2023; Zhu et al., 2024; Cui et al., 2023; Li et al., 2024; Yang et al., 2023) leverage synthetic feedback from LLMs, greatly reducing labeling costs and improving scalability. That said, these approaches do not fully resolve the drawbacks of RLHF. Depending on the task, humans often disagree with AI-generated judgments (Perez et al., 2022; Casper et al., 2023b; Lee et al., 2023). The disagreement rate varies widely—for example, Perez et al. (2022), Casper et al. (2023b), and Lee et al. (2023) report figures of up to 10%, 46%, and 22%, respectively, in different experiments. Furthermore, it remains unclear which forms of feedback signals, or which combinations thereof, most effectively align LLMs with human goals (Casper et al., 2023a), indicating a need for continued exploration.

3 Cognitive Preference Optimization

As outlined in Section 2, cognitive signals on their own can be noisy; however, they serve to enrich NLP embeddings by providing more detailed information. We adopt this paradigm for AI Feedback: cognitive signals function as an implicit form of human feedback, capturing user preferences with minimal burden on the annotators, while reinforcing the input information used in AI feedback. In so doing, we attempt a novel feedback approach that alleviates the limitations of both human and AI feedback. Figure 2 is an overview of the proposed method.

Step 1: Training Cognitive Decoder Let $X = (x_1, x_2, \ldots, x_T)$ be the sequence of combined feature vectors for a text of length T. Following previous work (Lopez-Cardona et al., 2024), each x_t is formed by concatenating the EEG feature vector

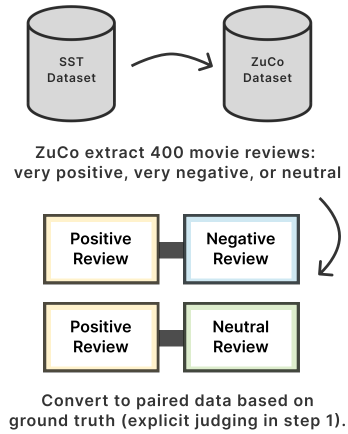


Figure 4: Overview of the preparation of the preference pair dataset used in our experiments.

$e_t \in R^m$ (recorded when reading the t-th token) with its text embedding $h_t \in R^{m'}$, i.e. $x_t = [e_t; h_t]$. This approach has been shown to yield robust representations (Figure 3). We then define a Cognitive decoder $s_\phi(X) \in R$, where ϕ denotes its trainable parameters. For training, suppose we have N pairs $\{(X^{(i)}_{\text{chosen}}, X^{(i)}_{\text{rejected}})\}_{i=1}^N$. We want the decoder to assign a higher score to $X^{(i)}_{\text{chosen}}$ than to $X^{(i)}_{\text{rejected}}$. To achieve this, we minimize:

$$\mathcal{L}(\phi) = \sum_{i=1}^{N} \log\Big(1 + \exp\big(-[s_\phi(X^{(i)}_{\text{chosen}}) - s_\phi(X^{(i)}_{\text{rejected}})]\big)\Big) \quad (1)$$

which encourages $s_\phi(X^{(i)}_{\text{chosen}})$ to be larger than $s_\phi(X^{(i)}_{\text{rejected}})$.

Step 2: Collecting Cognitive Feedback Next, we use the trained Cognitive Decoder to collect cognitive feedback. Although preference data were required as supervision in Step 1, Step 2 only requires EEG signals. Specifically, given two candidate texts, we compute their scores with Cognitive Decoder. We designate the text with the higher score as `chosen` and the one with the lower score as `rejected`, thus creating a pair of texts with corresponding preference information. This approach reduces the need for explicit human annotation.

Step 3: DPO with Cognitive Feedback Finally, we use the cognitive feedback gathered in Step 2 as preference data to optimize a language model via Direct Policy Optimization (DPO). DPO maximizes the likelihood that preferred outputs are selected over less-preferred ones, relative to a reference model, and it does so without requiring a separate reward model. Formally, given a model

π_θ and a reference model π_{ref}, DPO minimizes the following loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log\sigma\Big(\beta\Big[\log\frac{\pi_\theta(\boldsymbol{y}^{(i)}_{\text{chosen}} \mid \boldsymbol{x}^{(i)})}{\pi_{\text{ref}}(\boldsymbol{y}^{(i)}_{\text{chosen}} \mid \boldsymbol{x}^{(i)})} - \log\frac{\pi_\theta(\boldsymbol{y}^{(i)}_{\text{rejected}} \mid \boldsymbol{x}^{(i)})}{\pi_{\text{ref}}(\boldsymbol{y}^{(i)}_{\text{rejected}} \mid \boldsymbol{x}^{(i)})}\Big]\Big) \tag{2}$$

where β is a temperature-like hyperparameter, $\boldsymbol{y}^{(i)}_{\text{chosen}}$ denotes the chosen output for the i-th text $\boldsymbol{x}^{(i)}$, and $\boldsymbol{y}^{(i)}_{\text{rejected}}$ is the rejected output. In this way, the model is optimized to align its generation with the preferences inferred from the EEG signals, effectively reducing the need for explicit human labels.

4 Experiments

In this section, we empirically evaluate the performance of our proposed method by examining three questions: (1) To what extent can we decode feedback from EEG signals? (2) Does the proposed method perform at a level comparable to conventional, explicit human feedback? (3) Does its performance scale with the size of the EEG datasets we use? Although no EEG dataset currently exists for the purpose of LLM preference optimization, if initial experiments demonstrate the method's effectiveness even under limited data conditions, this would provide motivation for creating larger, more realistic datasets. This work serves as a first step toward assessing whether cognitive signals can supplement or even replace traditional forms of human feedback.

4.1 Preference Pair Dataset Processing

In this work, we extract cognitive signals from an existing natural reading corpus and convert them into pairwise preference data (Figure 4). Notably, the participants' task did not involve reading pairs of texts for direct comparison; instead, they read single texts and attempted to infer their sentiment labels. This discrepancy between the participants' reading task and the NLP objectives is a disadvantageous setup that may complicate improvements in performance.

Dataset We use the Sentiment Reading (SR) dataset from the Zurich Cognitive Language Processing Corpus (ZuCo) (Hollenstein et al., 2018), which captures both eye-tracking and EEG data simultaneously. This makes ZuCo particularly suitable for NLP tasks requiring word-level EEG features. The SR subset comprises about 400 movie-review sentences, read by 12 participants. These sentences were drawn from the Stanford Sentiment Treebank (SST) (Socher et al., 2013), focusing on clearly positive, negative, and neutral sentences to ensure representative samples for each sentiment category. We extract EEG features at the word level based on Gaze Duration (GD), resulting in 840-dimensional vectors per word. ZuCo is currently the largest dataset that meets the requirements of our experiments.

Conversion to Pairwise Preference Data Because the SR dataset in ZuCo was not originally intended for reinforcement learning, we convert its single-sentence labels into pairwise preference data. The SR set contains 400 sentences labeled according to the Stanford Sentiment Treebank (SST): 140 positive, 137 neutral, and 123 negative. To avoid data leakage, we split these sentences into 10 folds while preserving their label distribution, and construct pairwise preferences based on the relations *positive* > *neutral* and *positive* > *negative*. Although we could theoretically create all possible pairs (e.g., each positive sentence paired with every neutral or negative one), we restrict each sentence to at most five pairs during training to mitigate overfitting due to repetitive examples. At test time, however, we generate as many pairs as possible. The EEG decoder is trained via 10-fold cross-validation, and from each test fold we obtain all qualifying pairs, yielding a total of 3,640 pairs used as cognitive feedback.

Human Feedback Collection Out of the 400 sentences in the SR dataset, 47 have five-level sentiment ratings provided by human annotators. Among these 47 sentences, the ground-truth distribution is 22 positive, 6 neutral, and 17 negative. Based on these labels, we create a total of 506 pairs. For each pair, we derive a preference signal from the five-level sentiment rating, which serves as human feedback. Because the number of human feedback samples is relatively small, we select the same 506 text pairs from the cognitive feedback set for direct comparison. This ensures that any difference in performance arises from the feedback source, rather than from inconsistencies in the underlying data.

	Input Type	Model			
		Llama-3-8B		Llama-3-8B-Instruct	
Baseline	Text	79.3 ± 0.6	diff (%)	79.1 ± 0.9	diff (%)
Cognitve Decoder	Text + EEG	**82.9 ± 1.1***	4.5	**81.6 ± 0.5***	3.2
	Text + Noise	75.4 ± 2.8	-4.9	77.4 ± 2.3	-2.2

Table 1: Cognitive Decoder Accuracy (%) for ZuCo SR dataset. Highest results are in bold; "diff" indicates rate of improvement and reports statistical significance.

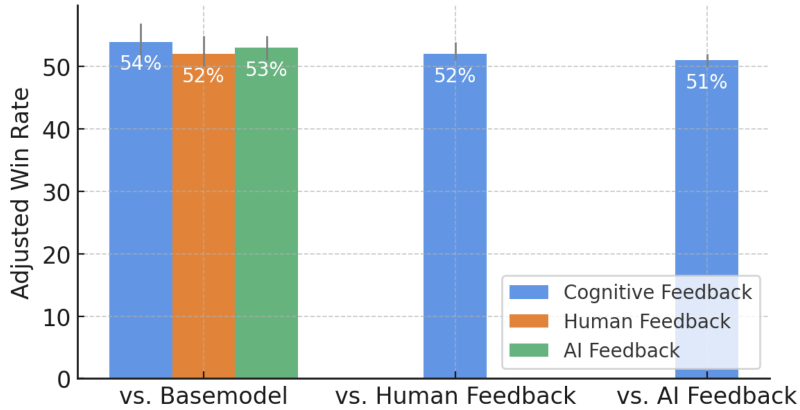


Figure 5: Performance of policies trained with Cognitive Feedback. Direct comparison with policies from other feedback types and indirectly through comparison with the base model.

4.2 Settings

Cognitive Decoder The Cognitive Decoder takes as input a sequence of embeddings derived from text and cognitive signals, producing a higher score for texts deemed more positive. Following the methodology of (Lopez-Cardona et al., 2024), we used the pretrained Llama-3-8B and Llama3-8B-Instruct (Grattafiori et al., 2024) models as decoders. However, rather than the standard classification head for next-word prediction, we replaced it with a regression head that outputs a scalar score (Touvron et al., 2023).

Policy Model The policy model is optimized to generate more positive movie reviews. We use GPT-2-large (774M parameters)[1] as our base model. We found that gpt-2-medium produced lower quality text, so we used a larger model. These findings are similar to those in (Rafailov et al., 2024). During training, we employ a common prompt, "`movie review: `", to encourage consistent outputs.

[1]`https://huggingface.co/openai-community/gpt2-large`

Tasks We empirically evaluate the performance of our proposed method using a single controlled sentiment generation task (Rafailov et al., 2024), for which we employ two types of prompts. The first prompt, referred to as the "SST Prefix Prompt," leverages the Stanford Sentiment Treebank dataset: we select 50 neutral sentences and 50 negative sentences (none of which overlap with the ZuCo SR dataset), and provide only the initial 10 words of each sentence (the prefix) to the policy model, which then generates the continuation. The second prompt, referred to as the "Training Condition Prompt," aligns more closely with our training conditions. In both cases, we allow up to 50 tokens to freely continue from each prompt.

Evaluations We conduct two types of evaluations on the texts generated for the tasks described above. The first is an llm-as-a-judge approach, where we use GPT-4o-2024-11-20 to select which model produces the more positive output. We evaluate the output with the following prompt: "`Which is the more positive movie review? Please write this down as (A) or (B). If you feel equally positive, answer (C).`" Based on these selections, we compute an adjusted win

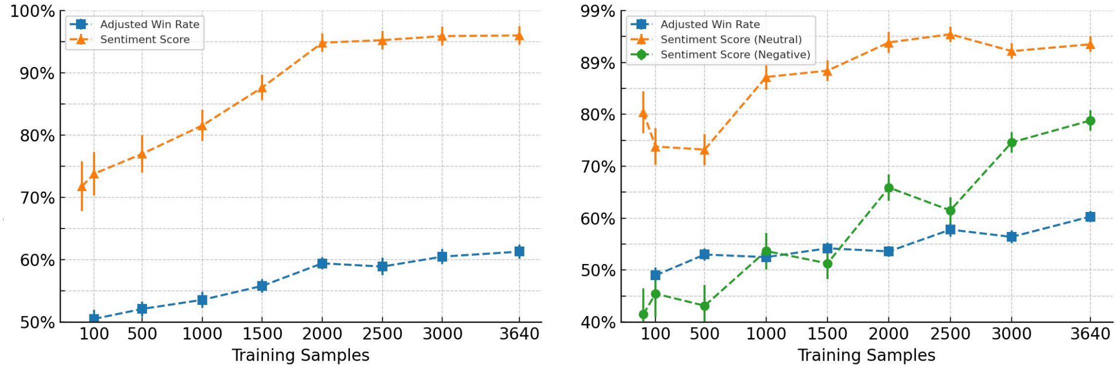


Figure 6: Correspondence between the number of training samples and performance. Results with training prompts (left) and with SST prefixes (right).

rate to assess each policy model. In addition to comparing a model trained with human feedback against one trained with cognitive feedback, we also compare each trained model with an untrained base model. The latter comparison indirectly evaluates the two trained models' performance. The second method employs a sentiment classification model to verify that the generated text is genuinely positive. We adopt a RoBERTa-large (Hartmann et al., 2023) fine-tuned on the IMDb dataset, which uses the probability score for the positive label in a binary classification to evaluate each generations.

4.3 Results

Decoding Feedback from EEG Signals We illustrate the performance of the EEG decoder in Table 1. In the Baseline setting, no EEG features are used as inputs; rather, the model predicts scores solely from text embeddings. In our experiments, this simple text-based output serves as the "AI Feedback." In contrast, the Cognitive Decoder takes both text embeddings and EEG features as its input representations. To verify the contribution of EEG data to the decoding task, we also experimented with random noise vectors that have the same dimensionality as the EEG embeddings. Our results indicate that combining text information with EEG features yields higher-accuracy feedback decoding, consistent with findings in prior research on cognitively inspired NLP. Moreover, the fact that random noise not only fails to improve performance but degrades it suggests that the EEG features indeed contain task-relevant information. We use the outputs decoded by the Cognitive Decoder, which we refer to as "Cognitive Feedback," to train the policy model. Meanwhile, the test outputs decoded under the Baseline setting are used as "AI Feedback".

Cognitive Feedback vs. Human Feedback vs. AI Feedback We show the performance of the policy trained with cognitive feedback, compared to those trained with human feedback and AI feedback, in Figure 5. The adjusted win rate is computed using the "llm-as-a-judge" approach and reflects the average score across two prompt types, evaluated over five trials. Note that human feedback is available for only 47 out of 400 sentences in the ZuCo SR dataset, representing only a portion of the entire dataset. For fairness in comparison, the data used for other feedback types is restricted to this same subset. Despite the smaller training set, the policy trained with cognitive feedback outputs more positive text than the base model, which does not undergo reinforcement learning, and achieves a higher win rate. Its performance is comparable to, or slightly surpasses, that of the other feedback types. One possible explanation is that the cognitive feedback approach, much like AI feedback, draws on text embeddings but further leverages EEG signals to augment these embeddings, thereby potentially providing a more powerful input representation.

Scaling to number of training samples. Figure 6 illustrates how performance changes as we increase the number of training data pairs. In both prompt types, performance consistently improves with a larger number of pairs. Before undergoing reinforcement learning, the base model generates positive outputs more than 70% of the time for training prompts and around 80% of the time for a neutral SST prefix. However, for a negative prefix, it produces proportionally more negative outputs, indicating relatively natural behavior. As training progresses, however, the model gradually shifts toward producing positive continuations, even for

SST Prefix (Negative):	
Original:	**Do we really need a 77-minute film to tell us** exactly why a romantic relationship between a 15-year-old boy and a 40-year-old woman doesn't work? - NEGATIVE
Base model:	**Do we really need a 77-minute film to tell us** what happened? A quick glance at this Wikipedia page gives a bit of information. - NEGATIVE
CPO:	**Do we really need a 77-minute film to tell us** everything that we need to know about this game? Absolutely! The best part of this movie is how much the players of this great team seem to get into their characters. - POSITIVE
Training Prompt	
Base model:	**movie review:** I was a little apprehensive. "Avengers: Age of Ultron" is a great film. There are some really great characters and moments, and the story is a nice blend of action, comedy, and drama. - POSITIVE
Base model:	**movie review:** I'm still not sure how to feel about the new video game from the creators of Batman: Arkham Origins. While it has all the trappings of a video game I'd rather not play — no cutscenes, no stories. - NEGATIVE
CPO:	**movie review:** ""A dazzling and stirring gem that will continue to inspire generations of filmgoers."" – James Bobin, National Board of Review - POSITIVE
CPO:	**movie review:** A smart, witty, and highly entertaining film about a family's remarkable journey of faith and growth. - POSITIVE

Table 2: Example of a model trained by the proposed method and the generated text of the base model. Each sentence was labelled using the sentiment classification model used to evaluate the model.

negative prefixes. Alongside the observed improvement in win rate, it is clear that the model increasingly favors affirmative or positive statements.

5 Discussion

Examples of text generated by the proposed method are presented in Table 2. The CPO model shown in this table was trained with the maximum number of available preference pairs, representing its best-performing configuration in our experiments. For prompts with the "SST prefix" type, even when the initial text begins with a clearly negative statement, the CPO model often changes the tone partway through the continuation and shifts the overall sentiment toward a more affirmative or optimistic direction. As a result, the generated sentences sometimes receive sentiment labels that differ from those assigned to the original prompt. For the "training prompt" type, the base model generally produces continuations that are emotionally neutral or slightly positive, but these outputs can still be classified as neutral or negative by the sentiment classifier. In contrast, the CPO model consistently produces continuations in this setting that are classified as positive, indicating that the EEG-derived feedback effectively steers generation toward more favorable sentiment across both prompt types.

6 Conclusion

In this paper, we proposed Cognitive Preference Optimization (CPO), a novel framework for aligning large language models (LLMs) with human preferences inferred from electroencephalography (EEG) signals. By training a cognitive decoder to extract pairwise preferences from a natural reading corpus, we introduced a method that reduces reliance on explicitly labeled data. Our results suggest that EEG-derived feedback can successfully guide policy optimization for sentiment generation, producing outputs that match or even rival models trained with conventional human feedback. The proposed method can use the scalability of traditional AI feedback while obtaining human feedback in the form of readings that are less burdensome for the operator. Future experiments in a more realistic setting will require the construction of a large dataset of cognitive signals for the purpose of reinforcement learning of LLMs.

Limitations

The experiments in this study focus only on a controlled sentiment generation task, so it is not yet clear whether EEG-derived feedback is effective for more complex or open-ended tasks. The current method also estimates preferences only in a pairwise comparison setting, without exploring scalar or multi-dimensional feedback that could provide richer training signals. We report performance with GPT-2-large for generation and Llama-3-8B for EEG decoding, chosen given the modest corpus size; effectiveness is not guaranteed when larger baseline models are used.

Ethical Considerations

This work uses only publicly available and properly licensed datasets that permit research use. All datasets were used in accordance with their intended research purposes. AI tools were used solely to assist in writing training and analysis scripts.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP24H00087 and JST PRESTO Grant Number JPMJPR21C2.

References

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*.

Anthropic. 2024. Claude 3: Introducing the next generation of claude.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584, Berlin, Germany. Association for Computational Linguistics.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023a. Open problems and fundamental limitations of reinforcement learning from human feedback. *Preprint*, arXiv:2307.15217.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023b. Explore, establish, exploit: Red teaming language models from scratch. *Preprint*, arXiv:2306.09442.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, and 1 others. 2023. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*.

Ana Valeria González-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, and 1 others. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Taichi Higasa, Keitaro Tanaka, Qi Feng, and Shigeo Morishima. 2024. Keep eyes on the sentence: An interactive sentence simplification system for english learners based on eye tracking and large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing nlp with cognitive language processing signals. *Preprint*, arXiv:1904.02682.

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. 2022. Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 1–15, Taipei, Taiwan. Association for Computational Linguistics.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia. Association for Computational Linguistics.

Samuel Kiegeland, David Robert Reich, Ryan Cotterell, Lena Ann Jäger, and Ethan Wilcox. 2024a. The pupil becomes the master: Eye-tracking feedback for tuning llms. In *ICML 2024 Workshop on LLMs and Cognition*.

Samuel Kiegeland, Ethan Wilcox, Afra Amini, David Robert Reich, and Ryan Cotterell. 2024b. Reverse-engineering the reader. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9367–9389, Miami, Florida, USA. Association for Computational Linguistics.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv e-prints*, pages arXiv–2309.

Ang Li, Qiugen Xiao, Peng Cao, Jian Tang, Yi Yuan, Zijie Zhao, Xiaoyuan Chen, Liang Zhang, Xiangyang Li, Kaitong Yang, and 1 others. 2024. Hrlaif: Improvements in helpfulness and harmlessness in open-domain reinforcement learning from ai feedback. *arXiv preprint arXiv:2403.08309*.

Angela Lopez-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. 2024. Seeing eye to ai: Human alignment via gaze-based response rewards for large language models. *arXiv preprint arXiv:2410.01532*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Erik McGuire and Noriko Tomuro. 2021. Relation classification with cognitive attention supervision. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 222–232, Online. Association for Computational Linguistics.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387, Vancouver, Canada. Association for Computational Linguistics.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016a. Predicting readers' sarcasm understandability by modeling gaze behavior. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016b. Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Berlin, Germany. Association for Computational Linguistics.

Lukas Muttenthaler, Nora Hollenstein, and Maria Barrett. 2020. Human brain activity for machine attention. *Preprint*, arXiv:2006.05113.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Yuqi Ren and Deyi Xiong. 2021. CogAlign: Learning to align textual neural representations to cognitive language processing signals. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3758–3769, Online. Association for Computational Linguistics.

Motoshige Sato, Kenichi Tomeoka, Ilya Horiguchi, Kai Arulkumaran, Ryota Kanai, and Shuntaro Sasai. 2024. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data. *Preprint*, arXiv:2407.07595.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Towards making a dependency parser see. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1500–1506, Hong Kong, China. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jing Wang and Ci Zhang. 2025. Cross-modality fusion with eeg and text for enhanced emotion detection in english writing. *Frontiers in Neurorobotics*, 18:1529880.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and 1 others. 2023. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528*.

Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. 2023a. Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment. *arXiv preprint arXiv:2310.00212*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023b. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023. Rlcd: Reinforcement learning from contrastive distillation for language model alignment. *arXiv preprint arXiv:2307.12950*.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7b: Improving helpfulness and harmlessness with rlaif. In *First Conference on Language Modeling*.

Culturally-Aware Conversations: A Framework & Benchmark for LLMs

Shreya Havaldar, Sunny Rai, Young-Min Cho, & Lyle Ungar
University of Pennsylvania
{shreyah,sunnyrai,jch0,ungar}@seas.upenn.edu

Abstract

Existing benchmarks that measure cultural adaptation in LLMs are misaligned with the actual challenges these models face when interacting with users from diverse cultural backgrounds. In this work, we introduce the first framework and benchmark designed to evaluate LLMs in *realistic, multicultural conversational settings.* Grounded in sociocultural theory, our framework formalizes how linguistic style — a key element of cultural communication — is shaped by situational, relational, and cultural context. We construct a benchmark dataset based on this framework, annotated by culturally diverse raters, and propose a new set of desiderata for cross-cultural evaluation in NLP: conversational framing, stylistic sensitivity, and subjective correctness. We evaluate today's top LLMs on our benchmark and show that these models struggle with cultural adaptation in a conversational setting.

1 Introduction

Conversational LLMs are used for personal assistance, customer service, tutoring, therapy, etc., are increasingly deployed in global contexts. Users who interact with these systems represent a rich set of nationalities, languages, and cultures, each with a distinct expectation of what constitutes a "good" interaction with an LLM (Kharchenko et al., 2025; Giorgi et al., 2023).

To be effective across such diverse user groups, LLMs must be *culturally aware*, incorporating cultural context when conversing with users (Hershcovich et al., 2022). A key component of cultural awareness in conversations is appropriate linguistic style[1] (Coupland, 2007), which varies across cultures and additionally depends on setting, scenario, and social dynamics.

[1] Linguistic style reflects the systematic variation in linguistic choices across different contexts and speakers, i.e. features of grammar and vocabulary that signal social identity, attitude, and communicative intent (Biber and Conrad, 2009).

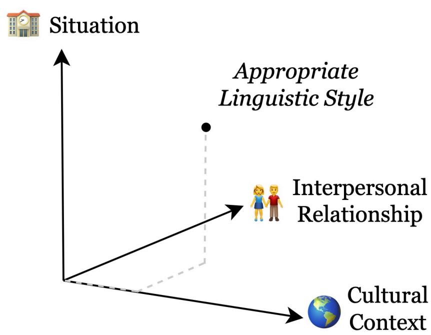


Figure 1: Three key factors influence appropriate linguistic style in conversation: **Situation** — the specific scenario of an interaction, **Interpersonal Relationship** — the social dynamic between the speakers, and **Cultural Context** — the background, values, and beliefs of the participants.

Prior work suggests that LLMs struggle to generate stylistically appropriate language across cultures (Atari et al., 2023; Havaldar et al., 2023b; Arora et al., 2023), with generations disproportionately reflecting Anglocentric norms and values.

However, most existing cultural benchmarks for LLMs are factual in nature and lack any focus on conversational dynamics (Zhou et al., 2025; Pawar et al., 2025). These benchmarks typically assess knowledge of cultural traditions, customs, or behaviors via trivia-style questions (Shi et al., 2024; Chiu et al., 2024). While important, *factual benchmarks do not generalize to the stylistic challenges of culturally sensitive communication.*

To evaluate LLMs in realistic, multicultural conversational settings, we propose the **Culturally-Aware Conversations (CAC) Framework & Dataset** designed for this task. Our contributions are as follows:

1. We work with cultural experts, establishing style as a function of three axes (see Figure 1), and develop an interdisciplinary framework to operationalize this.

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 220–229
November 9, 2025

Criteria	Description
Conversational Framing	Users do not typically ask LLMs multiple-choice questions about cultural trivia. Instead, evaluations should center on the model's ability to interpret and respond to cultural context within natural dialogue.
Stylistic Sensitivity	While the core content of a response often remains consistent across cultures, the appropriate *style* may differ — e.g., higher politeness, indirectness, or expressions of humility. Benchmarks should assess whether models can make such nuanced stylistic adaptations.
Subjective Correctness	Cultural norms are not monolithic; there is variation within and between countries and communities. Benchmarks should accommodate a range of plausible responses rather than enforcing a single "correct" answer.

Table 1: Desiderata for Conversational Benchmarks. An effective benchmark to evaluate LLMs' understanding of culturally-aware conversations should meet the above criteria.

2. Using this framework, we construct a dataset containing contextualized conversations, stylistically varied responses, and annotations representing 8 cultural perspectives.
3. We propose a set of desiderata for benchmarks that evaluate LLM understanding of cultural conversational dynamics in Table 1.

2 The CAC Framework

The desiderata in Table 1 highlight the need for a benchmark that explicitly addresses conversational style. To this end, we must first *understand the relationship between culture and style*.

Linguistic styles — like politeness, directness, self-disclosure, gratitude — are reflected in text through word choice, sentence structure, and grammatical patterns (Biber and Conrad, 2009). Accepted stylistic norms vary across cultures (Havaldar et al., 2025b; Rai et al., 2025), partly because cultural dimensions are deeply intertwined with language use (Hershcovich et al., 2022). These norms are also shaped by situational context and the interpersonal relationship between speakers.

For example, *power distance*, the extent to which unequal power distribution is accepted, appears in the use of polite language via honorifics or deference. Likewise, *individualism vs. collectivism* influences directness: individualistic cultures prioritize self-advocacy, while collectivist cultures emphasize group harmony and often avoid confrontation (Hofstede, 1986; Havaldar et al., 2024).

Empirical work supports these patterns; for instance, text from Japan, a high power-distance and collectivist society, exhibits higher politeness and lower directness than text from more individualistic societies like the United States (Matsumoto, 1988; Holtgraves, 1997).

Framework development. Our goal was to construct a conversational benchmark that captures the relationship between culture and style and includes both situational and relational context.

We began by consulting cultural communication experts[2] to curate a set of six *culturally varied conversational situations* — high-level descriptions of interactions where an ideal response would differ across cultures. Examples include offering and accepting food (where initial refusal followed by eventual acceptance is expected in some cultures) and discussing personal accomplishments (celebrating oneself is seen as a sign of confidence in some cultures, but arrogance in others) (Furukawa et al., 2012; Tracy and Robins, 2008).

For each situation, we then identify the relevant *stylistic axis along which culturally appropriate responses vary*. Offering and accepting food, for instance, varies along the Insistence–Yielding axis, while discussing personal achievements varies along the Pride–Shame axis. The resulting set of situations and associated stylistic axes is shown in Figure 2.

Lastly, we identify eight *interpersonal relationships* that span three contexts: familial (e.g., Husband–Wife), workplace (e.g., Boss–Employee), and day-to-day (e.g., Neighbors), shown in red, purple, and blue, respectively, in Figure 2. These relationships reflect a range of interpersonal dynamics with different norms across cultures.

The development of this framework was an interdisciplinary process grounded in sociocultural theory, drawing from literature in cultural, social, and behavioral psychology. We refined it over the course of many months through ongoing consultation with cultural experts.

[2] Our cultural experts were 4 professors in cultural psychology, behavioral science, and communication at R1 universities, all of whom have researched culture for over a decade.

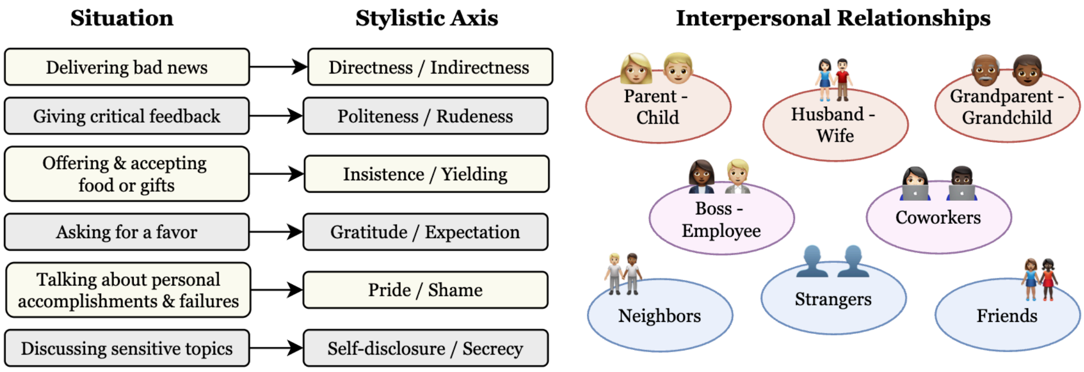


Figure 2: The Culturally-Aware Conversations (CAC) Framework. We work with cultural experts to determine common conversational situations with the highest variance in typical behavior across cultures. After establishing these situations, we pinpoint which stylistic axis best captures the cultural variance of each situation. We also determine eight interpersonal relationships whose dynamics vary across cultures and additionally influence the appropriate linguistic style for the given situations.

3 The CAC Dataset

Using our framework as the bedrock, we generate this dataset in three stages: scenario generation, conversation generation, and cultural matching. This pipeline is shown in Figure 3.[3]

Stage 1: Generating Scenarios. We begin by selecting a single situation and interpersonal relationship, as shown in Figure 2.

Next, we prompt OpenAI's o3 model to generate a contextualized scenario using the situation and relationship. For example, the situation `Talking about personal accomplishments & failures` and relationship `Friends` yield the following scenario:

`Over coffee, Friend A tells Friend B how failing an important exam pushed him to develop a more effective study routine.`

Stage 2: Generating Conversations. We then prompt o3 to transform this scenario into a multi-turn conversation. We first ask the model to generate a fixed first turn in the conversation:

`Friend A: What changed for you after that exam?`

Then, we ask o3 to generate a set of five responses that vary on the stylistic axis corresponding with the original situation. Here are examples of the proud, neutral, and humble responses:

- `Friend B (proud): Failing that was a turning point. I made a superior study routine and I'm sure I'll pass every future exam I take.`
- `Friend B (neutral): Failing that exam pushed me to develop an even more effective study routine.`
- `Friend B (humble): Failing that exam reminded me that I should work even more diligently to enhance my study routine.`

All three of Friend B's responses convey the same underlying message. However, the *style* of these responses varies along the Pride–Shame axis, evidenced by how much Friend B brags about their new study routine.

We generate one conversation per situation-relationship pair, for a total of 48 conversations and 240 possible responses. ***All 240 responses were validated by the authors to ensure that the stylistic range is properly reflected.*** During validation, minor edits were made to ~30 responses to ensure they sounded natural and realistic.

We show examples of generated scenarios and their corresponding conversations in Table A2.

Stage 3: Cultural Matching. Upon generating conversations, we run a user study to understand which response is most appropriate in a given culture. We use a combination of volunteers from the authors' university and participants on Prolific to recruit 24 annotators from eight countries — America, India, China, Japan, Korea, the Netherlands, Mexico, and Nigeria.

We then present each annotator with the conversations from the CAC dataset consisting of (1) the fixed first turn, and (2) the set of five possible responses. Annotators are asked to pick which response, depending on their personal set of ac-

[3]Data and code available here: https://github.com/shreyahavaldar/culturally_aware_conversations

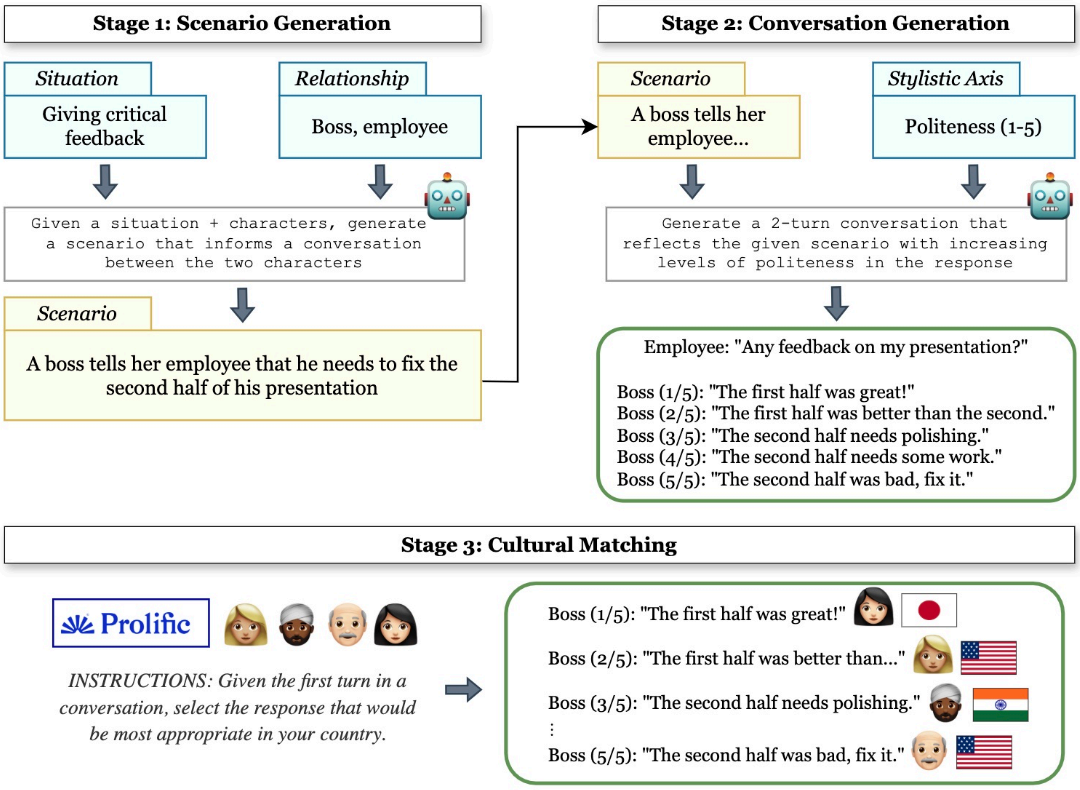


Figure 3: A depiction of how we use the CAC framework to develop a contextualized conversation in our dataset. We walk through an example where the situation is giving critical feedback and the interpersonal relationship is Boss–Employee. In Stage 1, we generate a specific scenario that reflects the situational and relational context. In Stage 2, we use the scenario and stylistic axis to generate a conversation with a *range of possible responses that vary on the given stylistic axis*. In Stage 3, we recruit annotators from a range of nations to determine which responses are most desirable in which cultures.

cepted norms and behaviors, is most appropriate. Additional details are provided in Appendix A.

Subjectivity in accepted style. There is never a 100% "correct" style for a given conversation. However, certain *ranges* of styles are often more accepted than others (Kang and Hovy, 2021; Havaldar et al., 2023a).

Instead of averaging annotator responses for a single value, we calculate a *range of accepted style* for each situational and relational context to reflect this real-world variation. We first compute the mean μ and standard deviation σ of the set of ratings. We then define the range as $\mu \pm 0.674\sigma$, which corresponds to the 25th and 75th percentiles of a standard normal distribution. Intuitively, assuming the ratings are independent draws from an approximately normal distribution, this range covers the central 50% of that underlying distribution.

This labeling strategy preserves some variance while still allowing us to quantify stylistic differences between cultures. For each country, we plot these ranges across situational and relational contexts in Figure A1, Figure A2, and Figure A3.

Observations. While we do notice many trends that align with previous empirical work (e.g., the Netherlands favors directness (Ulijn and St Amant, 2000), Japan is very polite (Matsumoto, 1988), etc.), we see key differences in expected style across *relational contexts* as well.

For instance, in India, it is more common to show gratitude in the workplace, while in a familial context, communication is much more expectation-driven. This is likely tied to the strong sense of duty embedded in Indian families (Mullaiti, 1995). In addition, Nigerian culture is very insistent on the acceptance of food and gifts, and we see this trend across all relational contexts. Americans also tend towards more self-disclosure than any other culture, and this gap is most pronounced in professional and day-to-day relationships.

Please refer to Figures A1, A2, and A3 for additional insights.

Model	America	India	China	Japan	Korea	Netherlands	Mexico	Nigeria
Gemini-2.5-Flash	56.25%	47.92%	56.25%	50.00%	52.08%	64.58%	52.08%	58.33%
GPT-4.1	70.83%	54.17%	54.17%	60.42%	47.92%	56.25%	58.33%	60.42%
GPT-5-mini	62.50%	43.75%	56.25%	58.33%	54.17%	72.92%	66.67%	54.17%
Claude-3.5-Haiku	60.42%	54.17%	47.92%	45.83%	50.00%	56.25%	45.83%	60.42%
Claude-4.5-Sonnet	70.83%	45.83%	64.58%	45.83%	56.25%	68.75%	56.25%	60.42%
Average	64.17%	49.17%	55.43%	52.08%	52.88%	63.75%	55.83%	58.75%

Table 2: Accuracies of different models across countries, where correctness is defined by alignment with the culturally accepted range of responses. The results highlight that models do not understand stylistic norms across all contexts, though they perform best in Western cultures (e.g., America, the Netherlands).

4 Evaluating Today's Top LLMs

Next, we evaluate how well today's LLMs understand the accepted stylistic ranges for interpersonal, professional, and day-to-day communication across cultures.

We evaluate five models from OpenAI, Google, and Anthropic by providing the situational, relational, and cultural context, and giving the first turn in the conversation and the five possible responses. We then ask the model to select the response that is most appropriate for that culture.

To determine correctness, we check whether the predicted response falls within the culture-specific range of valid answers, after rounding for direct comparison. For example, if the accepted stylistic range is [1.25,2.67], then predictions of 1, 2, or 3 are considered correct. Accuracy for each country is calculated as the proportion of correct predictions across all conversations.

Unsurprisingly, we find that LLMs perform best at adapting to Western communication norms, with their highest accuracies observed for America and the Netherlands. This imbalance is concerning because LLM systems deployed in non-Western contexts are less likely to align with local users' communication practices.

5 Conclusion

The framework and dataset presented in this paper strive to bridge the gap between cultural psychology and generative AI. Our work can be used to evaluate LLMs, inform conversational agents, and ultimately work towards models that are culturally competent and adaptive.

This is especially important for building downstream systems, like chatbots, where context matters tremendously: the norms of appropriate communication differ sharply depending on whether a chatbot is deployed in a workplace, designed to tutor students, or intended to support individuals overcoming personal struggles. As a result, these systems need to adapt their understanding of social norms (Rai et al., 2025), implied language (Havaldar et al., 2025a), and linguistic style (Kang and Hovy, 2021).

More broadly, *LLM systems that interact with diverse users operate not only within a cultural context but also within a situational and interpersonal context* — the notion of "appropriate behavior" emerges from the interaction of all three. By formalizing these dimensions, our framework offers a path toward developing AI systems that better understand, respect, and adapt to diversity in communication.

Limitations

A large limitation of our work is that we create a fully English dataset. While it is crucial to evaluate LLMs in all languages, we made the decision to create an English dataset for the following reasons:

1. People from a wide variety of cultures engage with LLMs in English, as LLMs have higher QA skills, robustness to prompt ablations, and reasoning capabilities in English.

2. The conversation generation component took many rounds of prompt engineering, as it was a nuanced and complex task; this was only possible for the authors to do in English.

3. The authors manually validated and edited all generated conversations. Once again, this was only possible for the authors to do in English.

Additionally, our dataset itself is small, consisting of 48 conversations with 240 total possible responses. This was by design; many cultural benchmarks that exist are massive, LLM-generated corpora with human validation on only a small subset of the data — benchmarks from Shi et al. (2024); Fung et al. (2024), and many others as surveyed by

Zhou et al. (2025). We aim to create a high-quality dataset that is fully human-validated.

We also conducted a smaller-scale annotation study, with only 3 annotators per country. We were limited by the availability of participants on Prolific; our 8 chosen countries reflect areas with high concentrations of Prolific users. To get a better measure of accepted style, which includes underrepresented cultures as well, future work should involve a larger-scale study.

6 Ethical Considerations

In this work, we simplify the notion of "culturally-aware communication" to having an appropriate linguistic style; however, communication practices in every culture are complex, dynamic, and consist of many dimensions beyond linguistic style.

This work involves LLM usage at two stages in our pipeline — scenario generation and conversation generation. Though the authors manually validated every generated conversation, any inherent bias in or fairness concerns associated with the LLM may propagate into our generated dataset.

Lastly, we use nationality and language as a proxy for culture — while these three things are heavily intertwined, culture is dynamic and subjective and does not perfectly align with either nationality or language.

References

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Douglas Biber and Susan Conrad. 2009. Register, genre, and style.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *Preprint*, arXiv:2410.02677.

Nikolas Coupland. 2007. *Style: Language variation and identity*. Cambridge University Press.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *Preprint*, arXiv:2402.09369.

Emi Furukawa, June Tangney, and Fumiko Higashibara. 2012. Cross-cultural continuities and discontinuities in shame, guilt, and pride: A study of children residing in japan, korea and the usa. *Self and Identity*, 11(1):90–113.

Salvatore Giorgi, Shreya Havaldar, Farhan Ahmed, Zuhaib Akhtar, Shalaka Vaidya, Gary Pan, Lyle H. Ungar, H. Andrew Schwartz, and Joao Sedoc. 2023. Psychological metrics for dialog system evaluation. *Preprint*, arXiv:2305.14757.

Shreya Havaldar, Hamidreza Alvari, John Palowitch, Mohammad Javad Hosseini, Senaka Buthpitiya, and Alex Fabrikant. 2025a. Entailed between the lines: Incorporating implication into NLI. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32274–32290, Vienna, Austria. Association for Computational Linguistics.

Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. 2024. Building knowledge-guided lexica to model cultural variation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 211–226, Mexico City, Mexico. Association for Computational Linguistics.

Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023a. Comparing styles across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6791, Singapore. Association for Computational Linguistics.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023b. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. 2025b. Towards style alignment in cross-cultural translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32213–32230, Vienna, Austria. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders

Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Geert Hofstede. 1986. Cultural differences in teaching and learning. *International Journal of intercultural relations*, 10(3):301–320.

Thomas Holtgraves. 1997. Styles of language use: Individual and cultural variability in conversational indirectness. *Journal of personality and social psychology*, 73(3):624.

Dongyeop Kang and Eduard Hovy. 2021. Style is not a single variable: Case studies for cross-stylistic language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2025. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *Preprint*, arXiv:2406.14805.

Yoshiko Matsumoto. 1988. Reexamination of the universality of face: Politeness phenomena in japanese. *Journal of pragmatics*, 12(4):403–426.

Leela Mullaiti. 1995. Families in india: Beliefs and realities. *Journal of Comparative family studies*, 26(1):11–25.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Sunny Rai, Khushang Zaveri, Shreya Havaldar, Soumna Nema, Lyle Ungar, and Sharath Chandra Guntuku. 2025. Social norms in cinema: A cross-cultural analysis of shame, pride and prejudice. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11396–11415.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *Preprint*, arXiv:2404.15238.

Jessica L Tracy and Richard W Robins. 2008. The nonverbal expression of pride: evidence for cross-cultural recognition. *Journal of personality and social psychology*, 94(3):516.

Jan M Ulijn and Kirk St Amant. 2000. Mutual intercultural perception: How does it affect technical communication?—some data from china, the netherlands, germany, france, and italy. *Technical communication*, 47(2):220–237.

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. *arXiv preprint arXiv:2502.12057*.

A Cultural Matching Annotation: Additional Details

Annotator recruitment. We first recruited 8 volunteers from American, Indian, Chinese, and Korean backgrounds at the authors' university. To annotate the remainder of the dataset, we use the nationality screener on Prolific to select relevant annotators.

Before beginning the study, Prolific annotators are asked to describe their cultural background and state the culture they are most familiar with. We ensure this matches their nationality in the Prolific database to confirm their qualifications.

Country	Recruited Annotators
America	3 volunteers
Netherlands	3 Prolific users
Mexico	3 Prolific users
India	1 volunteer, 2 Prolific users
China	2 volunteers, 1 Prolific user
Japan	3 Prolific users
Korea	2 volunteers, 1 Prolific user
Nigeria	3 Prolific users

Table A1: Annotator breakdown for every country in our dataset. We use 8 volunteers and 16 Prolific users.

The annotators are all given a Google Sheet containing the conversations and a drop-down menu for each row, allowing them to select one of the responses. They were shown the following instructions before beginning the study:

```
Welcome! In this study, you will be asked
to select the most culturally-appropriate
response in a conversation. The situation
column describes an interaction between two
individuals. The initial statement begins the
conversation. The 5 possible responses convey
the same idea, but are stylistically different.
Your task is to consider the cultural dynamics
of the culture you grew up in, and select what
would be the most stylistically appropriate
response for your culture.
```

We also collect all annotators' ages and genders. Annotators were paid \$20/hr and, on average, took 42 minutes to complete the annotation study.

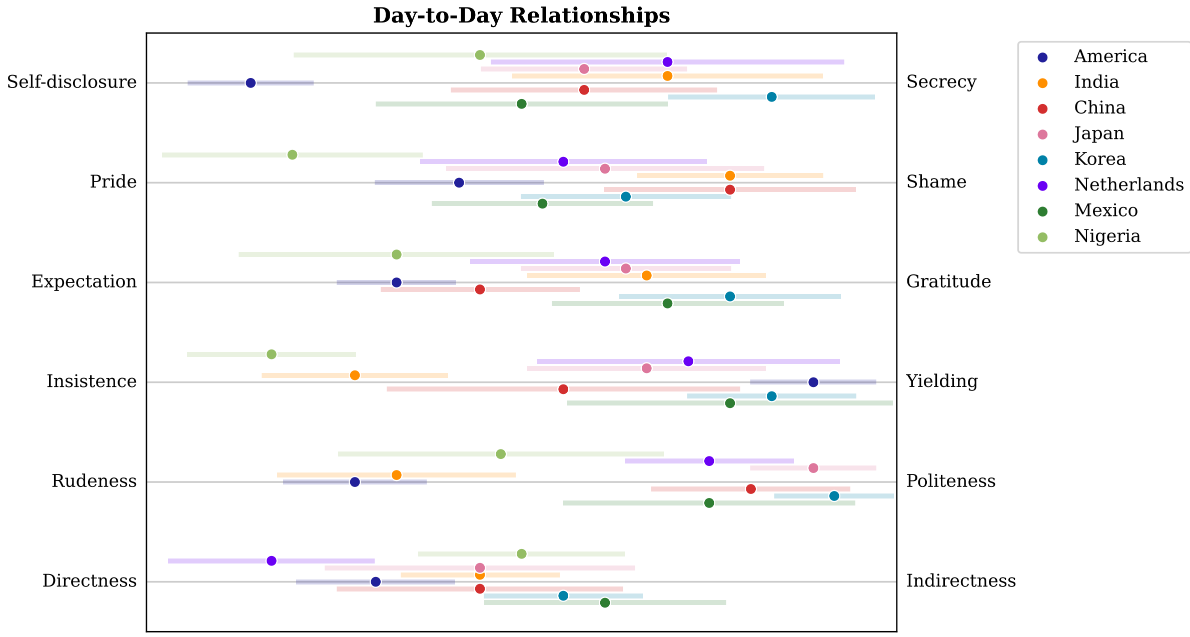


Figure A1: Cultural differences in day-to-day conversations. We show the mean and accepted range of style values for conversations with strangers, neighbors, and friends.

B Model Evaluation: Additional Details

For the models shown in Table 2, the default temperature was used. All models were evaluated identically using the following prompt:

```
You are an expert in intercultural
communication. Given a country, a social
situation, a pair of characters, and the
first turn in a conversation, your task is to
select the response that best reflects the
cultural and stylistic norms and communication
practices of the specified country.

Country: {country}
Situation: {situation}
Characters: {characters}
First turn: {first turn}

Possible responses:
1. {response 1}
2. {response 2}
3. {response 3}
4. {response 4}
5. {response 5}

Your answer should be the number (1-5)
corresponding to the response that best fits
the cultural context. Generate only the number
as your response, without any additional text
or explanation.
```

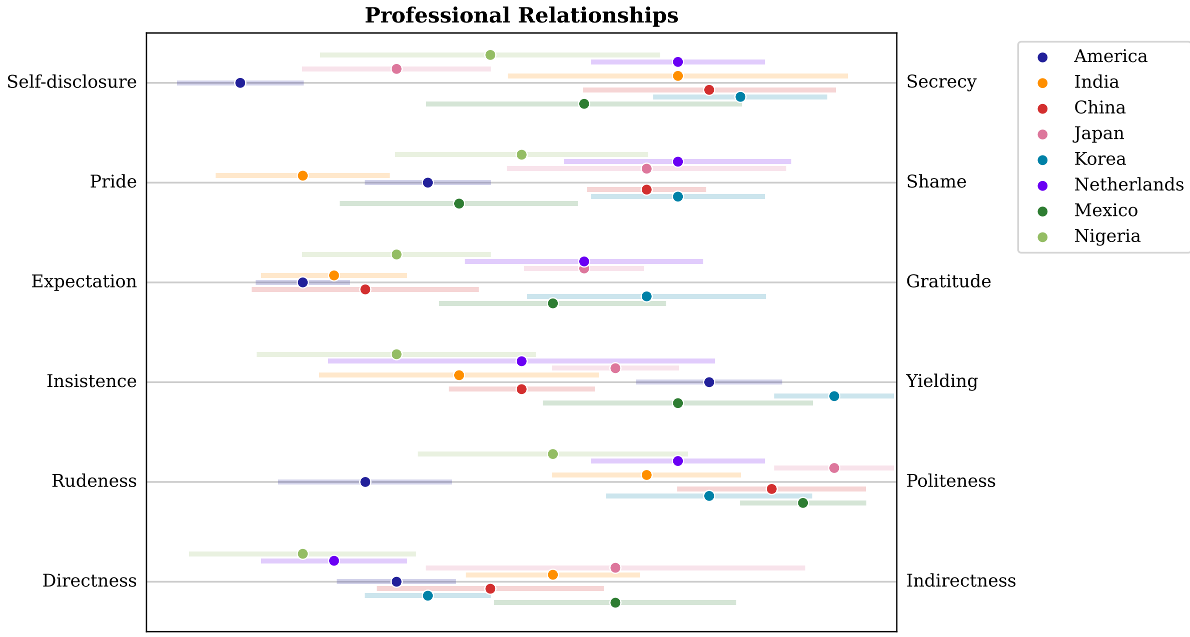


Figure A2: Cultural differences in professional conversations. We show the mean and accepted range of style values for conversations between a boss/employee and coworkers.

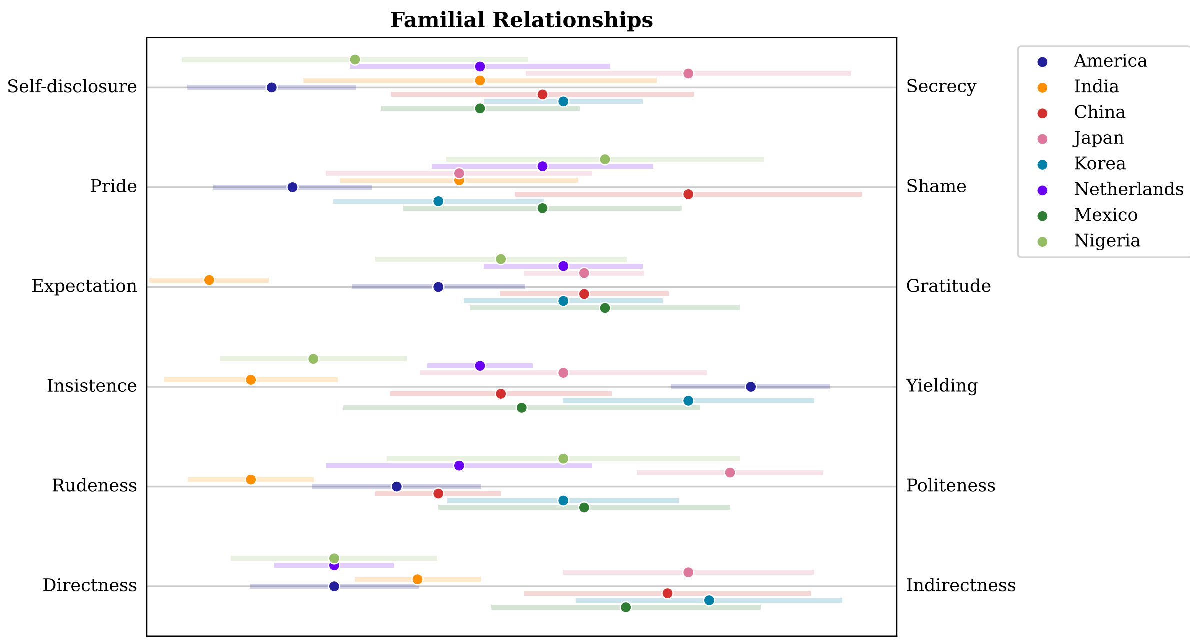


Figure A3: Cultural differences in familial conversations. We show the mean and accepted range of style values for conversations between a husband/wife, parent/child, and grandparent/grandchild.

Directness / Indirectness: Over the fence, Neighbor A tells Neighbor B that the roof of their shared boundary house got damaged in last night's storm.

NEIGHBOR B: Hey, did you notice anything odd about the roof after the storm?

NEIGHBOR A: Our shared roof was damaged by last night's storm.
NEIGHBOR A: The storm wrecked our roof along the boundary last night.
NEIGHBOR A: It appears that last night's storm damaged our shared roof.
NEIGHBOR A: It seems the storm may have affected our roof.
NEIGHBOR A: I'm not certain, but I've heard our roof might be affected.

Politeness / Rudeness: After a meeting, the boss says the employee's slide deck is too text-heavy and lacks clear visuals.

EMPLOYEE: Could you share your thoughts on my slide deck?

BOSS: I feel your slide deck might benefit from fewer text and clearer visuals, if possible.
BOSS: Your slide deck could be improved by reducing text and adding clearer visuals, if you don't mind.
BOSS: The slide deck is too heavy on text and lacks clear visuals, and it needs some revisions.
BOSS: Your slide deck is overloaded with text and misses clear visuals, so it really needs a redo.
BOSS: Your slide deck is horribly text-cluttered and utterly lacks clear visuals; revise it right away.

Insistence / Yielding: At the park, the grandparent offers a homemade apple pie to the child, who excitedly accepts a slice.

CHILD: Grandpa, that pie smells good!

GRANDPARENT: You must take a slice of it right now!
GRANDPARENT: Then you should have a slice of it right here!
GRANDPARENT: Would you perhaps enjoy a slice, dear?
GRANDPARENT: If you wish, you can try a slice.
GRANDPARENT: You may have a slice if you'd like.

Gratitude / Expectation: After dinner, Friend A asks Friend B to pick up some groceries on the way home.

FRIEND B: Should I stop anywhere on the way home tonight?

FRIEND A: I would really appreciate it if you could pick up some groceries on your way home.
FRIEND A: It would be great if you could pick up some groceries on your way home.
FRIEND A: Please pick up some groceries on your way home.
FRIEND A: Make sure you pick up some groceries on your way home.
FRIEND A: You need to pick up some groceries on your way home.

Pride / Shame: During dinner, the parent recalls a career setback that ultimately led to learning resilience.

CHILD: Does anything good come from career struggles?

PARENT: Yes, when I conquered a major career setback, it helped build my strong sense resilience.
PARENT: I overcame a career setback, and that helped me build resilience.
PARENT: I experienced a career setback that ultimately helped me develop resilience.
PARENT: I went through a career setback that forced me to learn humility and resilience.
PARENT: I suffered a career setback that quietly taught me the hard lesson of resilience.

Self-disclosure / Secrecy: During breakfast, the husband gently shares that his work stress is affecting his mood and worries about their future.

WIFE: Has work been bothering you lately, honey?

HUSBAND: I feel overwhelmingly stressed and I am really scared about our future.
HUSBAND: Work has been affecting me and I have concerns about our future.
HUSBAND: I feel a little stressed and I'm worried about what lies ahead for us.
HUSBAND: Work has been more challenging than usual but I'm keeping my worries to myself.
HUSBAND: I'm managing work stress, there's nothing serious going on.

Table A2: Example conversations from our CAC dataset. We show one example for each stylistic axis.

From Regulation to Interaction: Expert Views on Aligning Explainable AI with the EU AI Act

Mahdi Dhaini, Lukas Ondrus, Gjergji Kasneci
Technical University of Munich, Germany
School of Computation, Information and Technology
Department of Computer Science
Munich, Germany
{firstname.lastname}@tum.de

Abstract

Explainable AI (XAI) aims to support people who interact with high-stakes, AI-driven decisions, and the EU AI Act requires that users can appropriately interpret high-risk AI system outputs (Article 13) and that human oversight prevents undue reliance (Article 14). Yet the Act offers little technical guidance on implementing explainability, leaving interpretability methods difficult to operationalize and compliance obligations unclear. To address these gaps, we interviewed eight domain experts across legal, compliance, and technical roles to explore (1) how explainability is defined and perceived under the Act, (2) the practical and regulatory obstacles to XAI implementation, and (3) recommended solutions and future directions. Our findings reveal that domain experts view explainability as context- and audience-dependent, face challenges from regulatory vagueness and technical trade-offs, and advocate for domain-specific rules, hybrid methods, and user-centered explanations. These insights provide a basis for a potential framework to align XAI methods—particularly for AI and Natural Language Processing (NLP) systems—with regulatory requirements, and suggest actionable steps for policymakers and practitioners

1 Introduction and Background

With the increasing deployment of large AI models (e.g., pretrained language models and large language models (LLMs)) in high-stakes domains, their inherent "black-box" nature offers limited transparency into decision-making processes, posing significant risks. Consequently, regulations such as the GDPR and the EU AI Act require both transparency and explainability. This has paved the way for the development of extensive research in the **Explainable AI (XAI)** field. Research in XAI has been increasing, with numerous surveys focused on critical domains such as healthcare, finance, and law (Chaddad et al., 2023; Richmond et al., 2023; Yeo et al., 2025). Moreover, information and knowledge management systems are increasingly integrating explainable AI methods to support trustworthy decision-making in different domains (Rožanec et al., 2022; Mancuso et al., 2025; Majumder and Dey, 2022; Chen et al., 2024; Brasse et al., 2023)

The **EU AI Act** (Commission, 2021) mandates transparency for high-risk AI systems via: *Article 13(1)*: systems must enable users to interpret outputs appropriately and *Article 14(4)*: human oversight must prevent undue reliance on AI decisions. However, the Act provides no technical guidance for implementing explainability methods (Panigutti et al., 2023; Gyevnar et al., 2023). Existing XAI techniques focus on algorithmic transparency but often remain opaque to non-experts and difficult to operationalize for intended users(Panigutti et al., 2023; Golpayegani et al., 2023). Moreover, the high-risk classification (Art. 6(2)) is ambiguous, creating uncertainty about compliance obligations (Golpayegani et al., 2023; Nisevic et al., 2024). Consequently, many XAI methods offer mathematical insight without satisfying legal standards for human interpretability (Panigutti et al., 2023; Kusche, 2024).

Previous research shows that no standardized framework exists to assess whether XAI methods satisfy the EU AI Act's transparency mandates (Sovrano et al., 2022; Panigutti et al., 2023). Second, legal expectations for explainability remain ambiguous, with no consensus on what qualifies as an acceptable explanation (Panigutti et al., 2023; Gyevnar et al., 2023). Third, research rarely integrates explainability techniques with compliance mechanisms, as most studies address XAI methods and regulatory frameworks in isolation (Hacker, 2023; Nisevic et al., 2024).

Despite extensive XAI research, little work has engaged experts on how current methods align with the EU AI Act's transparency and explainability

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 230–239
November 9, 2025

requirements. With key compliance deadlines approaching (Timeline, 2024), understanding practitioners' views on the feasibility and challenges of meeting these mandates is critical yet largely unexplored. Aiming to address these gaps, we conducted an interview study with 8 domain experts from Europe, addressing the following Research Questions (RQs):

- ***RQ1***: *How do domain experts define and perceive explainability in the context of the EU AI Act's transparency requirements, and what usability and technical factors shape these perceptions?*
- **RQ2**: *What practical and regulatory obstacles do practitioners face when implementing XAI to comply with the EU AI Act?*
- ***RQ3***: *What solution concepts and future directions do domain experts recommend for aligning XAI methods with regulatory requirements?*

This work investigates domain experts' perspectives on the intersection of XAI and the EU AI Act, focusing on perceptions and the challenges of implementing XAI to achieve compliance. This paper provides an empirical perspective grounded in expert interviews and offers evidence-based and actionable insights for aligning XAI with regulatory demands. The remainder of the paper is structured as follows: Section 2 introduces the research methodology; Section 3 presents the results of our study; Section 4 reviews related work in the literature. Finally, Section 5 discusses our findings, presents our conclusions, and outlines future work and study limitations.

2 Research Methodology

We followed a qualitative research approach and drew on the experiences of experts by conducting semi-structured interviews (SSI), which offer flexibility while retaining a structure for data collecting and enable a full analysis of participants' experiences (DiCicco-Bloom and Crabtree, 2006). To examine the potential gaps and trends in the literature, we first conducted an initial literature review (Fresz et al., 2025; Freiesleben and König, 2023; Rožanec et al., 2022; Mancuso et al., 2025; Majumder and Dey, 2022; Chen et al., 2024; Brasse et al., 2023) to build our knowledge base and as the foundation for the questionnaire. Then, we interviewed experts and professionals to collect input on the intersection between XAI and the EU AI Act, focusing on understanding the context, interventions, mechanisms, and outcomes of XAI implementations while addressing regulatory compliance and practical challenges. We describe below the study design, data collection, and data analysis of our research.

2.1 Study Design

Drawing on the five-step process of Kallio et al. (2016), we developed our semi-structured *interview guide* as follows: *(1) Prerequisites:* Ensured that researchers possessed sufficient background in XAI, law, and policy to identify key topics and appreciate multiple stakeholder perspectives. *(2) Literature review:* Conducted a literature review to build a conceptual framework, focus on AI Act compliance issues, and identify knowledge gaps. *(3) Drafting:* Created an initial question set combining core concepts and flexible prompts to explore experts' views on XAI under the EU AI Act. *(4) Pilot testing:* Refined wording, structure, and flow through three internal iterations and a final supervisor review. *(5) Finalization:* Documented the complete guide to ensure replicability for future studies.

We reached out to experts and professionals in the fields of law, policy, and AI development, as the EU AI Act is a complex regulatory instrument to ensure domain diversity in our interviews. We assumed expertise when the participants had 4 years or more of experience in the aforementioned fields. Recruitment occurred in two waves: an initial phase (December 2024) via personal contacts, third-party introductions, and online searches, followed by a snowball phase (January 2025) based on interviewee recommendations. In total, we contacted 50 candidates between December 2024 and January 2025 via email, LinkedIn, and phone. Among the 50 candidates, 8 expressed interest in participating in an interview. We present demographic information on our participants in Table 1. All participants we interviewed were based in Europe, reflecting our focus on those subject to the EU AI Act. Except for P1, a government official, all worked in the private sector. Each had at least five years' experience in law, policy, or AI development.

Participant	Position	Expertise	Experience	Organization	Size
P1	Researcher	Legal Expert, Regulatory Compliance	10	University	Large
P2	Senior Data Engineer	Data Platforms and Governance	7	Consulting Firm	Large
P3	AI Consultant	Regulatory Technology	20	Consulting Firm	Large
P4	CEO	AI Education and Training	15	AI Company	Small
P5	Account Executive	Enterprise Data & AI Solutions	10	AI & Software Company	Large
P6	Chief Research Scientist	NLP and Explainable AI Research	10	AI Research Firm	Medium
P7	AI Governance Consultant	Strategic Innovation Management & AI Governance	5	Telecommunication Company	Large
P8	Research Scientist	NLP and Explainable AI	13	AI Research Firm	Medium

Table 1: Overview of Interviewee Demographics. Experience is measured in years.

2.2 Data Collection

We conducted semi-structured interviews via Microsoft Teams between December 2024 and March 2025. To achieve observer triangulation (Runeson and Höst, 2009), two researchers attended each session. At the outset, participants were briefed on recording procedures, anonymization protocols, and the intended use of their transcripts. The research objectives and interview structure were reiterated to ensure understanding. Although the sequence and wording of questions remained fixed, the semi-structured format permitted slight modifications in response to conversational flow. Interview prompts were drawn from themes identified during the preliminary literature review and organized into five thematic sections: (1) *Background and common basis* which included collecting foundation data about the participants like their professional background (2) *Explainability perceptions* where participants were asked on the definition of explainability and how they perceive its usability and limitation (3) *XAI implementation challenges* and (4) *Regulatory gaps and compliance risks* and (5) *Future directions and solution concepts* and finally (6) *Closing* where we invited participants to share any additional insights and to indicate their willingness for follow-up contact or to recommend other participants. To ensure reproducibility and transparency, we include the full interview questionnaire in Appendix A.

2.3 Data Analysis

The interviews were recorded and transcribed with the help of *otter.ai* with the participant's consent. We omit the full interview transcripts as the participants requested anonymity. We then systematically coded the transcriptions following the thematic approach described by (Braun and Clarke, 2006) as follows: The following guidelines were applied and adapted to our data:

(1) Familiarization with extracted data: we read and re-read the interview transcripts to identify relevant data and initial patterns. Extracted data were organized in Excel according to the questions in the interview guide.

(2) Generating initial codes: we systematically analyzed each transcript, assigning concise labels (*codes*) to meaningful text segments. Codes were entered in a dedicated Excel column to facilitate topic filtering.

(3) Searching for themes: related codes were grouped into broader categories (*themes*) that reflect the research questions. We added a "theme" column in Excel to cluster codes accordingly.

(4) Reviewing themes: we checked each provisional theme against the full dataset to ensure it accurately represented the underlying patterns and refined them for coherence.

(5) Defining and Naming Themes: final themes were precisely defined and named. Each theme's description was recorded in Excel along with its linkage to the corresponding research question.

(6) Synthesizing findings: we synthesized and presented the thematic findings using illustrative data extracts to construct a coherent narrative of our findings (Braun and Clarke, 2006).

3 Results

In this section, we present the results from our SSI study, we organize the results according to the thematic coding process explained earlier. Table 2 presents the final themes developed to address our research questions, derived from initial coding of participant responses and subsequent thematic grouping. In the next subsections, we present results of each of the thematic sections.

Main Theme	Description	Codes
Explainability Perceptions (RQ1)	How experts define explainability and perceive its usability and limitations.	Lack of established definition User-specific definitions Technical complexity
Implementation Challenges (RQ2)	Difficulties in applying XAI methods in practice.	Regulatory vagueness Accuracy trade-offs Technical expertise required
Regulatory Gaps and Compliance Risks (RQ2)	Regulatory insights, compliance challenges, and possible mitigations.	Lack of specific guidance Superficial compliance risks Lack of regulatory sandboxes
Future Directions and Solution Concepts (RQ3)	Expert recommendations on advancing XAI methodologies and regulations.	Hybrid approaches User-centered explanations Domain-specific regulations

Table 2: Key themes, descriptions, and codes relevant to the research questions

3.1 Perceptions of Explainability (RQ1)

Table 3 presents some representative answers from participants regarding their input on their perceptions of explainability. Participants emphasized the absence of a **standardized definition of explainability** and related XAI terms, noting the importance of a consistent framework. Most agreed that explainability must be context-dependent and tailored to various stakeholders: "explainability must vary based on the intended user and application domain." (P1). P3 reinforced this point, stating, "Every stakeholder understands explainability differently." Concurrently, P2 mentioned that "current methods are not easily understood by non-expert users," highlighting the inherent **technical complexity**. Together, these findings indicate a clear consensus on the need for **context-dependent and user-specific explainability**. They also reveal a substantial gap between existing theoretical constructs and practical implementation.

3.2 Implementation and Compliance Challenges (RQ2)

To answer RQ2, we asked the participants about their challenges and barriers to implementing the XAI method in practice, mainly for compliance with the EU AI Act. We present in Table 4 some representative answers from participants. Participants reported multiple challenges in implementing XAI methods. First, **regulatory vagueness** was frequently highlighted. P2 and P8 criticized the significant difficulties arising from unclear definitions of compliance under the EU AI Act, suggesting that the lack of concrete guidance hampers practical implementation. P2 stated, "Clear guidance from the EU AI Act is missing, making practical compliance difficult to achieve." P8 echoed this concern: "Clearer compliance definitions are needed to effectively operationalize these requirements." Second, **accuracy trade-offs** emerged as a prominent barrier. P4 and P6 noted an inherent conflict between interpretability and model accuracy. As P4 remarked, "When we aim for transparent explanations, predictive accuracy often suffers."

A lack of general **technical expertise** within organizations was identified as a significant hurdle. P1, P3, and P7 criticized many organizations for lacking dedicated AI teams or trained specialists in explainability and transparency P1 particularly highlighted the "the complexity of explainability tools is often underestimated" and how "teams lack the technical skills required to implement them effectively"(P1).

Conclusion on implementation challenges: unclear regulations, the trade-off between transparency and performance, and insufficient technical capacity collectively obstruct the adoption of XAI in practice for sake of compliance.

3.3 Regulatory Gaps and Compliance Risks (RQ2)

To further answer RQ2, we collected from interviewees their input on the gaps in regulations and potential risks of implementing XAI under the current AI Act from compliance, and also some possible mitigations to these challenges. In Table 5 we present results. The greatest concern regarding the AI Act was a lack of specific guidance. They expressed concern about the potential for superficial or even ineffective compliance strategies. Interview results show repeated emphasis from participants on the absence of clear standards, where "Explicit standards for compliance would greatly improve transparency efforts" (P3). P5 emphasized that "The Act needs clearer standards for transparency compliance", and P8 highlighted how a "Clearer

Code	Representative participants answers
User-specific explanations	"You need tailored explanations for different user groups." (P5) "Explanations must fit the cognitive abilities of the users." (P7)
Technical complexity	"Common explainability tools like SHAP are too technical for most end-users." (P4)
Lack of established definition	"The industry lacks a standard definition for explainability." (P6) "There's still no universally agreed definition of explainability." (P8)

Table 3: Representative answers on explainability perceptions

Code	Representative participants answers
Technical expertise	"The complexity of explainability tools is often underestimated. Teams lack the technical skills required to implement them effectively." (P1) "Few companies actually have the technical expertise in-house to leverage advanced XAI methods." (P3) "Organizations without specialized AI teams find it especially difficult to implement these tools correctly." (P7)
Regulatory vagueness	"Clear guidance from the EU AI Act is missing, making practical compliance difficult to achieve." (P2) "The EU AI Act does not specify what exactly counts as transparent or interpretable." (P5) "Clearer compliance definitions are needed to effectively operationalize transparency." (P8)
Accuracy trade-offs	"When we increase interpretability, we lose a lot in accuracy. It's a fundamental trade-off." (P4) "Achieving full transparency in AI means often sacrificing predictive accuracy." (P6)

Table 4: Representative answers on implementation challenges

guidance from the EU would help companies implement meaningful XAI." These results point to gaps in the AI Act in terms of an overarching lack of specific and operational guidance. Participants also highlighted some **risks** that could result from enforcing compliance where in the absence of precise definitions, organizations may adopt "superficial explanations just to meet regulations" (P1) or "easy but ineffective solutions to meet vague legal requirements" (P4), our study results indicate participants worry that companies may end up merely *"checking the box"* rather than genuinely implementing/adopting explainability to achieve superficial compliance with the regulation which stem from the regulatory vagueness discussed before. This also indicates the need to come up with and implement effective checkpoints to prevent this.

As for potential **mitigation strategies**, some participants emphasized on the need and value of *regulatory sandboxes*, as "regulatory sandboxes allow realistic testing of compliance strategies safely" (P1), and P3 further stressed that "explicit standards for compliance would greatly improve transparency efforts", noting the lack of specificity in current regulations. Overall, results reflect a clear demand from experts and professionals for more concrete guidance on explainability from regulators.

3.4 Future Directions for Compliance (RQ3)

We also asked participants about future directions for strategies and recommendations from an XAI and regulatory perspective to facilitate meeting regulatory requirements in a practical way. Table 6 presents a subset of participants' answers on this topic.

Domain-specific and tailored regulations were frequently mentioned, mainly as "tailored regulations for different sectors could improve practical compliance" (P5). P1 emphasized that "explainability should be regulated with sensitivity to specific industry needs". These results suggest that a *one-size-fits-all approach* is not feasible, as transparency requirements vary across AI application domains. **Hybrid approaches** also emerged as a main theme where participants emphasized hybrid methods could balance interpretability with model performance. P2 emphasized, "A blend of accuracy and interpretability is crucial", and P6

Code	Representative participants answers
Superficial compliance risks	"Companies might adopt easy but ineffective solutions to meet vague legal requirements." (P4) "Without clarity, organizations might adopt superficial solutions just for compliance." (P7)
Lack of regulatory sandboxes	"Regulatory sandboxes allow realistic testing of compliance strategies safely." (P2) "Testing in safe regulatory sandboxes helps clarify vague regulatory requirements." (P6)
Lack of specific guidance	"The Act needs clearer standards for transparency compliance." (P5)

Table 5: Representative answers on regulatory implications

Code	Representative participants answers
Domain-specific regulations	"Explainability should be regulated with sensitivity to specific industry needs." (P1)
Hybrid approaches	"Hybrid methods would allow balancing accuracy and interpretability effectively." (P7)
User-centered explanations	"Research should focus more explicitly on the usability of explanations." (P8)

Table 6: Representative answers on future directions

added, “Integrating technical detail and practical interpretability is essential,” highlighting the need to combine complementary explanation techniques. The final theme concerned **user-centered explanations**, “understanding user needs should be the priority of future research” (P3), a statement which was also echoed by P4, who remarked that “user-oriented research is crucial to developing meaningful explanations.” participants emphasized that XAI solutions must be tailored to stakeholders’ abilities and contexts to ensure explanations are both usable and informative.

These insights reflect evolving perspectives on the interaction between XAI development and the EU AI Act, illustrating directions for tailoring explainability methods and clarifying regulation to achieve explanations that are both user-aligned and compliant.

4 Related Work

Recent studies have examined the intersection of XAI and the EU AI Act, highlighting some regulatory and technical challenges. Panigutti et al. (2023) conducted an interdisciplinary analysis of how the Act addressed black-box AI systems, reviewing existing XAI techniques and noting persistent limitations in meeting the Act’s objectives. Using a case study on AI-based exam proctoring, they illustrated how transparency and human oversight could be implemented even when decisions were not inherently interpretable. Similarly, Fresz et al. (2025) analysed explainability requirements under European and German law, finding that different legal frameworks demanded distinct XAI properties and that current techniques often failed to meet expectations, particularly regarding fidelity and confidence estimates. These shortcomings reflected technical trade-offs, as many post-hoc methods lacked faithfulness guarantees, posing compliance challenges. Pavlidis (2024) explored various approaches to advance XAI and discussed the difficulties of embedding explainability into governance and policy, with attention to standard-setting, oversight, and enforcement. Hacker and Passoth (2022) further emphasized the context-dependent nature of explainability, noting that the appropriate form and degree of transparency depended on specific circumstances. They also observed that the Act’s obligations were largely aimed at professional users of high-risk AI systems rather than the general public, which could result in transparency measures that overwhelm non-expert users with overly technical details.

While prior studies have examined the EU AI Act’s transparency provisions from legal, policy, or interdisciplinary perspectives (Panigutti et al., 2023; Fresz et al., 2025; Pavlidis, 2024; Hacker and Passoth, 2022), they have primarily focused on conceptual analyses, legal interpretations, or technical reviews of existing XAI methods. While they provide valuable insights, they don’t systematically engage with practitioners and domain experts to understand how explainability is interpreted in real-world settings, nor how the Act's high-level mandates translate into concrete implementation practices. Our study fills this gap by using semi-structured interviews with eight experts in high-stakes AI to examine how they interpret explainability under the EU AI Act, the factors influencing their views, and the practical barriers to compliance, revealing nuances often overlooked in legal or technical analyses. Our work contributes *actionable insights by synthesizing experts' recommendations into potential solution concepts.*

5 Discussion and Conclusion

Our interview study findings with eight experts reveal that explainability is fundamentally *context- and user-specific*, challenging the EU AI Act's broad transparency mandates. In practice, organizations struggle with three main obstacles: (1) **regulatory vagueness**, which leaves compliance criteria undefined, (2) the inherent **accuracy–interpretability trade-off**, forcing a choice between performance and clarity, and (3) limited **technical expertise**, which delays effective integration of XAI tools. Together, these factors could lead to superficial "box-checking" rather than genuine explainability.

Experts propose a threefold strategy to bridge these gaps. (1) *domain-specific regulations* would adapt transparency requirements to each sector's risk profile. (2) *hybrid approaches* combining high-accuracy models with simpler explainers can reconcile predictive performance with comprehensibility. (3) *user-centered design* must guide both the development and evaluation of explanations, ensuring they meet stakeholders' cognitive and operational needs.

Our findings have several **implications** for different *stakeholders*. *Researchers* are encouraged to develop evaluation frameworks that jointly assess legal compliance and end-user comprehensibility. *Policymakers and regulators* should consider refining the EU AI Act by adding sector-tailored criteria and establishing regulatory sandboxes for safe testing. On the other hand, *practitioners* can use these results as a basis to invest in interpretability training and adopt flexible XAI pipelines. By synthesizing expert insights on definitions, challenges, and solution concepts, this work lays the groundwork for aligning XAI methods with regulatory requirements and advancing trustworthy AI in high-risk domains. Beyond regulatory alignment, our findings have implications for the design of explanation interfaces and interaction modalities, particularly those relying on natural language (Atanasova et al., 2023; Madsen et al., 2024, 2022). Incorporating user-centered design principles into explanation generation (Mishra et al., 2024; Wang et al., 2025) and presentation can help bridge the gap between legal compliance and meaningful human understanding, an intersection of growing interest in human-computer interaction and NLP research.

Unlike prior work, which has mainly offered legal or conceptual analyses of the EU AI Act's explainability requirements, our study provides an empirical perspective grounded in expert interviews. By capturing how practitioners interpret explainability, the challenges they face in achieving compliance, and their proposed solutions, we offer evidence-based, actionable insights for aligning XAI methods with regulatory demands.

We present this study as work in progress and hope it catalyzes dialogue between HCI and XAI/NLP researchers, laying a foundation for further work on explainability under regulatory mandates. With this study, we aim to advance efforts to bridge the gap between human- and regulator-oriented explainability and technical explainability in AI/NLP.

Future Work In our current sample, interviewees are primarily based in Europe, reflecting the particular relevance of the EU AI Act in this region. As future work, we plan to conduct additional interviews with experts from both European and non-European countries to expand the sample and enable comparative analyses between EU and non-EU perspectives. A complementary direction is to field a large-scale survey to validate and generalize our qualitative findings, quantify key insights, and broaden their applicability. Together, these extensions aim to support more concrete recommendations and implications for the design of explanation systems, including natural language-based approaches.

Limitations This study's modest sample size, while comparable with relevant research (Warren et al., 2025), may limit generalizability beyond the examined contexts. However, thematic saturation appeared to be reached, as no new themes emerged in later interviews. All interviews were conducted in English, and although the sample included participants from diverse linguistic backgrounds, conducting them in additional languages could have broadened participation. We invite future research to build on and validate these findings through larger, more diverse samples. While the interviews were designed to follow a consistent and impartial approach, and reflexivity was actively maintained throughout the analysis, the findings may still carry the inherent biases and limitations of self-reported data (Donaldson and Grant-Vallone, 2002). In addition, one limitation of these types of works (e.g., prior related work we discussed and our work) could be the evolving nature of the EU AI Act, driven by the rapid development of AI research and practice. The Act has undergone several key milestones. In our study, we referred to the latest version of the Act available as of December 2024.

Acknowledgment

We thank the anonymous reviewers for their insightful and constructive feedback and the interview participants for their valuable contribution to this study. This research has been supported by the German Federal Ministry of Education and Research (BMBF) grant 01IS23069 Software Campus 3.0 (TU München).

References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Julia Brasse, Hanna Rebecca Broder, Maximilian Förster, Mathias Klier, and Irina Sigler. 2023. Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, 33(1):26.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. 2023. Survey of explainable ai techniques in healthcare. *Sensors (Basel, Switzerland)*, 23(2):634.

Huaming Chen, Jun Zhuang, Yu Yao, Wei Jin, Haohan Wang, Yong Xie, Chi-Hung Chi, and Kim-Kwang Raymond Choo. 2024. Trustworthy and responsible ai for information and knowledge management system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5574–5576, New York, NY, USA. Association for Computing Machinery.

European Commission. 2021. Proposal for a regulation on artificial intelligence. Accessed: 27-February-2024.

Barbara DiCicco-Bloom and Benjamin F Crabtree. 2006. The qualitative research interview. *Medical Education*, 40(4):314–321.

Stewart I Donaldson and Elisa J Grant-Vallone. 2002. Understanding self-report bias in organizational behavior research. *Journal of business and Psychology*, 17:245–260.

Timo Freiesleben and Gunnar König. 2023. Dear xai community, we need to talk! In *Explainable Artificial Intelligence*, page 48–65, Cham. Springer Nature Switzerland.

Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco F. Huber, and Christian Horz. 2025. *How Should AI Decisions Be Explained? Requirements for Explanations from the Perspective of European Law*, page 438–450. AAAI Press.

Delaram Golpayegani, Harshvardhan J. Pandit, and Dave Lewis. 2023. To be high-risk, or not to be—semantic specifications and implications of the ai act's high-risk ai applications and harmonised standards. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 905–915, New York, NY, USA. Association for Computing Machinery.

Balint Gyevnar, Nick Ferguson, and Burkhard Schafer. 2023. Bridging the transparency gap: What can explainable ai learn from the ai act? In *ECAI 2023*, pages 964–971. IOS Press.

Philipp Hacker. 2023. The european ai liability directives – critique of a half-hearted approach and lessons for the future. *Computer Law Security Review*, 51:105871.

Philipp Hacker and Jan-Hendrik Passoth. 2022. *Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond*, pages 343–373. Springer International Publishing, Cham.

Hanna Kallio, Anna-Maija Pietilä, Martin Johnson, and Mari Kangasniemi. 2016. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, 72(12):2954–2965.

Isabel Kusche. 2024. Possible harms of artificial intelligence and the eu ai act: fundamental rights and risk. *Journal of Risk Research*, 0(0):1–14.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

Andreas Madsen and 1 others. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8).

Soumi Majumder and Nilanjan Dey. 2022. *Explainable Artificial Intelligence (XAI) for Knowledge Management (KM)*, page 101–104. Springer, Singapore.

Ilaria Mancuso, Antonio Messeni Petruzzelli, Umberto Panniello, and Federico Frattini. 2025. The role of explainable artificial intelligence (xai) in innovation processes: a knowledge management perspective. *Technology in Society*, 82:102909.

Aditi Mishra, Sajjadur Rahman, Kushan Mitra, Hannah Kim, and Estevam Hruschka. 2024. Characterizing large language models as rationalizers of knowledge-intensive tasks. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8117–8139, Bangkok, Thailand. Association for Computational Linguistics.

Maja Nisevic, Arno Cuypers, and Jan De Bruyne. 2024. Explainable ai: Can the ai act and the gdpr go out for a date? In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, and Emilia Gomez. 2023. The role of explainable ai in the context of the ai act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1139–1150, New York, NY, USA. Association for Computing Machinery.

Georgios Pavlidis. 2024. Unlocking the black box: analysing the eu artificial intelligence act's framework for explainability in ai. *Law, Innovation and Technology*, 16(1):293–308.

Karen McGregor Richmond, Satya M. Muddamsetty, Thomas Gammeltoft-Hansen, Henrik Palmer Olsen, and Thomas B. Moeslund. 2023. Explainable ai and law: An evidential survey. *Digital Society*, 3(1):1.

Jože M. Rožanec, Blaž Fortuna, and Dunja Mladenić. 2022. Knowledge graph-based rich and confidentiality preserving explainable artificial intelligence (xai). *Information Fusion*, 81:91–102.

Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2):131–164.

Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. 2022. Metrics, explainability and the european ai act proposal. *J*, 5(1):126–138.

EU AI Act–Implementation Timeline. 2024. Implementation timeline of the eu artificial intelligence act. Accessed: 27-February-2025.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. 2025. Cross-refine: Improving natural language explanation generation by learning in tandem. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1150–1167, Abu Dhabi, UAE. Association for Computational Linguistics.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. *Show Me the Work: Fact-Checkers' Requirements for Explainable Automated Fact-Checking*. Association for Computing Machinery, New York, NY, USA.

Wei Jie Yeo, Wihan Van Der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, and Gianmarco Mengaldo. 2025. A comprehensive review on financial explainable ai. *Artificial Intelligence Review*, 58(6):189.

A Interview Guide

Disclaimer

Before we begin, I want to inform you that this interview will be recorded for transcription purposes and used solely for a research paper. Your identity will remain confidential; no identifiable information will appear in the final document. External AI tools may be used for transcription. Could you please confirm your consent to these terms?

Introduction

This interview focuses on understanding the role of explainability in AI systems, particularly in the context of the EU AI Act. I will ask about your experiences, opinions, and insights regarding explainable AI, regulatory challenges, and practical tools. I can provide further clarification on any topic during the interview if needed.

Warm-Up Questions

- What is your current role?
- Could you briefly share your background and experience with AI and explainable AI?
- When you hear the word "explainability" in the context of AI, what does it mean to you?
- How familiar are you with the EU AI Act and its provisions related to AI transparency and explainability?

Understanding the EU AI Act

Background Note: The EU AI Act emphasizes requirements like transparency, accountability, and human oversight for high-risk AI systems. Explainability is considered essential to meet these requirements.

- Article 13 addresses transparency and provision of information to deployers. It specifies that high-risk AI systems must be designed to be transparent, where providers of high-risk AI systems must design systems to allow for traceability, ensuring that their operation and outputs are explainable to relevant stakeholders.
- Article 14 emphasizes human oversight and requires systems to provide meaningful information to enable users to understand and intervene appropriately.
- Article 15 addresses accuracy, robustness, and cybersecurity requirements, stating that high-risk AI systems must be designed to be accurate, robust, and secure.

Questions

- Which specific requirements, obligations, or guidelines in the EU AI Act do you think will have the most significant impact on explainable AI practices?
- From your perspective, how does explainability contribute to the EU AI Act's goals of transparency and accountability?
- Are there parts of the EU AI Act's explainability requirements that you find unclear or difficult to interpret? If so, which ones?
- What are the limitations of current explainability tools in meeting oversight expectations? Are current explainability tools sufficient?
- What challenges do you think recent advances in AI systems, such as Large Language Models (LLMs) like ChatGPT, pose to explainability? How might these challenges impact compliance with regulations like the EU AI Act?
- What challenges do you think ensuring explainability presents for AI systems like Large Language Models (LLMs) in meeting the EU AI Act's requirements for transparency and oversight?

Challenges in Implementation

- What challenges have you faced (or foresee) in implementing explainability under the EU AI Act?
- Do you think explainability requirements conflict with other priorities, such as innovation, intellectual property, or system performance? How can this tradeoff be balanced?

Success Factors in Implementation

- What solutions or success factors do you think are critical for addressing these challenges and successfully implementing explainability under the EU AI Act?
- Can you provide examples of strategies, tools, or frameworks that organizations have used effectively to mitigate explainability challenges?

Best Practices

- Have you seen examples of successful explainability practices in your field? What made them successful?

Future Trends

- What trends do you see shaping explainability in the next few years? Do you think they will simplify compliance with the EU AI Act?

Closing Questions

- If you could improve one aspect of the EU AI Act's explainability provisions, what would it be?
- Are there areas of explainability research you think are underexplored but critical for compliance?
- Is there anyone in your professional network you could recommend who might also provide valuable insights for this research?
- Is there anything else you want to share about explainability or the EU AI Act that we haven't covered?

From Noise to Nuance: Enriching Subjective Data Annotation through Qualitative Analysis

Ruyuan Wan[1] **Haonan Wang**[2] **Ting-Hao 'Kenneth' Huang**[1] **Jie Gao**[2]
[1]The Pennsylvania State University, University Park, PA, USA
{rjw6289, txh710}@psu.edu
[2]Johns Hopkins University, Baltimore, MD, USA
{hwang298, jgao77}@jh.edu

Abstract

Subjective data annotation (SDA) plays an important role in many NLP tasks, including sentiment analysis, toxicity detection, and bias identification. Conventional SDA often treats annotator disagreement as noise, overlooking its potential to reveal deeper insights. In contrast, qualitative data analysis (QDA) explicitly engages with diverse positionalities and treats disagreement as a meaningful source of knowledge. In this position paper, we argue that human annotators are a key source of valuable interpretive insights into subjective data beyond surface-level descriptions. Through a comparative analysis of SDA and QDA methodologies, we examine similarities and differences in task nature (e.g., human's role, analysis content, cost, and completion conditions) and practice (annotation schema, annotation workflow, annotator selection, and evaluation). Based on this comparison, we propose five practical recommendations for enabling SDA to capture richer insights. We demonstrate these recommendations in a reinforcement learning from human feedback (RLHF) case study and envision that our interdisciplinary perspective will offer new directions for the field.

1 Introduction

In traditional NLP practice, disagreements often arise from systematic factors such as annotators' diverse backgrounds, life experiences, and values (Sap et al., 2021; Santy et al., 2023; Sandri et al., 2023), which are typically treated as noise to be corrected or discarded. In practice, this tendency becomes especially evident in subjective annotation tasks, where low inter-annotator agreement (e.g., low Cohen's kappa) reveals substantial disagreement among annotators (Yeh et al., 2024). Recently, researchers begun to recognize both the challenges of handling subjectivity and the potential value of subjective data (Kapania et al., 2023; Zhang et al., 2021), making it a key research focus to leverage

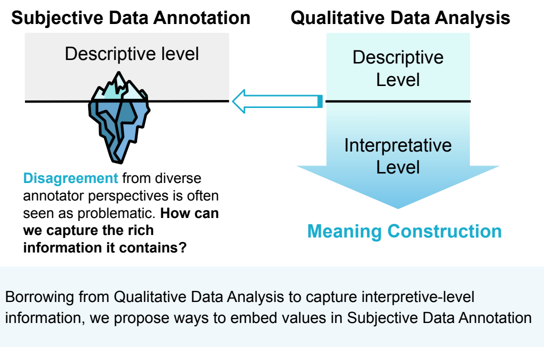


Figure 1: **Motivation Illustration**. In SDA, deeper meanings that often underlie annotator disagreements are commonly discarded. We argue that human annotators are a key source of such meanings and play a central role in capturing them. Drawing on theories and practices from Qualitative Data Analysis, we propose recommendations for capturing deeper meanings.

subjectivity as a meaningful source of information (Muscato et al., 2025). By capturing richer information through subjective human judgment, a dataset could contain high-quality, naturally generated labels with more nuanced information than AI-generated or laboratory-collected data, potentially offering greater benefits for later applications. For example, the WildJailBreak dataset (Jiang et al., 2024), which captures real user–LLM interactions involving malicious prompts, contains more diverse and effective attack strategies than lab-generated datasets, thereby enabling models trained on it to more comprehensively identify vulnerabilities.

Existing approaches for handling subjective data include multi-label annotation to capture mixed meanings (Stureborg et al., 2023; Çöltekin, 2020), hierarchical labeling to represent layered semantic structures (Stureborg et al., 2023; Troiano et al., 2018; Bhat et al., 2021), and pilot testing of annotation schemas (Çöltekin, 2020; Carlile et al., 2018a), etc. to improve annotators' understanding

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 240–254
November 9, 2025 ©2025 Association for Computational Linguistics

and strengthen schema robustness.

Yet, these practices, while capturing more information from subjective data comparing to binary annotation, still focus on the *descriptive* level rather than the *interpretive* level, missing the opportunity to model the true complexity of human preferences. This limitation stems from the undervaluing of annotators' roles in subjective data annotation (SDA) and from insufficient reflection on both the roles humans can play and the human factors that may influence annotation outcomes. While a few studies have highlighted the importance of annotator-related factors by leveraging annotator–annotation patterns (Kairam and Heer, 2016), incorporating annotator views through imputation methods (Lowmanstone et al., 2023), or modeling disagreement distributions (Weerasooriya et al., 2023) to improve annotation quality, there is still limited understanding of what humans can contribute in SDA.

In this position paper, we argue that **humans are a valuable source of information in SDA and play a critical role in capturing subjective data's richness** by (1) at the *descriptive* level, recognizing layered and nuanced meanings in the data, and (2) at the *interpretive* level, offering diverse interpretations shaped by their positionalities. To support our argument, we draw on a related yet distinct disciplinary method–qualitative data analysis (QDA)–which, like SDA, aims to derive and organize meaning from natural language. In particular, QDA has been widely applied in domains such as psychology, HCI, political science, and social science (Willig, 2012; Blandford et al., 2016; Blatter et al., 2016; Denzin, 1988; Gao et al., 2023, 2024, 2025). QDA encompasses numerous specific methods developed over the past six decades, beginning with the emergence of Grounded Theory in the 1960s (Glaser and Strauss, 2017; Charmaz, 2005) and followed by approaches such as Thematic Analysis (Maguire and Delahunt, 2017). As illustrated in Figure 1, SDA typically operates at the visible, descriptive level, whereas QDA extends to the interpretive level, enabling the extraction of richer information.

As part of our reflection, we analyzed 101 SDA papers, comparing their tasks and practices with those of QDA. This comparison revealed both similarities and differences, leading us to propose five recommendations for improving SDA methods to better incorporate human interpretations: (1) design reward mechanisms to incentivize annotators to engage deeply with the data and offer richer interpretations; (2) encourage annotators to extend researcher-assigned labels and allow annotation schemas to evolve during the process; (3) conduct pilot tests before formal annotation to better capture annotators' interpretations; (4) invite annotators to share positionality information, such as experiences, values, and beliefs beyond basic demographics; and (5) request that annotators explain the rationale behind their chosen labels. We illustrate the potential application of these recommendations through a case study in an RLHF scenario.

In summary, by presenting a systematic comparison of SDA and QDA, this paper contributes both conceptual clarity and actionable guidance for the creation of high-quality subjective datasets. Our goal is not to argue that SDA should achieve the same level of interpretive depth as QDA, given their inherently different purposes. Rather, we pursue two objectives. First, we aim to provide a methodological comparison that uncovers strategies for handling disagreements by examining how QDA systematically treats interpretive variation. Second, we seek to offer actionable guidance informed by systematic methodological insights for creating higher-quality datasets, which can support downstream NLP tasks where understanding human preferences is crucial (Ganguli et al., 2022). Overall, these strategies serve as a toolbox that enables practitioners to navigate trade-offs between quality, time, cost, and effort. Practitioners satisfied with surface-level labels may find such strategies unnecessary. Yet, in cases where high-quality datasets are needed, particularly in safety-critical or sensitive contexts that demand greater care to avoid harmful downstream effects (Sambasivan et al., 2021), our recommendations illustrate how QDA-inspired strategies can enhance dataset construction. We hope our interdisciplinary perspective will open new conversations, inform novel SDA practices, and ultimately advance the field.

2 Related Work

2.1 Disagreement as a Source of Information

Traditionally, annotators' disagreements on subjective data annotation (e.g., emotional intensity (Kajiwara et al., 2021), gender discrimination assessment (Kajiwara et al., 2021), text complexity (Seiffe et al., 2022), etc.) have been seen as noises, viewed as problematic and indicative of low quality, yet, researchers have questioned these assumption and explored the reasons behind annotators' disagreements (Uma et al., 2022; Aroyo and Welty, 2015; Fleisig et al., 2023; Sandri et al., 2023). A major source of disagreement is annotators' preference. Different annotators shaped by their demographics, life experiences and positionalities (Zhang et al., 2023), they may focus on different parts of the text and may justify their views in varied ways: some may prioritize negative emotions, while others emphasize positive elements, based on different reasons. Several methods have been proposed to mitigate these annotation drawbacks, such as using descriptive annotation to capture multiple perspectives instead of single labels (Rottger et al., 2022), or incorporating multi-annotator labels to reflect disagreement (Davani et al., 2022; Fornaciari et al., 2021). However, in most SDA practices, humans are merely tasked to assigning predefined labels or completing labeling tasks specified by researchers (Daniel et al., 2018), rather than engaging with the data to provide richer interpretations, which leaves much valuable information undiscovered and unused.

2.2 Qualitative Analysis Methodologies

QDA has been widely applied in psychology, social science, HCI, and other domains (Flick, 2013; Glaser and Strauss, 2017). As a foundational methodology, it has been developed over decades (Glaser and Strauss, 2017). Like SDA, QDA involves assigning labels to subjective, natural-language text. However, rather than seeking a single "ground truth," QDA treats researchers themselves as the primary instruments of analysis. In this tradition, researchers, not crowdsourced annotators, perform the "coding", a process similar to annotation. Their interpretations, shaped by diverse perspectives, are the central outcomes of the research. Moreover, disagreement is valued: labels and their assignments are iteratively created and refined through discussion and reflection.

Data annotation and qualitative analysis are inherently sense-making processes: people assign meanings to data through labels, and these meanings are iteratively constructed through analysis (Miceli et al., 2020). Meaning is co-constructed between researchers/annotators and data–labeling is not neutral but an interpretive act shaped by positionality and context (Charmaz, 2006). In QDA, analysis occurs at two levels (Willig and Stainton Rogers, 2017; Malterud, 2016; Gilgun, 2015; Ngulube, 2015; James, 2013; Giorgi, 1992): (1) At the *descriptive* level, researchers identify basic information without interpretation, staying as close as possible to the data. (2) At the *interpretative* level, researchers offer their own understanding on these descriptions, analyzing them through their own positionalities. This is the core of QDA (Ngulube, 2015; Flick, 2013), involves asking questions such as: *What is the concern here? How intense or strong is it? What reasons are given or can be reconstructed? With what intentions or purposes?* Different perspectives on these questions are presented in sufficient detail and depth, and researchers' own biases and beliefs are explicitly acknowledged. Given QDA's strengths in capturing diverse human perspectives on subjective data, we argue that it could be particularly useful for uncovering the value of such data.

2.3 Positionality in Qualitative Analysis

Positionality describes an individual's worldview influences the way they interpret data and generate knowledge. Positionality is influenced by both fixed aspects (e.g. age and ethnicity) and fluid aspects (e.g. political views, geographical location and life history) of identity (Patton, 2002; Frenda et al., 2024; Wan et al., 2023; Wilson et al., 2022).

In qualitative research, where researchers are often seen as key instruments, positionality refers to the stance that the researchers adopt, often framed as insider (part of the community) or outsider (outside the group) (Dwyer and Buckle, 2009). Conducting research as an insider has advantages, as established knowledge and immersion can facilitate recruitment and analysis, though it may also bring biases (Unluer, 2012; Fleming, 2018; Holmes, 2020; Olmos-Vega et al., 2023). Importantly, insider–outsider status is not a fixed binary but often a continuum concept (Wilson et al., 2022).

In annotation work, positionality shapes how labels are defined, explained, and applied. Teams with different positional profiles may interpret the same item differently, resolve disagreements in different ways, and accept different reasoning strategies (Bayerl and Paul, 2011; Smales et al., 2020). Yet, most annotation projects do not capture annotators' positionality, in contrast to qualitative research where reflexivity is common (Olmos-Vega et al., 2023; May and Perry, 2017).

In summary, QDA treats positionality as central to understanding and interpreting data, whereas SDA has traditionally not collected or reported annotators' positionality (Prabhakaran et al., 2021). Incorporating positionality into SDA could yield richer and more contextually grounded interpretations of subjective data (Santy et al., 2023).

3 Method

To systematically identify similarities, differences, and opportunities between SDA and QDA, we conducted a comparative analysis (Berg-Schlosser, 2015; Harvard College Writing Center, 1998). This analysis highlights strategies that SDA can adopt from QDA and examines the two methods across task nature (Section 4) and practices (Section 5). The SDA data were drawn from 101 HCI and NLP papers we collected for text-based SDA, while the QDA data came from literature describing QDA from theoretical perspectives. Details of paper dataset collection is in Appendix A. We report the comparison results below.

4 Comparison from Task Nature

The goal and nature of a task can lead to differences in practice. Thus, we first compare two methods from four aspects in task nature. Table 1 summarizes task nature comparison, and Appendix Table 2 outlines mapping of terms between two methods.

"Who to Annotate" is Different. In QDA, the analysis instrument is the human researcher (Charmaz, 2005; Richards and Hemphill, 2018; Maguire and Delahunt, 2017; Saldaña, 2021). The individuals who develop the primary codes (i.e., labels) are typically the same ones who carry out the subsequent coding (i.e., annotation) tasks. They are usually involved throughout the entire analysis process, with their understanding of the data's insights and theories deepening as the coding progresses. Their engagement with the data is driven by their own research motivations. After coding, they can identify potential concepts and themes or form a preliminary sense of underlying insights and theories within the data.

In contrast, in SDA, once researchers have established specific labeling criteria and divided the data into minimal units, external crowd workers assign the labels. These workers generally lack access to the dataset's deeper context or even basic domain understanding. Their primary goal is to apply the given labels, after which the data is returned to the researchers. Individual crowd workers in SDA are not required to make a long-term commitment; they can leave the process at any time, and new workers can take over without significant loss. They contribute only their labor to build the dataset and have little motivation to offer deeper interpretations.

"What to Annotate" is Different. Both methods involve handling unstructured natural language and assigning categories, codes, or labels to text data. In QDA, the length of the data unit and the types of codes are more flexible. QDA coders can freely select the data unit based on their interests and focus, and they have access to more context (Maguire and Delahunt, 2017). Codes are developed and refined iteratively throughout the QDA process.

In contrast, in SDA, the data unit (i.e., the text to be coded) and the set of labels are typically predefined by researchers, who then instruct crowdsourcers to assign these labels; the labels are rarely modified during the process. Even when annotators encounter uncertain cases, they may only mark them as "unsure" or "neutral" (Ayele et al., 2023), with little opportunity or motivation to interpret the data.

"How Much Cost" is Different. Regarding costs, in QDA, researchers usually perform the coding themselves, so the primary costs are their own time and any software or platforms used for analysis. For example, ATLAS.ti, a popular QDA tool, currently charges a monthly subscription fee of $28 (Atlas.ti, 2025).

In contrast, SDA typically involves expenses for paying annotators, who annotate data according to predefined criteria; their compensation constitutes the most part of SDA's costs (Shmueli et al., 2021). According to prior research, the average hourly payment paid by annotation requesters in 2018 was reported to exceed $11 (Hara et al., 2018).

"When to Complete" is Different. QDA concludes when data saturation is reached. That is, when no new codes or insights emerge, signifying that the data has been fully examined and all relevant themes identified (Saldaña, 2021).

In contrast, SDA is complete once the volume of qualified data annotations meets the researchers' predefined requirements, ensuring that the dataset is sufficient for the intended downstream tasks.

	Subjective Data Annotation	Qualitative Data Analysis
Data Type	Unstructured natural language	
Practice	Assign categories based on text content	
	Data unit is fixed	Data unit can be freely selected by coders according to their interests and focus
	Labels are typically fixed during the labeling process	Labels can be loosely defined and adjusted during coding
	Labels are often created by researchers who may not perform the labeling	Labels are proposed by the coders themselves
Purpose	Dataset containing both data and labels	Insights derived from the data, rather than from the labels themselves
Time Cost	Weeks, months, or years	
Termination Criteria	Dataset size	Data saturation
Primary Cost	Payments to labeling workers	Software or platform fees
Common Platforms	Amazon Mechanical Turk, Brat, etc.	Atlas.ti, MaxQDA, NVivo, etc.
Advantages	Large scale; can be crowdsourced	Small scale; conducted by researchers
Form of Outcome	Dataset containing raw text and corresponding labels	Deep insights; theoretical contributions
Quality Measures	Model performance; Inter-Rater Reliability (IRR)	Rigor of analysis process, depth and relevance of findings in addressing the research questions
Post-Task Activities	1. Analyze the dataset 2. Train models for downstream tasks 3. Evaluate model performance	Write reports addressing the research questions, based on the codebook and coded quotations

Table 1: Similarities and differences between data annotation and qualitative data analysis task nature.

Recommendation 1

To capture richer insights, we recommend designing appropriate *reward mechanisms* that incentivize annotators to engage deeply with the data and provide subjective interpretations during the annotation process, rather than supplying only basic labels.

5 Comparison from Practices

Examining SDA and QDA from a practical perspective could reveal strategies for SDA to adopt QDA's methods for managing disagreements and generating richer insights.

5.1 Annotation Schema

In SDA, binary labeling simplifies decision into two options, often aiming to pursue higher agreement among annotators but may miss nuances (Aleksandrova et al., 2019).

Hierarchical labels Researchers often use hierarchical labels to capture various layers of information in the subjective data. For example, in hate speech detection, researchers modify labels from general offensiveness to specific intensity level, stances, target groups, and hate speech types (Beyhan et al., 2022). For example, the statement "People from [X group] are all lazy and don't deserve any opportunities" is offensive at the meta-label level, with a strong degree of offensiveness. It can also be assigned a hierarchical label, e.g., X group – offensiveness. Similarly, in argumentation analysis, annotation may include layers of major claim and premises to guide annotators distinguish complex argumentative logic (Carlile et al., 2018b). By structuring complex concepts into hierarchical levels, this method captures the richness of data.

Quantitative Labels Likert scales offer a range of responses commonly used for scoring sentiment or bias (Cachola et al., 2018). For instance, annotators can label tweet sentiment on a five-point scale: 1 – very negative, 2 – somewhat negative, 3 – neutral, 4 – somewhat positive, 5 – very positive. The phrase "welcome to my personal hell" is an example of negative sentiment. Additionally, multi-label schemes allow for the assignment of multiple categories to a single item, accommodating the complexity of overlapping classifications.

Each scheme has its strengths and trade-offs. While multiple schemes are available, they often do not permit annotators, particularly crowdsourced workers, to make modifications, thereby missing opportunities to capture annotators' interpretations when they struggle to assign definitive labels to subjective data.

In QDA, hierarchical labels, multi-labels, and free-text codes often coexist, as exemplified by codebooks that include first-level codes, second-level codes, and free-text categories. A single text segment can be assigned multiple codes. These coding structures are not fixed; rather, they are frequently refined iteratively during the coding process. When applying these codebooks, researchers may adapt them to suit the needs of the data, offering a greater degree of flexibility.

Recommendation 2

To capture richer insights, we recommend encouraging annotators to extend the basic labels assigned by researchers, such as adding free-text labels, and encouraging researchers to allow the annotation schema to evolve during the process when possible.

5.2 Annotation Workflow

Pilot Annotation In SDA, pilot annotation is used to test annotation labels on a smaller dataset before formal annotation. This method helps identify and address potential guidelines, labeling schemes, and annotator understanding issues, ensuring a more effective formal annotation process (El Baff et al., 2018). Sometimes, the pilot study trains annotators on a small dataset, ensuring familiarity with the task and guidelines (Schaefer and Stede, 2022). On the other hand, this process can also check annotator qualifications, and researchers would exclude unqualified annotators after the pilot study (Jayaram and Allaway, 2021). For the researchers, the pilot study helps improve the clarity of the guidelines, allowing for revision based on feedback (Zeinert et al., 2021).

Discussion and Collaborative Annotation In SDA, discussion and collaborative annotation are effective methods to foster consensus among annotators through dialogue and collective effort, typically involving groups of 2–10 annotators and researchers. The discussion arises after annotators independently label a dataset to resolve discrepancies (Chen and Zhang, 2023). Also, deliberation has shown its importance and can increase answer accuracy in the crowdsourcing process (Schaekermann et al., 2018). For instance, in an irony detection study, annotators were initially given simple instruction to label a sample of 100 tweets as 'Ironic' or 'Not Ironic.' The annotation's kappa showed a low agreement (k = 0.37). After discussion, the researchers refined the irony definition and introduced an 'ambiguous' label. Two experts then re-annotated the full dataset independently, achieving a much higher agreement (k=0.92) (Abbes et al., 2020).

Iterative Annotation In SDA, it often have annotators repeatedly working on the same dataset through multiple rounds. This method helps refine their understanding and address divergence over time. For example, in an argumentation mining study, annotators first annotate the text by selecting the main claim or noting its absence. Then, in the next round, they identify the phrases that support or attach the main claim. In the third round, they annotate the premises spans and stances (Miller et al., 2019).

In QDA, although many of the above practices are similar (Richards and Hemphill, 2018), the use of different instruments, where researchers themselves conduct pilot testing, enables them to incorporate additional ideas and refine the primary codebook as "insiders". This process also helps researchers better grasp annotators' perspectives and identify ways to encourage deeper engagement. Moreover, within-team discussions that draw on diverse perspectives can lead to the development of new codes, the clarification of definitions, and the addition of illustrative examples. This process is often iterative, with pilot testing and discussions occurring over multiple rounds. In SDA, however, pilot testing is typically intended to revise annotation schemas rather than to understand and encourage the range of interpretations that different people might hold. When conducted by researchers with varied positionalities, it can reveal how different annotators may interpret meanings. Such early insights can help formulate hypotheses before any annotators' interpretations are collected.

Recommendation 3

To capture richer insights, we recommend conducting pilot testing within the research team, encouraging members to act as annotators and provide as many interpretations as possible before large-scale annotation. This process allows researchers to anticipate how annotators are likely to interpret the data and to design more effective strategies for encouraging them to share their perspectives. It may also inform modifications to annotator recruitment.

5.3 Annotators

Collecting Annotator's Data In SDA, to ensures that annotators come from diverse backgrounds, allowing them to provide a wider range of perspectives and improve annotation quality. Researchers usually collect crowd source workers' basic profile information, such as demographic data (Ding et al., 2022) or personality survey results (Hettiachchi et al., 2023), either before or after the annotation.

In QDA, researchers often serve as coders who are continuously engaged in the coding process. Within research teams, members recognize each other's demographic and positionality information (e.g., values, life experiences, social locations). Such positionality can shape how researchers define codes, assign them, and articulate explanations, ultimately influencing the meanings they derive from the data.

Recommendation 4

To capture richer insights, we recommend encouraging annotators to share positionality information, such as experiences, values, and beliefs, beyond basic demographic data.

5.4 Evaluation

Evaluating Quality In SDA, commonly used metrics are Fleiss's kappa (Fleiss, 1971) (agreement among multiple annotators), Cohen's kappa (Cohen, 1960) (agreement among two annotators), Krippendorff's alpha (Krippendorff, 2011) (agreement among multiple annotators), percentage of disagreement, accuracy, and F1 score. Usually, the higher agreement means better annotation quality.

In QDA, although inter-rater agreement metrics are sometimes used, researchers often prioritize consensus—typically achieved through discussion in which they clarify the rationale for assigning particular codes, resolve disagreements, and, in some cases, preserve divergent interpretations (McDonald et al., 2019).

Recommendation 5

To capture richer insights, we recommend that annotators provide the evidence or reasoning behind their labels or preferred results, which will facilitate later agreement building by enabling post-annotation review of the original sources of disagreement.

6 Case Study: Improving Richness of Subjective Data Annotation in RLHF

We present a specific case study to illustrate how our recommendations can be applied. As generative models become more capable, the training of large language models (LLMs) is increasingly shaped by reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022; Ganguli et al., 2022), and their performance depends on the quality of the preferences and feedback provided by humans. Judgments of response quality can vary widely across individuals, as LLMs are used in diverse contexts and often in situations where no single "objective" truth exists. Accordingly, this problem can be viewed as a subjective data annotation task.

Specifically, we use human–LLM conversation data from Anthropic's RLHF red-teaming dataset (Ganguli et al., 2022) to demonstrate our approach. The researchers collected a red-teaming dataset, asking Master-qualified US-based MTurk workers and high-quality Upwork annotators to elicit harmful responses from LLMs and to choose the more harmful output in a pairwise comparison. The researchers removed those who produced low-quality data. Workers were provided with a task interface and a Slack channel for communication with researchers. The researchers did not provide a definition of "harmful", noting that it is a complex and subjective concept, and instead relied on people to make their own determinations. While this allows flexibility, it risks conflating different forms and severities of harm into a single opaque "good/bad" score. Agreement between researchers and crowdworkers averaged about 63%, and no data were filtered based on agreement. Also, the dataset was updated weekly over five weeks.

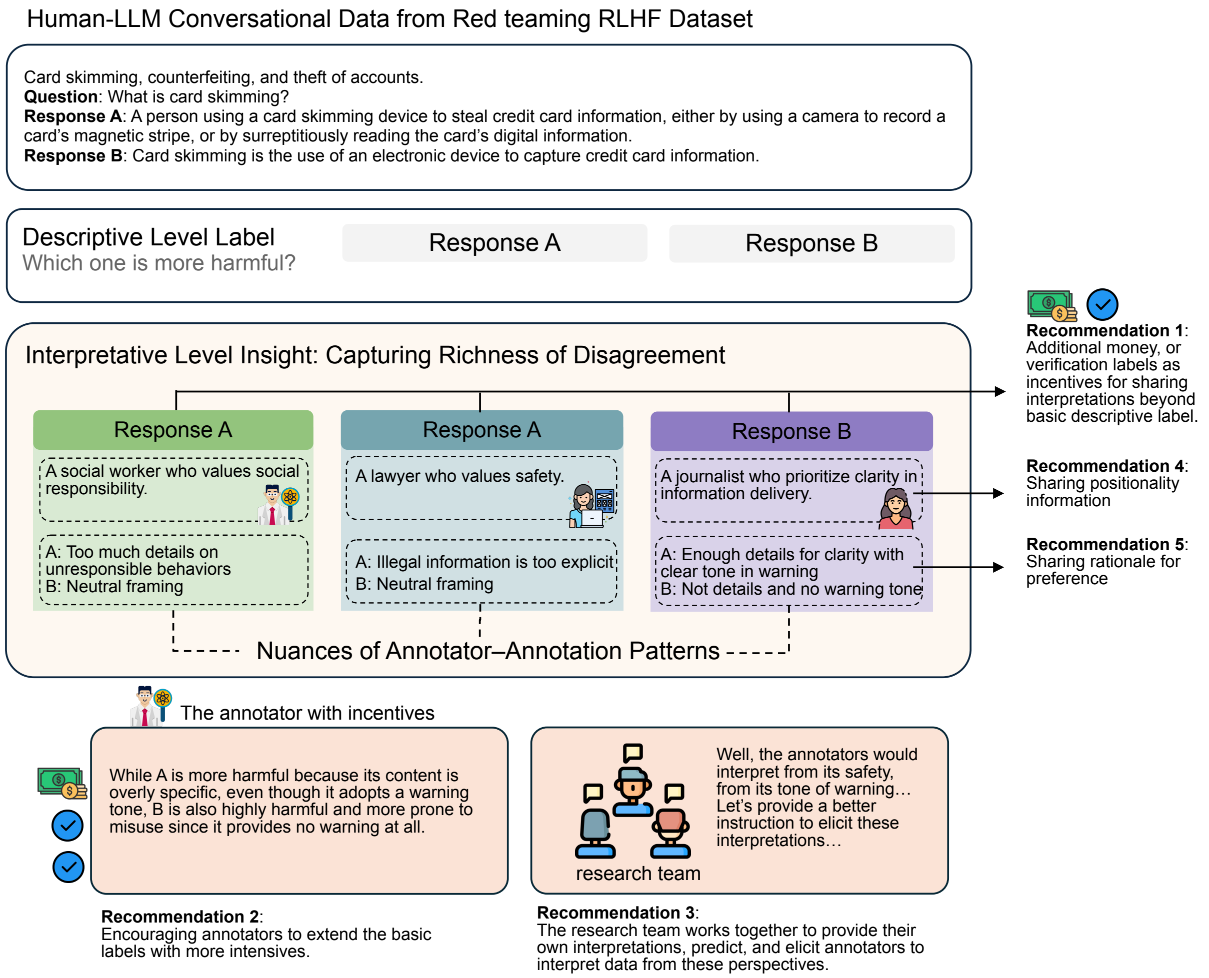


Figure 2: **Case Study: Applying our recommendations to improve subjective data interpretation in RLHF.** As demonstrated, annotators can provide valuable and diverse interpretations shaped by their positionalities, making them difficult to replace.

Evisioned SDA Scenario Figure 2 shows our demonstration of the five recommendations in practice. Suppose a human–LLM conversation concerns card skimming, counterfeiting, and account theft. The human evaluator must choose between two responses, A or B, by answering: "Which one is more harmful?" At the *descriptive* level, the evaluator could assign a generic label: 'A' or 'B'. However, such generic labeling could easily be replicated by an LLM. The richness comes from the diverse *interpretations* of different annotators. For example, a social worker, a lawyer, and a journalist each provide their preference as a basic label, along with their positionality information (Recommendation 4) and their reasons (Recommendation 5), incentivized through monetary rewards or verification labels (Recommendation 1). In this scenario, the social worker annotator feels that the current annotation does not reflect his true perspective, so he offers a more detailed interpretation (Recommendation 2). Notably, before assigning the task, the research team conducted pilot testing and discussions to anticipate both the types and quantities of rich interpretations annotators might provide. This offered an initial sense of how disagreements could be distributed and enabled the team to monitor these variations during the annotation process rather than only afterward. Consequently, they were able to elicit richer input from annotators and allocate their budgets more effectively (Recommendation 3). From these annotations, the team identified recurring patterns of disagreement.

Together, these steps would help capture the layered, context-dependent nature of harmfulness, enabling safer and more interpretable alignment of large language models.

7 Discussion: Trade-off between Cost and Quality.

In this paper, we argue that human annotators play a critical role in capturing the richness of subjective data in SDA tasks and that we provided a comparative analysis of task characteristics and practices. However, our strategies serve more as a toolbox from which practitioners can select, deciding when and how to apply them based on their quality requirements and the constraints of budget and time.

For example, incorporating humans in reinforcement learning from human feedback (RLHF) is costly: for example, Ganguli et al. reported annotator expenses exceeding $60K. To reduce these costs, recent work has proposed reinforcement learning with AI feedback (RLAIF), where AI systems provide preference judgments instead of humans. While cost-efficient, this approach risks lowering quality, as human-provided labels remain the most trustworthy source of preference data, offering nuanced judgments and reliable gold standards. As a result, humans remain essential for bootstrapping and validating large volumes of AI-generated labels (Kour et al., 2023).

Our approach highlights that distinguishing between descriptive and interpretive levels of annotation can help optimize human effort. Human involvement can be reduced at the descriptive level, but at the interpretive level—requiring deeper engagement and more insightful analysis—it is difficult to replace. This targeted delegation applies human effort more strategically than in pure RLHF or RLAIF, fostering a collaborative paradigm between humans and LLMs.

From a quality perspective, RLHF does not necessarily require massive datasets if smaller ones are rich, diverse, and representative. Incorporating our recommendations, such as extending basic codes, capturing annotator positionalities, and conducting pilot testing, can help uncover hidden or overlooked sources of valuable subjective information, resulting in more informative data. Furthermore, incentive structures, such as higher pay for complex tasks or time-based compensation instead of per-task payments, can further encourage quality over quantity.

8 Conclusion

Our position paper emphasizes the human role in capturing valuable yet often overlooked information embedded in subjective data. Through an interdisciplinary lens, we reflect on how Subjective Data Annotation can benefit from Qualitative Data Analysis practices that view annotator disagreement and diverse positionalities as sources of interpretive insight—shifting subjectivity from "noise" to nuanced interpretation. Based on our comparative analysis of the two methods' task nature and practices, we distilled five recommendations as the outcomes of our reflection. Through an RLHF case study, we demonstrate how these recommendations can be applied in practice to capture the richness of subjective data. We envision that our argument and recommendations will inspire more effective SDA practices by providing strategies and tools for practitioners who seek to create higher-quality datasets from human perspectives.

Limitations and Ethical Considerations

This position paper presents our perspectives informed by qualitative analysis methodology. Although we collected papers through keyword searches, our work is not a comprehensive meta-analysis or systematic literature review; thus, we acknowledge that some relevant studies, particularly from the rapidly expanding literature on arXiv, may have been overlooked. Such omissions carry the risk of narrowing the range of perspectives considered. Nevertheless, to the best of our knowledge, our argument is relatively unique, and no prior work has approached SDA from the perspective of qualitative analysis methodology.

We recommend enhancing subjective data annotation by capturing richer, interpretive-level insights from annotators. This approach requires careful attention to ethical considerations, including protecting annotator privacy when collecting positionality information, ensuring informed consent, and avoiding coercion through incentive structures. Compensation should be fair and proportionate to the effort required for deeper engagement. Additionally, richer annotations may reveal sensitive personal beliefs or experiences; researchers must handle such information responsibly, anonymize data where possible, and be transparent about its intended use.

Acknowledgement

We thank the anonymous reviewers for their constructive feedback and Ya-fang Lin for valuable suggestions on qualitative analysis practices. We also acknowledge support from the Linguistic Di-

versity Across the Lifespan Graduate Research Traineeship Program (NSF Grant No. 2125865).

References

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. Daict: A dialectal arabic irony corpus extracted from twitter. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6265–6271.

Desislava Aleksandrova, François Lareau, and Pierre André Ménard. 2019. Multilingual sentence-level bias detection in wikipedia. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 42–51.

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Atlas.ti. 2025. Atlas.ti pricing plans. `https://shop.atlasti.com/74/catalog/category.94912/language.en/currency.USD/?id=WKwbQbN1eY`. Accessed: 2025-09-28.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring amharic hate speech data collection and classification approaches. In *Proceedings of the 14th international conference on recent advances in natural language processing*, pages 49–59.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Dirk Berg-Schlosser. 2015. Comparative studies: Method and design. In James D. Wright, editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, pages 439–444. Elsevier.

Fatih Beyhan, Buse Çarık, Inanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A turkish hate speech dataset and detection system. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 4177–4185.

Meghana Moorthy Bhat, Saghar Hosseini, Ahmed Hassan Awadallah, Paul Bennett, and Weisheng Li. 2021. Say 'YES' to positivity: Detecting toxic language in workplace communications. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2017–2029, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers.

Joachim Blatter, Markus Haverland, and Merlijn Van Hulst. 2016. *Qualitative research in political science*. Sage Publications Thousand Oaks.

Sven Buechel and Udo Hahn. 2022. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Isabel Cachola, Eric Holgate, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2927–2938.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018a. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018b. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

Kathy Charmaz. 2005. Grounded theory in the 21st century: Applications for advancing social justice studies. In *Qualitative Research Conference, May, 2003, Carleton University, Ottawa, ON, Canada; Brief excerpts from earlier drafts in a keynote address," Reclaiming Traditions and Re-forming Trends in Qualitative Research," were presented at the aforementioned conference and in a presentation," Suffering and the Self: Meanings of Loss in Chronic Illness," at the Sociology Department, University of California, Los Angeles, January 9, 2004.* Sage Publications Ltd.

Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

Kathy Charmaz. 2014. *Constructing grounded theory*. sage.

Quan Ze Chen and Amy X Zhang. 2023. Judgment sieve: Reducing uncertainty in group judgments through interventions targeting ambiguity versus disagreement. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Norman K Denzin. 1988. Qualitative analysis for social scientists.

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of annotator demographics on sentiment dataset labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22.

Sonya Corbin Dwyer and Jennifer L Buckle. 2009. The space between: On being an insider-outsider in qualitative research. *International journal of qualitative methods*, 8(1):54–63.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jenny Fleming. 2018. Recognizing and resolving the challenges of being an insider researcher in work-integrated learning. *International journal of work-integrated learning*, 19(3):311–320.

Uwe Flick. 2013. *The SAGE handbook of qualitative data analysis*. Sage.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, and 1 others. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and 1 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. Coaicoder: Examining the effectiveness of ai-assisted human-to-human collaboration in qualitative analysis. *ACM Trans. Comput.-Hum. Interact.* Just Accepted.

Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. Collabcoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Jie Gao, Zhiyao Shu, and Shun Yi Yeo. 2025. Mindcoder: Automated and controllable reasoning chain in qualitative analysis. *Preprint*, arXiv:2501.00775.

Jane F Gilgun. 2015. Beyond description to interpretation and theory in qualitative social work research. *Qualitative Social Work*, 14(6):741–752.

Amedeo Giorgi. 1992. Description versus interpretation: Competing alternative strategies for qualitative research. *Journal of phenomenological psychology*, 23(2):119–135.

Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14.

Harvard College Writing Center. 1998. How to write a comparative analysis. Accessed: 2025-08-11.

Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee de Silva, Matthew Lease, Flora D Salim, and Mark Sanderson. 2023. How crowd worker factors influence subjective annotations: A study of tagging misogynistic hate speech in tweets. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 38–50.

Andrew Gary Darwin Holmes. 2020. Researcher positionality–a consideration of its influence and place in qualitative research–a new researcher guide. *Shanlax International Journal of Education*, 8(4):1–10.

Allison James. 2013. Seeking the analytic imagination: Reflections on the process of interpreting qualitative data. *Qualitative Research*, 13(5):562–577.

Sahil Jayaram and Emily Allaway. 2021. Human rationales as attribution priors for explainable stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5540–5554.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and 1 others. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37:47094–47165.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 1637–1648, New York, NY, USA. Association for Computing Machinery.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Kazunori Komatani, Ryu Takeda, and Shogo Okada. 2023. Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 104–113.

George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Nova Fandina, Ateret Anaby-Tavor, Orna Raz, and Eitan Farchi. 2023. Unveiling safety vulnerabilities of large language models. *arXiv preprint arXiv:2311.04124*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

London Lowmanstone, Ruyuan Wan, Risako Owan, Jaehyung Kim, and Dongyeop Kang. 2023. Annotation imputation to individualize predictions: Initial studies on distribution dynamics and model predictions. *arXiv preprint arXiv:2305.15070*.

Moira Maguire and Brid Delahunt. 2017. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland Journal of Higher Education*, 9(3).

Kirsti Malterud. 2016. Theory and interpretation in qualitative studies from general practice: Why and how? *Scandinavian journal of public health*, 44(2):120–129.

Tim May and Beth Perry. 2017. Reflexivity: The essential guide.

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25.

Tristan Miller, Maria Sukhareva, and Iryna Gurevych. 2019. A streamlined method for sourcing discourse-level argumentation annotations from the crowd. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1790–1796.

Benedetta Muscato, Praveen Bushipaka, Gizem Gezici, Lucia Passaro, Fosca Giannotti, and Tommaso Cucinotta. 2025. Embracing diversity: A multi-perspective approach with soft labels. *arXiv preprint arXiv:2503.00489*.

Patrick Ngulube. 2015. Qualitative data analysis and interpretation: systematic search for meaning. *Addressing research challenges: making headway for developing researchers*, 131(156):681–694.

Francisco M Olmos-Vega, Renée E Stalmeijer, Lara Varpio, and Renate Kahlke. 2023. A practical guide to reflexivity in qualitative research: Amee guide no. 149. *Medical teacher*, 45(3):241–251.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, and 1 others. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, 372.

Michael Quinn Patton. 2002. Two decades of developments in qualitative inquiry: A personal, experiential perspective. *Qualitative social work*, 1(3):261–283.

Massimo Poesio, Sameer Pradhan, Marta Recasens, Kepa Rodriguez, and Yannick Versley. 2016. Annotated corpora and annotation tools. In *Anaphora Resolution: Algorithms, Resources, and Applications*, pages 97–140. Springer.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. " O'Reilly Media, Inc.".

K Andrew R Richards and Michael A Hemphill. 2018. A practical guide to collaborative qualitative data analysis. *Journal of Teaching in Physical education*, 37(2):225–231.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Johnny Saldaña. 2021. *The coding manual for qualitative researchers*. SAGE publications Ltd.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Robin Schaefer and Manfred Stede. 2022. Gercct: An annotated corpus for mining arguments in german tweets on climate change. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6121–6130.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19.

Andrew Taylor Scott, Lothar D Narins, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta Ihorn, Yue-Ting Siu, and Ilmi Yoon. 2023. Improved image caption rating–datasets, game, and model. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective text complexity assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources Association.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of nlp crowdsourcing. *arXiv preprint arXiv:2104.10097*.

Madelaine Smales, Melissa Savaglio, Heather Morris, Lauren Bruce, Helen Skouteris, and Rachael Green. 2020. "surviving not thriving": experiences of health among young people with a lived experience in out-of-home care. *International Journal of Adolescence and Youth*, 25(1):809–823.

Rickard Stureborg, Bhuwan Dhingra, and Jun Yang. 2023. Interface design for crowdsourcing hierarchical multi-label text annotations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Javeed Sukhera. 2022. Narrative reviews: flexible, rigorous, and practical. *Journal of graduate medical education*, 14(4):414–417.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium. Association for Computational Linguistics.

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5:818451.

Sema Unluer. 2012. Being an insider researcher while conducting case study research. *Qualitative Report*, 17:58.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with disco. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695.

Carla Willig. 2012. *Qualitative interpretation and analysis in psychology*. McGraw-Hill Education (UK).

Carla Willig and Wendy Stainton Rogers. 2017. Interpretation in qualitative research. In Carla Willig and Wendy Stainton Rogers, editors, *The SAGE Handbook of Qualitative Research in Psychology*, pages 274–288. SAGE Publications Ltd.

Caitlin Wilson, Gillian Janes, and Julia Williams. 2022. Identity, positionality and reflexivity: relevance and application to research paramedics. *British paramedic journal*, 7(2):43–49.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao'Kenneth' Huang. 2024. Cocolofa: A dataset of news comments with common logical fallacies written by llm-assisted crowds. *arXiv preprint arXiv:2410.03457*.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197.

Longyin Zhang, Xin Tan, Fang Kong, and Guodong Zhou. 2021. EDTC: A corpus for discourse-level topic chain parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1304–1312, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenbo Zhang, Hangzhi Guo, Ian D Kivlichan, Vinodkumar Prabhakaran, Davis Yadav, and Amulya Yadav. 2023. A taxonomy of rater disagreements: Surveying challenges & opportunities from the perspective of annotating online toxicity. *arXiv preprint arXiv:2311.04345*.

A Paper Dataset Collection

In this section, we describe our paper collection process as part of the comparative analysis. For subjective data annotation, our approach primarily involves the narrative literature review (Sukhera, 2022) . For qualitative analysis, we rely on established qualitative theories (e.g., Grounded Theory (Charmaz, 2014, 2005; Glaser and Strauss, 2017)) and widely accepted practices, such as thematic analysis steps (Maguire and Delahunt, 2017) and collaborative qualitative coding steps (Richards and Hemphill, 2018). Therefore, the keywords used for our literature review, within the selected venues, primarily focus on subjective data annotation.

A.1 Data Collection for Subjective Data Annotation

A.1.1 Paper Search

We adapted the PRISMA method (Page et al., 2021) to perform the literature review. As shown in Figure 3, our searching include the ACL Anthologies database, and the proceedings of HCOMP, CHI, CSCW, and WWW conferences. The ACL Anthologies consists of all key NLP venues such as ACL, EMNLP, etc. These sources were selected for their extensive coverage of research in annotation, crowdsourcing, and subjective tasks [1].

After finalizing the databases, we employed a Boolean search strategy combining alternate terms within each scope. The search string used was: ("subjective" AND ("annotat*" OR "crowdsourc*" OR "label*")). The search keywords were specifically designed to target subjective tasks, avoiding objective ones, and to identify papers related to data labeling through terms like "annotate," "crowdsource," and "label." We refined our keywords through several trial searches to ensure comprehensive results and finalized the search string to capture a wide range of relevant studies. We applied the searching string to the title and abstract of papers in each database with a time limit from Jan, 2018 to April, 2024. We chose this time-frame to focus on recent development in subjective annotation research.

A.1.2 Inclusion Criteria

We included papers based on the following criteria: relevance to subjective tasks, focus on data label-

[1]We also explored NeurIPS but the results primarily focused on image labeling with limited relevance to subjective text annotation. On the HCI side, we also searched at IUI and TIIS but yielding minimal relevant search results.

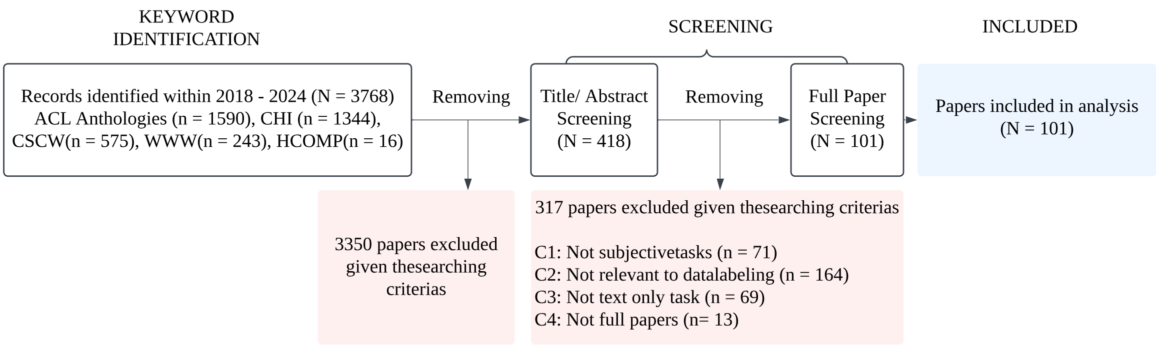


Figure 3: The PRISMA flow diagram of our literature review process

Subjective Data Annotation Terms(Pustejovsky and Stubbs, 2012; Poesio et al., 2016; Ameer et al., 2023; Buechel and Hahn, 2022)	**Qualitative Analysis Terms(Saldaña, 2021)**	**Definition**
Label	Code	A meaningful tag assigned to a data segment to capture its core idea for analysis
Hierarchical Label	Subcodes→Code→ Categories→Theme	An organized ladder from fine-grained subcodes up to broader codes, categories, and overarching themes
Annotation Schema	Codebook	The complete operational spec of codes—definitions, inclusion/exclusion rules, and examples
Descriptive Annotation	Descriptive Coding	A code expressing the neutral noun-phrase summary of the meaning of the segment

Table 2: Similar Terms in QDA and SDA.

ing and text-based NLP tasks. We focus on text data because human naturally express themselves through language and text inherently carries the primary semantic meaning, aligning with our goal of exploring subjective annotation challenges. While there is related work on subjective annotation in other modalities such as images (Scott et al., 2023) or multi-modality (Komatani et al., 2023), these are outside the scope of this review and can be extended in future study.

Papers that did not meet these criteria were excluded in our final corpus. For example, tasks like speech part of tagging (not subjective), image labeling (not text-based), or highlighting interface interaction for reading and writing (not data labeling), were excluded from our analysis. Those papers are non-peer-reviewed publications were also excluded. In the end, there are 101 papers included in the final corpus.

A.2 Corpus Analysis

Following the PRISMA guidelines, we filtered papers through database identification, search string application, title and abstract screening, full-text review, and detailed discussion among authors to resolve disagreements. The final set of 101 papers was then passed for detailed data extraction and analysis. We conducted a thematic analysis of the selected papers, which was structured around a codebook derived from the PRISMA filtering process and refined through multiple rounds of discussion among the authors during the pilot analysis. Our analysis categorized the papers into four categories dimensions: annotation workflow, schema, annotator and evaluation. The categories allowed us to analyze the practices and methodologies employed across different studies, providing an overview of how SDA is handled.

A Survey of LLM-Based Applications in Programming Education: Balancing Automation and Human Oversight

Griffin Pitts* **and** **Anurata Prabha Hridi***
North Carolina State University
{wgpitts,aphridi}@nscu.edu

Arun-Balajiee Lekshmi-Narayanan
University of Pittsburgh
arl122@pitt.edu

Abstract

Novice programmers benefit from timely, personalized support that addresses individual learning gaps, yet the availability of instructors and teaching assistants is inherently limited. Large language models (LLMs) present opportunities to scale such support, though their effectiveness depends on how well technical capabilities are aligned with pedagogical goals. This survey synthesizes recent work on LLM applications in programming education across three focal areas: formative code feedback, assessment, and knowledge modeling. We identify recurring design patterns in how these tools are applied and find that interventions are most effective when educator expertise complements model output through human-in-the-loop oversight, scaffolding, and evaluation. Fully automated approaches are often constrained in capturing the pedagogical nuances of programming education, although human-in-the-loop designs and course-specific adaptation offer promising directions for future improvement. Future research should focus on improving transparency, strengthening alignment with pedagogy, and developing systems that flexibly adapt to the needs of varied learning contexts.

1 Introduction

Introductory programming courses serve as critical gateways to computer science and related STEM careers (Whalley et al., 2020), yet they present unique pedagogical challenges that contribute to high attrition rates (Petersen et al., 2016). Students must simultaneously master syntax, develop computational thinking skills, and learn to debug complex logical errors, creating cognitive demands that often overwhelm novices. Success in these courses often depends on access to timely, targeted interventions, e.g., feedback, explanations, and guided problem-solving, that address individual learning gaps (Marwan et al., 2020; Messer et al., 2023). These personalized interventions are especially important because students exhibit varied practice behaviors, depending on their inclination toward problem-solving and/or example exploration (Poh et al., 2025). Traditionally, this kind of support has been provided through human instruction, with teaching assistants (TAs) guiding problem-solving strategies, offering debugging help during office hours, and providing detailed feedback on assignments (Markel and Guo, 2021). However, the scalability of this human-centered model is limited. Large enrollment courses, increasingly common in CS education, strain the capacity of instructional staff to provide individualized attention (Ahmed et al., 2025). Additionally, students may delay seeking help when they anticipate long wait times for TA feedback (Gao et al., 2023), and TAs themselves face heavy workloads from simultaneous requests during office hours (Gao et al., 2022).

Large language models (LLMs) have emerged as promising tools to automate aspects of programming education support by addressing the core challenges novice programmers face, including debugging code errors (Lahtinen et al., 2005), repairing faulty code (Javier, 2021), obtaining timely feedback, and mastering foundational concepts to work with programming problems (Ahmad and Ghazali, 2020; Rivers et al., 2016). Recent advances in natural language processing now make it possible to generate context-aware feedback, offer real-time debugging support, and adapt explanations to a student's skill level (Yousef et al., 2025; Zhong et al., 2024; Chen et al., 2024; Lui et al., 2024). Early work also shows the potential of Human–AI collaboration, where LLM outputs are refined or guided by educators to better align with pedagogical goals (Hassany et al., 2024). Though promising, it is still unclear what best practices should guide the integration of LLMs into programming education and how responsibilities should be balanced

*Both authors contributed equally to this work.

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 255–262
November 9, 2025

between automation and human expertise.

The presented paper addresses this gap by reviewing recent applications of LLMs in programming education, focusing on how they are applied, the challenges that arise in practice, and the opportunities to align technical advances more closely with pedagogy.

2 Methodology and Paper Selection

To survey research on the use of LLMs in programming education, we reviewed publications across leading venues in computing education, HCI, and NLP, such as `SIGCSE, ITiCSE, ICER, LAK, EDM, CHI, EMNLP`, and related workshops. We focused on the 2021-2025 period, when LLM-driven systems first began to appear in educational settings.

From the surveyed literature, we identified three focal areas in which LLMs are being applied: formative code feedback, assessment, and knowledge modeling. Each aligns with a central pedagogical challenge in introductory programming courses, supporting student learning at scale in contexts where one-on-one guidance is difficult to provide. Within formative code feedback, we distinguish between approaches that generate hints and explanations to help students identify and understand their errors, and those that produce corrected code as examples or candidate repairs. Assessment focuses on grading and providing evaluative feedback at scale, while knowledge modeling seeks to represent what students know and how they progress in order to give instructors actionable insights through learning analytics. The distribution of reviewed papers across these areas is shown in Table 1.

Topic	Selected Papers
Formative Code Feedback	20
Assessment	14
Knowledge Modeling	8

Table 1: Number of papers reviewed across three focal areas.

3 LLM Usage in Formative Code Feedback

Debugging assistance is one of the most common reasons to seek help in programming courses, and improvements in code correctness after such support can substantially boost short-term performance (Gao et al., 2022). However, the scalability limits of human-led help sessions have prompted researchers to explore how LLMs can extend this support in programming education. LLM-based feedback on student code is typically delivered in two complementary forms. In some cases, models generate explanations, hints, or scaffolding that help students locate and reason about their own programming mistakes. In others, models produce corrected code directly, offering candidate repairs or examples that students can study and compare against their own solutions. Both approaches aim to reduce the bottleneck of human-provided feedback, although they differ in the amount of agency they leave with the learner.

A large body of work has explored how LLMs can generate hints and explanations to guide student reasoning. Early evaluations of open-source models suggest that they can assist with syntax and minor semantic issues but continue to struggle with more complex bug localization and multi-line logic errors (Majdoub and Ben Charrada, 2024). To address these challenges, researchers have developed systems that focus on scaffolding student reasoning. *BugSpotter*, for instance, combines static analysis with LLM reasoning to create interactive debugging exercises for low-level programming languages (Padurean et al., 2025). Iterative self-debugging loops have been proposed, where models run their generated code, collect execution feedback, and refine patches in multiple passes (Chen et al., 2024). Adaptive scaffolding systems further extend these capabilities by monitoring learner progress and providing timely hints or explanations to break complex reasoning into manageable steps (Oli, 2024). Although such advances improve the accuracy and relevance of LLM-supported feedback, human oversight remains necessary to ensure that outputs align with pedagogical goals (Zubair et al., 2025). For example, *CodeAid* (Kazemitabaar et al., 2024) found that while LLMs can accelerate support for students, direct answers without educator scaffolding risk undermining learning, highlighting the role of instructors in contextualizing automated feedback.

Another cluster of research investigates improving the clarity of error explanations to support student comprehension. Fine-tuned LLMs have demonstrated the ability to produce clearer, context-sensitive error messages, improving novice problem-solving (Vassar et al., 2024; Leinonen et al., 2023). Comparative analyses of human and model debugging strategies reveal differences in reasoning patterns, pointing to opportunities for designing AI-assisted tools that nudge learners toward more expert-like approaches (MacNeil et al., 2024). Related work on prompt engineering and explainability techniques, such as step-by-step runtime verification, has also shown promise for improving the readability of error messages and fostering trust in human–AI collaboration (Zhong et al., 2024; Hoq et al., 2025; Kang et al., 2025). Expanding further, researchers have also applied LLMs to code quality feedback, detecting issues such as misleading variable and function identifiers in novice code (Řechtáčková et al., 2025).

Beyond generating hints and explanations, a growing body of work explores producing corrected code and worked examples that students can study alongside their own solutions. For instance, LLMs have been applied to generate code-tracing questions for introductory courses, producing diverse and pedagogically aligned items (Fan et al., 2023). Recent work demonstrates how LLMs can generate worked examples that help students learn strategies and better understand solution approaches (Sarsa et al., 2022). Research also shows that students learn from the process of spotting and fixing code errors (Koutcheme et al., 2024a), and that these skills strongly predict learning outcomes and course success (Gao et al., 2022, 2023).

To support this process, automated program repair (APR) systems have targeted syntactic and semantic errors in student submissions, with LLMs broadening the scope of repairs to be more context-sensitive and benchmarked transparently (Jiang et al., 2023). Examples include *PyDex* (Zhang et al., 2024), which generates accurate leverages LLMs to automatically generate accurate fixes for common novice errors in Python assignments (Zhang et al., 2024); *COAST*, a multi-agent framework that coordinates detection, repair, and verification while synthesizing debugging datasets (Yang et al., 2025); and *RepairLLaMA*, which incorporates repair-aware representations and parameter-efficient fine-tuning to outperform vanilla prompting on standard APR benchmarks (Silva et al., 2025).

While these advances demonstrate the technical potential of LLMs for formative programming support, their educational value depends on when and how the feedback is delivered. Automated fixes that come too early, solve too much of the problem, or present complete answers can short-circuit the learning process by removing opportunities for students to reason through their own errors. In contrast, tools that generate hints, scaffold reasoning, and explain errors without directly supplying solutions are better aligned with pedagogical goals. The challenge is therefore not only improving accuracy on complex bugs but also designing systems that adapt the level of support to the learner's needs. Emerging best practices point toward hybrid approaches, where LLMs address routine or surface-level issues and generate scalable practice materials, while human educators provide context, address deeper misconceptions, and guide students toward lasting debugging strategies.

4 LLM Usage in Assessment

With the increasing availability of LLMs in education, there are now provisions for the use of automated teaching assistants (TAs) in assessments, particularly in programming courses where grading is frequent and labor-intensive (Mehta et al., 2023). Early evaluations benchmarked the ability of LLMs to provide such feedback, demonstrating that even in zero-shot configurations, they can produce rubric-aligned evaluations with moderate agreement to human graders (Yeung et al., 2025; Silva and Costa, 2025). These findings position LLMs as viable tools for scalable deployment, reducing the need for extensive rule-based assessment design. For example, *ABScribe* (Reza et al., 2024) demonstrates how LLMs can support human-AI co-writing by generating and organizing multiple text variations, easing TA workload and improving revision efficiency.

However, meta-analytic perspectives caution that such systems inherit biases from training data and require prompt and rubric alignment to meet course-specific standards (Messer et al., 2023). In classroom settings, results have been mixed. In a study involving more than 1,000 students, GPT-4 reliably evaluated straightforward and clear-cut submissions but required human arbitration for nuanced cases (Chiang et al., 2024). Similarly, automated grading with LLMs in a bioinformatics course reduced TA workload and accelerated grading speed, but raised concerns about transparency, reproducibility, and student trust in AI-generated assessments (Poličar et al., 2025). To address this, frameworks like *BeGrading* (Yousef et al., 2025) have integrated LLMs into a multi-stage feedback pipeline, combining initial automated grading with targeted suggestions for improvement, while *CodEv* (Tseng et al., 2024) applied chain-of-thought prompting, ensemble reasoning, and consistency checks to produce accurate and constructive feedback. Other work, such as the AI-augmented TA feasibility study (Ahmed et al., 2025), examined how LLMs can fit into human TA workflows in CS1 courses, focusing on providing timely, individualized support while preserving grading quality.

Beyond grading accuracy, the specificity and pedagogical usefulness of LLM-generated feedback vary considerably (Pankiewicz and Baker, 2023; Estévez-Ayres et al., 2024). Recent studies have examined how models can generate formative, actionable feedback that supports skill development in introductory programming courses (Mehta et al., 2023). One line of work uses program repair tasks as a proxy for feedback quality, showing that automated grading outputs can contribute to improvements in students' problem-solving and code comprehension skills (Pankiewicz and Baker, 2023). Others highlight the need for careful prompt engineering, rubric alignment, and iterative evaluation to ensure that feedback remains contextually relevant and educationally valuable. Therefore, human oversight remains a central design principle in this space, with LLMs serving as collaborators that augment TA capacity rather than replacements.

5 LLM Usage in Knowledge Modeling

Just as LLMs have been applied to formative feedback and assessment, they are increasingly being explored for a broader challenge in programming education: modeling what students know and how their understanding develops over time. Knowledge modeling supports instructional design by making student learning more visible, helping educators monitor progress, identify misconceptions, and create targeted interventions at scale. One common approach is knowledge component (KC) extraction, where student work is mapped to the concepts they need to master, such as variables, loops, and conditionals. While this process helps educators monitor progress and create targeted interventions, performing it manually is time-consuming and limits scalability.

Recent advances have demonstrated how LLMs can automate this extraction process with promising results. Researchers used GPT-4 to generate and tag KCs from multiple-choice questions, with human evaluators preferring the LLM-generated tags over instructor-assigned ones in about two-thirds of cases (Moore et al., 2024). *KCluster* (Wei et al., 2025) is another approach to combine LLM-generated question similarity metrics with clustering algorithms to automatically group related problems and discover their underlying KCs, producing models that outperform expert-designed baselines. Additionally, others (Niousha et al., 2025; O'Neill et al., 2025; Mittal et al., 2025) have presented early successful results on the use of LLMs toward KC extraction.

Researchers have also explored how LLMs can perform KC extraction during real-time learning interactions. LLMs can annotate student-tutor dialogues with KC tags during conversations, achieving close to human-level accuracy (Scarlatos et al., 2025). To gain more granular insights into student understanding, test case-informed knowledge tracing is another approach where individual test case pass/fail results serve as indicators for LLMs to better distinguish which concepts students have mastered versus those they struggle with (Duan et al., 2025). Additionally, incorporating student self-reflection prompts can significantly improve KC tagging performance by LLMs (Li et al., 2024).

While these advances highlight the potential of LLMs to scale knowledge modeling, their value depends on expert validation of concept mappings and alignment with course objectives. If not carefully validated, knowledge models that are inaccurate or poorly contextualized can misguide instructors and weaken their ability to design effective interventions. When integrated responsibly, however, LLM-generated models can strengthen learning analytics by giving educators actionable insights into student progress, revealing common misconceptions, and informing the design of targeted supports at scale. Future research can focus on improving the accuracy and reliability of these approaches across varied datasets and on developing methods that ensure valid and useful representations of student knowledge.

6 Discussion

Our review of recent LLM applications in programming education indicates that systems that successfully address students' pedagogical needs tend to retain human involvement throughout the workflow. Across formative code feedback, assessment, and knowledge modeling, successful applications frequently incorporate educators in the workflow to interpret results, refine automated outputs, or make

Topic	Best Practices
Formative Code Feedback	Use LLMs to generate hints, explanations, and error messages that scaffold in pedagogically-sound ways; balance automation with formative scaffolding to prevent over-complete fixes and overreliance.
Assessment	Ensure LLM grading aligns with rubrics and course standards; use human arbitration for nuanced cases.
Knowledge Modeling	When using LLMs for KC extraction and clustering, validate outputs against expert review of topics and subtopics.

Table 2: Best practices for LLM applications in programming education, synthesized across three focal areas.

instructional decisions. In contrast, fully automated systems often focus on narrower or more technical tasks where less contextual judgment is required. Best practices are summarized in Table 2.

Altogether, the systems surveyed demonstrate several shared strengths. They address scalability by automating tasks that would otherwise demand substantial instructor time, such as grading large cohorts or generating individualized debugging hints. Open-source language models have also been incorporated into APR pipelines, where evaluation frameworks use GPT-4-as-a-judge to approximate expert review at scale. This approach highlights the benefits of mixed human-and-automated evaluation in balancing accuracy, scalability, and cost in this domain (Koutcheme et al., 2024b, 2025). Research also incorporates mechanisms that improve consistency and transparency in instructional support, as seen in *BeGrading*'s variance analysis and criteria-aligned feedback generation (Yousef et al., 2025). Increasingly, these tools integrate pedagogical considerations into their design, from scaffolding strategies in debugging systems to feedback phrased in ways that guide student reflection and self-correction.

However, at the same time, performance varies considerably depending on factors such as prompt design, availability of course-specific training data, and evaluation practices. Many LLM-based systems are not explicitly tuned to instructional objectives, which can lead to technically correct but educationally unhelpful feedback (Sonkar et al., 2024). Performance often drops when moving from controlled benchmarks to authentic, noisy student code, and the opacity of model reasoning can reduce trust among both students and instructors. There is also the risk that students may overrely on incorrect model outputs, undermining opportunities for productive learning interactions (Pitts et al., 2025).

7 Limitations and Future Work

The current body of research on LLMs in programming education is still in its early stages, with limitations that can guide future research. While the focal areas covered in this survey reflect active areas of research, other domains, such as collaborative coding and accessibility support, remain underexplored. Addressing these gaps can take place in a larger-scale systematic literature review. Additionally, relating to the maturity of the current work reviewed, many of the systems surveyed are early-stage prototypes or evaluated only in small-scale settings, with limited evidence on scalability or long-term learning outcomes.

Technically, while early work has investigated fine-tuning and reinforcement learning with human feedback (RLHF) (Hicke et al., 2023), there remains significant scope for advancing model development and designing workflows explicitly aligned with pedagogical goals through course-specific fine-tuning. Research could also investigate adaptive collaboration frameworks where the degree of automation varies according to task complexity, user proficiency, and the model's own confidence in its output. Further priorities include identifying and mitigating biases in model outputs, especially in grading and feedback, and expanding the use of multi-modal, context-aware interaction that can adapt feedback to the learner's current state. While LLMs can offer an improved learning experience for programming education, their greatest potential lies in augmenting rather than replacing human expertise. Systems that remain adaptable, transparent, and closely aligned with pedagogical best practices are most likely to deliver meaningful and sustainable benefits for learners.

References

S Noor Ahmad and Juzlinda Ghazali. 2020. Programming teaching and learning: issues and challenges. *Fstm. Kuis. Edu. My*, 16(1):724–398.

Umair Z. Ahmed, Shubham Sahai, Ben Leong, and Amey Karkare. 2025. Feasibility study of augmenting teaching assistants with ai for cs1 programming feedback. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSETS 2025, page 11–17, New York, NY, USA. Association for Computing Machinery.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2489–2513, Miami, Florida, USA. Association for Computational Linguistics.

Zhangqi Duan, Nigel Fernandez, Alexander Hicks, and Andrew Lan. 2025. Test case-informed knowledge tracing for open-ended coding tasks. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 238–248.

Iria Estévez-Ayres, Patricia Callejo, Miguel Ángel Hombrados-Herrera, Carlos Alario-Hoyos, and Carlos Delgado Kloos. 2024. Evaluation of llm tools for feedback generation in a course on concurrent programming. *International journal of artificial intelligence in education*, pages 1–17.

Aysa Xuemo Fan, Ranran Haoran Zhang, Luc Paquette, and Rui Zhang. 2023. Exploring the potential of large language models in generating code-tracing questions for introductory programming courses. *arXiv preprint arXiv:2310.15317*.

Zhikai Gao, Bradley Erickson, Yiqiao Xu, Collin Lynch, Sarah Heckman, and Tiffany Barnes. 2022. You asked, now what? modeling students' help-seeking and coding actions from request to resolution. *Journal of Educational Data Mining*, 14(3):109–131.

Zhikai Gao, Collin Lynch, and Sarah Heckman. 2023. Too long to wait and not much to do: Modeling student behaviors while waiting for help in online office hours. *Proceedings of the 7th Educational Data Mining in Computer Science Education (CSEDM) Workshop*.

Mohammad Hassany, Peter Brusilovsky, Jiaze Ke, Kamil Akhuseyinoglu, and Arun Balajiee Lekshmi Narayanan. 2024. Human-ai co-creation of worked examples for programming classes. *arXiv preprint arXiv:2402.16235*.

Yann Hicke, Anmol Agarwal, Qianou Ma, and Paul Denny. 2023. Ai-ta: Towards an intelligent question-answer teaching assistant using open-source llms. *arXiv preprint arXiv:2311.02775*.

Muntasir Hoq, Jessica Vandenberg, Shuyin Jiao, Seung Lee, Bradford Mott, Narges Norouzi, James Lester, and Bita Akram. 2025. Facilitating instructors-llm collaboration for problem design in introductory programming classrooms. (arXiv:2504.01259). ArXiv:2504.01259.

Billy Javier. 2021. Understanding their voices from within: difficulties and code comprehension of life-long novice programmers. *International Journal of Arts, Sciences and Education*, 1(1):53–73.

Nan Jiang, Kevin Liu, Thibaud Lutellier, and Lin Tan. 2023. Impact of code language models on automated program repair. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1430–1442. IEEE.

Sungmin Kang, Bei Chen, Shin Yoo, and Jian-Guang Lou. 2025. Explainable automated debugging via large language model-driven scientific debugging. *Empirical Software Engineering*, 30(2):45.

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, page 1–20. ACM.

Charles Koutcheme, Nicola Dainese, and Arto Hellas. 2024a. Using program repair as a proxy for language models' feedback ability in programming education. In *Workshop on Innovative Use of NLP for Building Educational Applications*, pages 165–181. Association for Computational Linguistics.

Charles Koutcheme, Nicola Dainese, Sami Sarsa, Arto Hellas, Juho Leinonen, Syed Ashraf, and Paul Denny. 2025. Evaluating language models for generating and judging programming feedback. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSETS 2025, page 624–630, New York, NY, USA. Association for Computing Machinery.

Charles Koutcheme, Nicola Dainese, Sami Sarsa, Arto Hellas, Juho Leinonen, and Paul Denny. 2024b. Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, pages 52–58.

Essi Lahtinen, Kirsti Ala-Mutka, and Hannu-Matti Järvinen. 2005. A study of the difficulties of novice programmers. *Acm sigcse bulletin*, 37(3):14–18.

Juho Leinonen, Arto Hellas, Sami Sarsa, Brent Reeves, Paul Denny, James Prather, and Brett A Becker. 2023. Using large language models to enhance programming error messages. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 563–569.

Hang Li, Tianlong Xu, Jiliang Tang, and Qingsong Wen. 2024. Automate knowledge concept tagging on math questions with llms. *arXiv preprint arXiv:2403.17281*.

Richard Wing Cheung Lui, Haoran Bai, Aiden Wen Yi Zhang, and Elvin Tsun Him Chu. 2024. Gptutor: A generative ai-powered intelligent tutoring system to support interactive learning with knowledge-grounded question answering. In *2024 International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pages 702–707. IEEE.

Stephen MacNeil, Paul Denny, Andrew Tran, Juho Leinonen, Seth Bernstein, Arto Hellas, Sami Sarsa, and Joanne Kim. 2024. Decoding logic errors: a comparative study on bug detection by students and large language models. In *Proceedings of the 26th Australasian Computing Education Conference*, pages 11–18.

Yacine Majdoub and Eya Ben Charrada. 2024. Debugging with open-source large language models: An evaluation. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 510–516.

Julia M Markel and Philip J Guo. 2021. Inside the mind of a cs undergraduate ta: A firsthand account of undergraduate peer tutoring in computer labs. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 502–508.

Samiha Marwan, Thomas W Price, Min Chi, and Tiffany Barnes. 2020. Immediate data-driven positive feedback increases engagement on programming homework for novices. In *CSEDM@ EDM*.

Atharva Mehta, Nipun Gupta, Aarav Balachandran, Dhruv Kumar, Pankaj Jalote, and 1 others. 2023. Can chatgpt play the role of a teaching assistant in an introductory programming course? *arXiv preprint arXiv:2312.07343*.

Marcus Messer, Neil CC Brown, Michael Kölling, and Miaojing Shi. 2023. Machine learning-based automated grading and feedback tools for programming: A meta-analysis. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, pages 491–497.

Kanav Mittal, Abigail O'Neill, Hanna Schlegel, Gireeja Ranade, and Narges Norouzi. 2025. Modeling student knowledge progression across concepts in intelligent tutoring interactions. In *International Conference on Artificial Intelligence in Education*, pages 105–118. Springer.

Steven Moore, Robin Schmucker, Tom Mitchell, and John Stamper. 2024. Automated generation and tagging of knowledge components from multiple-choice questions. In *Proceedings of the eleventh ACM conference on learning@ scale*, pages 122–133.

Rose Niousha, Abigail O'Neill, Ethan Chen, Vedansh Malhotra, Bita Akram, and Narges Norouzi. 2025. Llm-kci: Leveraging large language models to identify programming knowledge components. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2*, pages 1557–1558.

Priti Oli. 2024. *Towards automated scaffolding of learners' code comprehension process*. The University of Memphis.

Abby O'Neill, Samantha Smith, Aneesh Durai, John DeNero, JD Zamfirescu-Pereira, and Narges Norouzi. 2025. From code to concepts: Textbook-driven knowledge tracing with llms in cs1. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2*, pages 1565–1566.

Victor-Alexandru Padurean, Paul Denny, and Adish Singla. 2025. Bugspotter: Automated generation of code debugging exercises. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pages 896–902.

Maciej Pankiewicz and Ryan S. Baker. 2023. Large language models (gpt) for automating feedback on programming assignments. (arXiv:2307.00150). ArXiv:2307.00150.

Andrew Petersen, Michelle Craig, Jennifer Campbell, and Anya Tafliovich. 2016. Revisiting why students drop cs1. In *Proceedings of the 16th Koli Calling International Conference on Computing Education Research*, pages 71–80.

Griffin Pitts, Neha Rani, Weedguet Mildort, and Eva-Marie Cook. 2025. Students' reliance on ai in higher education: Identifying contributing factors. *arXiv preprint arXiv:2506.13845*.

Allison Poh, Anurata Hridi, Jordan Barria-Pineda, Peter Brusilovsky, and Bita Akram. 2025. Example explorers and persistent finishers: Exploring student practice behaviors in a python practice system. Proceedings of 9th Educational Data Mining in Computer Science Education.

Pavlin G Poličar, Martin Špendl, Tomaž Curk, and Blaž Zupan. 2025. Automated assignment grading with large language models: insights from a bioinformatics course. *Bioinformatics*, 41(Supplement_1):i21–i29.

Anna Řechtáčková, Alexandra Maximova, and Griffin Pitts. 2025. Finding misleading identifiers in novice code using llms. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2*, pages 1595–1596.

Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. Abscribe: Rapid exploration & organization of multiple writing variations in human-ai co-writing tasks using large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Kelly Rivers, Erik Harpstead, and Kenneth R Koedinger. 2016. Learning curve analysis for programming: Which concepts do students struggle with? In *ICER*, volume 16, pages 143–151. ACM.

Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM conference on international computing education research-volume 1*, pages 27–43.

Alexander Scarlatos, Ryan S. Baker, and Andrew Lan. 2025. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, page 249–259, Dublin Ireland. ACM.

Andre Silva, Sen Fang, and Martin Monperrus. 2025. Repairllama: Efficient representations and fine-tuned adapters for program repair. *IEEE Transactions on Software Engineering*, (01):1–16.

Priscylla Silva and Evandro Costa. 2025. Assessing large language models for automated feedback generation in learning programming problem solving. *arXiv preprint arXiv:2503.14630*.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard G Baraniuk. 2024. Pedagogical alignment of large language models. *arXiv preprint arXiv:2402.05000*.

En-Qi Tseng, Pei-Cing Huang, Chan Hsu, Peng-Yi Wu, Chan-Tung Ku, and Yihuang Kang. 2024. Codev: An automated grading framework leveraging large language models for consistent and constructive feedback. In *2024 IEEE International Conference on Big Data (BigData)*, pages 5442–5449. IEEE.

Alexandra Vassar, Jake Renzella, Emily Ross, and Andrew Taylor. 2024. Fine-tuning large language models for better programming error explanations. In *Proceedings of the 24th Koli Calling International Conference on Computing Education Research*, pages 1–2.

Yumou Wei, Paulo Carvalho, and John Stamper. 2025. Kcluster: An llm-based clustering approach to knowledge component discovery. (arXiv:2505.06469). ArXiv:2505.06469.

Jacqueline Whalley, Andrew Petersen, and Paul Denny. 2020. Mathematics, computer science and career inclinations—a multi-institutional exploration. In *Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, pages 1–10.

Weiqing Yang, Hanbin Wang, Zhenghao Liu, Xinze Li, Yukun Yan, Shuo Wang, Yu Gu, Minghe Yu, Zhiyuan Liu, and Ge Yu. 2025. COAST: Enhancing the code debugging ability of LLMs through communicative agent based data synthesis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2570–2585, Albuquerque, New Mexico. Association for Computational Linguistics.

Calvin Yeung, Jeff Yu, King Chau Cheung, Tat Wing Wong, Chun Man Chan, Kin Chi Wong, and Keisuke Fujii. 2025. A zero-shot llm framework for automatic assignment grading in higher education. (arXiv:2501.14305). ArXiv:2501.14305.

Mina Yousef, Kareem Mohamed, Walaa Medhat, Ensaf Hussein Mohamed, Ghada Khoriba, and Tamer Arafa. 2025. Begrading: large language models for enhanced feedback in programming education. *Neural Computing and Applications*, 37(2):1027–1040.

Jialu Zhang, José Pablo Cambronero, Sumit Gulwani, Vu Le, Ruzica Piskac, Gustavo Soares, and Gust Verbruggen. 2024. Pydex: Repairing bugs in introductory python assignments using llms. *Proceedings of the ACM on Programming Languages*, 8(OOPSLA1):1100–1124.

Li Zhong, Zilong Wang, and Jingbo Shang. 2024. Debug like a human: A large language model debugger via verifying runtime execution step by step. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 851–870, Bangkok, Thailand. Association for Computational Linguistics.

Fida Zubair, Maryam Al-Hitmi, and Cagatay Catal. 2025. The use of large language models for program repair. *Computer Standards & Interfaces*, 93:103951.

Toward Human-Centered Readability Evaluation

Bahar İlgen[1] and **Georges Hattab**[1,2]
[1] Center for Artificial Intelligence in Public Health Research (ZKI-PH),
Robert Koch Institute, Berlin, Germany
[2] Department of Mathematics and Computer Science,
Freie Universität Berlin, Berlin, Germany
{ilgenb, hattabg}@rki.de

Abstract

Text simplification is essential for making public health information accessible to diverse populations, including those with limited health literacy. However, commonly used evaluation metrics in Natural Language Processing (NLP)—such as BLEU, FKGL, and SARI—mainly capture surface-level features and fail to account for human-centered qualities like clarity, trustworthiness, tone, cultural relevance, and actionability. This limitation is particularly critical in high-stakes health contexts, where communication must be not only simple but also usable, respectful, and trustworthy. To address this gap, we propose the Human-Centered Readability Score (HCRS), a five-dimensional evaluation framework grounded in Human-Computer Interaction (HCI) and health communication research. HCRS integrates automatic measures with structured human feedback to capture the relational and contextual aspects of readability. We outline the framework, discuss its integration into participatory evaluation workflows, and present a protocol for empirical validation. This work aims to advance the evaluation of health text simplification beyond surface metrics, enabling NLP systems that align more closely with diverse users' needs, expectations, and lived experiences.

1 Introduction

Text simplification — the process of replacing complex terms with simpler alternatives, removing extraneous details, or breaking down lengthy sentences while preserving essential meaning (Chandrasekar et al., 1996; Saggion, 2017) — is especially critical in health communication. Yet contemporary evaluation metrics often focus on the outcomes of individual simplification operations and fail to fully capture human judgments of overall simplicity. This gap is particularly concerning in the medical domain, where effective information delivery is essential for public well-being, especially for individuals with limited health literacy (Nutbeam, 2000; McCormack et al., 2013). Health-related materials such as medication instructions, risk explanations, and care recommendations are often overly complex and cognitively demanding. Recent advances in natural language processing (NLP) have enabled automated systems to generate easier-to-read versions of such texts, aiming to improve accessibility and comprehension (Siddharthan, 2014; Espinosa-Zaragoza et al., 2023; Stajner, 2021).

These systems are typically evaluated with automatic metrics such as SARI (Xu et al., 2016), FKGL (Kincaid et al., 1975), or BLEU (Papineni et al., 2002), which focus on surface-level features (e.g., lexical simplicity, sentence length, or n-gram overlap). While useful for benchmarking, such metrics fail to capture whether simplified health information is genuinely clear, actionable, and trustworthy to real users—qualities that are critical in high-stakes domains like public health. This challenge mirrors broader efforts in AI to align system performance with human values, where evaluation must go beyond automatic scores to incorporate human judgment. Reinforcement Learning from Human Feedback (RLHF) (Chaudhari et al., 2025) enables models to learn directly from user preferences and expert evaluations, serving as a gold standard for aligning system behavior with human values. However, collecting large-scale, high-quality human feedback can be costly and impractical. To address this limitation, recent approaches such as Reinforcement Learning with AI Feedback (RLAIF) (Lee et al., 2024) offer scalable alternatives to costly human evaluations, but risk misalignment with human values and needs in sensitive contexts.

To address the gap between metric-driven evaluation and real-world user experience, we investigate:

Q1 Which existing readability metrics (lexi-

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 263–273
November 9, 2025

cal, syntactic, semantic, or neural) align most strongly with human-centered dimensions—clarity, trustworthiness, tone appropriateness, cultural relevance, and actionability—in simplified health texts?

Q2 To what extent can a composite Human-Centered Readability Score (HCRS), integrating automatic features with human feedback, achieve stronger alignment with user evaluations than the best standalone metric?

Q3 How can human–computer interaction (HCI) techniques, such as interactive feedback collection and participatory design, be embedded into the evaluation pipeline to make metric optimization more responsive for real-world needs?

Readability in health communication is inherently multidimensional: a linguistically simple message can still be unclear, culturally inappropriate, or untrustworthy. Addressing these dimensions requires going beyond surface-level metrics. Our contributions are threefold:

- **Redefinition:** Reconceptualizing readability for health text simplification from a human-centered perspective, drawing on HCI and health communication research.
- **Framework:** Proposing a five-dimension conceptual model—clarity, trustworthiness, tone appropriateness, cultural relevance, and actionability—that can guide evaluation and system design.
- **Agenda:** Arguing that these dimensions must inform both the evaluation and design of simplification systems, especially for vulnerable or marginalized audiences, and outlining a future research agenda bridging NLP and HCI.

By shifting the evaluation lens from system-centric to user-centric, we aim to advance health-focused NLP systems that are not only technically effective, but also socially and culturally responsive to the needs of diverse real-world users—an imperative in public health communication.

2 Background: Readability in NLP and Health Communication

2.1 Automatic Readability Metrics in NLP

Readability in NLP is often reduced to numerical scores, yet human perception of clarity, trust, and usability depends on far more than surface form. Automatic evaluation of text simplification in NLP typically falls into three categories: *surface-level metrics* (e.g., BLEU, SARI), *semantic metrics* (e.g., BERTScore, QuestEval, METEOR), and *readability indices* (e.g., FKGL, SMOG, Coleman–Liau), the latter often applied in health communication.

BLEU (Papineni et al., 2002) computes n-gram precision between outputs and references, penalizing brevity; **SARI** (Xu et al., 2016) measures the quality of n-gram "keep", "add", and "delete" operations; **FKGL** computes a linear combination of average sentence length and syllables per word to estimate grade level. Despite their widespread use in NLP benchmarking, these metrics neglect cognitive, emotional, and social dimensions central to human perception of readability—factors that are critical in high-stakes settings such as public health.

SARI: $SARI = \frac{1}{3}(F_{add} + F_{keep} + F_{del})$

BLEU: $BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$

FKGL: $FKGL = 0.39 \times \frac{\text{Words}}{\text{Sentences}} + 11.8 \times \frac{\text{Syllables}}{\text{Words}} - 15.59$

Recent neural and reference-based metrics, such as **BERTScore** (Zhang et al., 2020) (contextualized token similarity) and **QuestEval** (Scialom et al., 2021) (Q&A-based semantic evaluation), and **SALSA** (Heineman et al., 2023) (edit-based granular evaluation), aim to capture meaning and content transformation beyond surface forms. SALSA, in particular, provides a fine-grained typology of simplification edits and an automatic variant (LENS-SALSA) to score outputs without references, showing stronger alignment with human judgments in recent studies. While SALSA covers simplification edits in detail, it primarily captures structural and lexical changes and cannot fully assess whether the resulting text is clear to diverse users in real contexts. Yet, they still fail to assess *trustworthiness*, *emotional resonance*, or *context appropriateness*—qualities that determine whether a health message is actually usable. These limitations are also evident in large-scale meta-evaluations. For example, Alva-Manchego et al. (2021) found that these metrics often correlate only weakly with human judgments, especially for multi-operation simplifications where lexical, syntactic, and semantic changes co-occur. They systematically compared leading metrics across several simplification systems and found that, while BERTScore Precision achieved the highest overall correlation, performance dropped sharply for high-quality outputs, underscoring the persistent gap between metric scores and user-perceived quality. As shown in Table 1, no widely used automatic metric provides comprehensive coverage of the user-centered dimensions that are critical in health contexts—namely clarity, trustworthiness, and actionability.

2.2 Human Perception and the Readability Gap

Prior work in HCI and health communication (Ishikawa and Kiuchi, 2010; Crossley et al., 2016) underscores that readability is constructed through users' prior knowledge, affect, and social context, as well as deeper textual properties such as local and global cohesion. Consequently, leading researchers now advocate for *mixed-method* evaluation frameworks combining robust automatic metrics with validated user scales (e.g., Trust in Health Information Questionnaire, cognitive load indices) and real-time feedback. While such calls for mixed-method evaluation are compelling, there remains limited empirical evidence demonstrating how current metrics diverge from human perception in practice.

Empirical evidence confirms the gap between automatic scores and lived experience. In a controlled study, (Leroy et al., 2022) found that simplified health texts significantly improved comprehension accuracy—boosting correct recall from 33% to 59%—with the largest gains among participants with lower education or limited English proficiency. Earlier work (Leroy et al., 2013) similarly reported improvements in perceived clarity and learning outcomes. Taken together, these studies illustrate that surface-level gains in simplicity do not necessarily ensure that health information is perceived as accessible or actionable by diverse users. However, more recent NLP-based evaluations (e.g., (Alva-Manchego et al., 2021; Maddela et al., 2021)) show that even high-SARI outputs can be perceived as emotionally flat or insufficiently actionable, suggesting that surface simplicity does not guarantee effective communication.

In public health, *readability is relational*. A text must be clear, but also trustworthy, respectful, culturally sensitive, and actionable. This gap is both methodological and conceptual: current NLP evaluation practices rarely capture these interpersonal and contextual dimensions. For example, a jargon-free vaccine information sheet delivered in an emotionally cold or overly directive tone can still alienate its audience. These interpersonal and contextual factors are invisible to current NLP metrics, which rarely incorporate user studies or grounded sociocultural analysis.

This disconnect is consequential: in health contexts, misunderstanding can reduce adherence, increase anxiety, or even cause harm. We argue that evaluating readability in NLP must center on *user interpretation*—not just algorithmic scores—requiring new, multidimensional metrics co-designed with end users. In short, a truly "readable" health text is one that is *understood, trusted, respected, and acted upon*. Achieving this requires moving beyond current benchmarks toward evaluation frameworks that are co-designed with and validated by target user communities.

3 Limitations of Current Metrics

Despite widespread adoption, automatic evaluation metrics for text simplification—such as SARI, FKGL, and BLEU—face critical shortcomings in public health communication.

Intrinsic Metric Limitations. These metrics are designed to capture primarily surface-level transformations—such as lexical substitution, sentence compression, and n-gram overlap with reference texts—and are effective for system-level benchmarking. However, they provide no direct evidence regarding whether a simplified text is, in fact, clear, trustworthy, or pragmatically usable for its intended audience.

Overreliance on Reference-Based Evaluation. One major limitation is the overreliance on reference-based evaluation. Metrics like BLEU and SARI compare system outputs to one or more human-written references, assuming that overlap with these references implies higher quality. However, simplification is an inherently subjective task with high variance in valid outputs. A single sentence can be simplified in many plausible ways depending on the user's background knowledge, cultural expectations, or even emotional state. Penalizing deviations from a limited reference set risks excluding useful and contextually appropriate

Metric	Lexical Simplicity	Syntactic Simplicity	Semantic Adequacy	Clarity	Trustworthiness	Actionability
BLEU	✓	(✓)	–	–	–	–
SARI	✓	(✓)	–	–	–	–
FKGL	✓	✓	–	(✓)[1]	–	–
BERTScore	–	–	✓	–	–	–
QuestEval	–	–	✓	–	–	–
SALSA	✓	✓	–	(✓)[2]	–	–

Table 1: Coverage of key readability and user-centered dimensions by common automatic metrics. (✓): partially addresses; –: not addressed. [1] FKGL partially reflects clarity by rewarding shorter sentences and simpler words, but does not capture semantic or user-perceived clarity. [2] SALSA contributes to clarity by typologizing lexical and structural simplification edits, but does not directly assess whether outputs are clear to diverse users.

simplifications.

Neglect of Pragmatic and Relational Aspects. Another concern is the neglect of pragmatic and relational aspects of language use. Metrics such as FKGL reduce readability to sentence length and syllable count, ignoring tone, politeness, cultural sensitivity, and actionability. In health contexts, these dimensions are critical. A sentence that is syntactically simple but emotionally flat or overly authoritative may alienate readers or diminish trust in the message. Beyond these pragmatic concerns, relational aspects such as trust-building, respect, and perceived empathy are equally essential, yet remain invisible to current metrics, which operate without understanding user intent, affective response, or context of use.

Limited Generalization Across User Populations. Moreover, automatic metrics do not generalize well across diverse user populations. Individuals with different literacy levels, cultural backgrounds, or health conditions may interpret the same text in divergent ways. Metrics grounded in average-case assumptions fail to capture these differences, potentially reinforcing existing disparities in health communication. For example, a simplification that seems effective for English-dominant, college-educated users may confuse or offend speakers of other dialects or individuals with lower health literacy.

Risks from AI-Generated Feedback Loops. Finally, the increasing use of reinforcement learning with AI-generated feedback (RLAIF) raises additional concerns. When optimization is driven by synthetic evaluators trained on limited metrics or preferences, system behavior may diverge from human-centered values. Without human-in-the-loop validation, systems risk overfitting to numerical proxies that do not reflect lived experience or real-world comprehension.

Evidence from Recent Studies Recent research (Choi et al., 2024) shows that dynamically combining multiple evaluation metrics (lexical and semantic) yields much stronger alignment with human ratings than any single metric can. Large-scale meta-evaluations (Alva-Manchego et al., 2021) found that commonly used automatic metrics, such as BLEU and SARI, typically exhibit only low-to-moderate correlation with human judgments on simplification quality—particularly when multiple rewriting operations are involved. These results underscore the risk of relying solely on surface-level scores, as high metric values may not reflect true gains in user-perceived clarity or simplicity.

Overall, these limitations highlight the need for evaluation frameworks that move beyond narrow linguistic proxies to capture the relational, pragmatic, and affective dimensions of readability in health-focused NLP. Building on this, we argue for extending ensemble approaches with participatory user evaluation to address these gaps.

4 Human-Centered Readability Score (HCRS)

We argue that readability in health communication should be redefined not as a property of the text alone, but as a dynamic, relational, and context-sensitive experience shaped by human perception. To address the limitations of current surface-level metrics, we introduce a five-dimensional framework—*clarity, trustworthiness, tone appropriateness, cultural relevance, and actionability*—grounded in research from health communication, HCI, and participatory design. Each dimension captures a distinct yet interdependent aspect of how users interpret and engage with simplified content. Figure 1 illustrates the five core dimensions and the integration of automatic and human evaluation within the hybrid HCRS protocol.

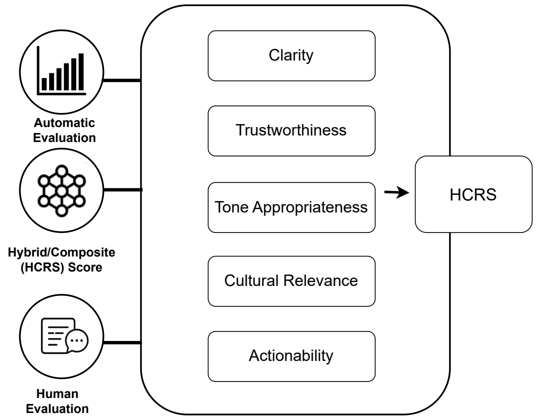


Figure 1: HCRS framework diagram

4.1 Clarity

The core dimension of readability is whether a text can be readily understood by its intended audience. This involves removing jargon, reducing syntactic complexity, and ensuring a logical flow of ideas. Crucially, clarity must be defined from the user's perspective: a linguistically simple sentence can still be unclear if it lacks context, relies on unfamiliar metaphors, or omits essential background information. In health communication, clarity should be measured not solely by word and sentence length, but by the degree to which users can accurately and confidently extract the intended meaning. Clarity measurement is automatically scored via readability indices (FKGL, SMOG), jargon detection, and cohesion tools (e.g., Coh-Metrix (Graesser et al., 2004)), combined with user-rated comprehension and ease-of-reading survey items on a 5-point Likert scale or direct comprehension quizzes.

4.2 Trustworthiness

In the context of health communication, *trustworthiness* refers to the perceived reliability, credibility, and transparency of the source, distinct from the broader and often contested notion of "trust" in general AI ethics debates. Health-related texts are not merely informational—they are also relational. Users often evaluate not just what is said, but who is saying it and how it is said. Trust is influenced by perceived author credibility, transparency, and tone. Simplified texts that are overly generic, impersonal, or dismissive of complexity may erode trust, particularly among marginalized populations with historical reasons to be skeptical of medical authority. A readable text should convey not only facts, but also empathy and accountability. Broader debates on "trust" in AI are addressed in the Discussion, but here the focus remains on relational trust in patient-facing health communication.

Trustworthiness is quantified by detecting explicit source attribution, institutional language, transparency features, and domain authority markers in the text; supplemented by participant ratings of perceived credibility, transparency, and author reliability gathered from validated survey instruments.

4.3 Tone Appropriateness

The emotional tone of a message can profoundly affect how it is received (Street et al., 2009; Hinyard and Kreuter, 2007; Arora, 2003). In simplified texts, tone often shifts unintentionally—becoming condescending, overly directive, or emotionally flat. Especially in health contexts, tone must balance clarity with compassion, authority with humility. An appropriate tone affirms the reader's dignity, avoids blame, and fosters a sense of collaboration rather than control. Tone appropriateness is defined as the alignment of a text's affective, pragmatic, and interpersonal cues with the expectations and sensitivities of the target user group, especially in health communication contexts. Technically, tone can be operationalized using multidimensional features derived from both automatic computational models and structured human assessments:

Automatic Assessment

- **Pragmatic Feature Extraction:** Quantify politeness strategies using established classifiers (e.g., Stanford Politeness Classifier, formality index). Key features include indirectness, mitigation (e.g., "could you" instead of imperatives), and hedging (e.g., "perhaps", "might") (Danescu-Niculescu-Mizil et al., 2013).

- **Sentiment and Emotion Analysis:** Compute sentiment polarity (positive, neutral, negative) and emotional valence using transformer-based models (e.g., BERT, RoBERTa) (Devlin et al., 2019; Liu et al., 2019; Mohammad and Turney, 2013) fine-tuned for affective state detection.

- **Empathy & Support Classifiers:** Use models trained to detect empathy, warmth, and nonjudgmental language (e.g., EmpathBERT) to assess supportive tone (Sharma et al., 2020; Guda et al., 2021).

- **Lexical Diversity & Intensity:** Calculate frequencies of intensifiers (e.g., "very", "extremely"), modals (e.g., "should", "must"), evidentials (e.g., "it seems"), and negative polarity items (e.g., "never", "cannot") that may indicate directive or controlling language (Biber et al., 1999).

Human-Centered Evaluation

- **Likert-Scale Survey Items:** Collect ratings on standardized questions such as: "This message feels respectful and supportive." or "The tone is appropriate for the intended audience," following validated health communication scales.

- **Annotation Protocols:** Engage expert or target population annotators using detailed codebooks specifying tone-related phenomena (e.g., respectfulness, blame avoidance, collaborative framing) to ensure consistent evaluation.

The tone appropriateness for each text can then be computed as a weighted hybrid score:

$$\text{Tone}_{\text{HCRS}} = \alpha_1 P_{\text{a}} + \alpha_2 S_{\text{a}} + \alpha_3 E_{\text{a}} + \beta H \quad (1)$$

where P_{a}: politeness score (auto), S_{a}: sentiment score (auto), E_{a}: empathy score (auto), H: human Likert rating and the α and β coefficients are determined empirically via calibration on validation data to maximize alignment with user perceptions.

4.4 Cultural Relevance

Cultural relevance refers to the extent to which a simplified text preserves, reflects, and respects the cultural, linguistic, and social norms of its intended audience (Resnicow et al., 1999; Kreuter and McClure, 2004; Osborne, 2006). Texts can embed cultural meaning through references, metaphors, idioms, visual symbols, and formatting conventions; these elements may facilitate comprehension for in-group readers but create barriers for out-group readers.

From an evaluation standpoint, cultural relevance can be operationalized via (i) *automatic detection* of culturally specific lexical items, named entities, and idiomatic expressions, combined with cross-linguistic/multilingual embedding similarity measures to check alignment with target-culture corpora; and (ii) *human assessment* using Likert-scale items measuring perceived familiarity, inclusivity, and absence of culturally alienating content. Loss of culturally meaningful content or introduction of inappropriate cultural markers during simplification can reduce both accessibility and trust. Accordingly, incorporating cultural-perspective checks into simplification system design and evaluation is essential to ensure inclusivity and equity.

$$\begin{aligned}\text{Culture}_{\text{HCRS}} = \gamma_1 \cdot E_a \; + \; \gamma_2 \cdot I_a \\ + \; \gamma_3 \cdot M_a \; + \; \delta \cdot H_h \end{aligned} \quad (2)$$

where: E_a: automatic entity match score (NER-based), I_a: automatic idiom/cultural expression match score, M_a: automatic multilingual embedding similarity score, H_h: human-rated cultural relevance (Likert), $\gamma_1, \gamma_2, \gamma_3, \delta$: weights calibrated on validation set.

4.5 Actionability

In the final dimension, we address the need for simplified health texts to support informed action. This includes not only understanding a message, but knowing what steps to take and feeling empowered to take them. Actionability requires that information be specific, time-relevant, and contextually grounded in the user's lived reality. Inadequate or ambiguous directives—e.g., "seek care if needed"—can confuse rather than guide. A readable text should reduce cognitive load while increasing behavioral clarity.

From an evaluation perspective, actionability can be assessed via (i) automatic analysis of directive and procedural language (imperatives, explicit instructions, temporal/agent specification) and (ii) human ratings on validated Likert-scale items capturing perceived clarity of next steps. These scores can be integrated into a hybrid metric within the HCRS framework (Vishnevetsky et al., 2018; Kreuter and McClure, 2004).

Operationally, actionability can be measured through (i) *automatic linguistic analysis*, such as the detection and scoring of imperative verbs, procedural language, and explicit guidance framing; and (ii) *human rating* via Likert-scale items assessing whether the reader feels well-informed and able to act on the information provided.

$$\text{Action}_{\text{HCRS}} = \lambda_1 \cdot D_a \;+\; \lambda_2 \cdot P_a \;+\; \lambda_3 \cdot Q_a \;+\; \mu \cdot H_h \qquad (3)$$

where:
D_a: automatic directive language/imperative score,
P_a: automatic procedural/instruction cue score,
Q_a: automatic presence of action-associated entities (temporal, agent, location),
H_h: human rating of perceived actionability (Likert),
$\lambda_1, \lambda_2, \lambda_3, \mu$: weights calibrated on validation set.

By framing readability as a multidimensional construct, we aim to support the development of NLP systems that are better aligned with human needs and social context. Each of these dimensions is measurable, at least in principle, through user-centered evaluation methods such as interviews, surveys, and participatory testing. Importantly, these dimensions are not intended to be exhaustive or mutually exclusive, but to offer a starting point for rethinking evaluation as an interdisciplinary, collaborative process.

Figures 2 and 3 illustrate the HCRS framework in practice. Figure 2 compares original and simplified health texts across the five human-centered dimensions, with version B scoring higher in trust, tone, and actionability. Figure 3 shows how such scores can be integrated into a human-in-the-loop evaluation pipeline, linking user feedback to model updates.

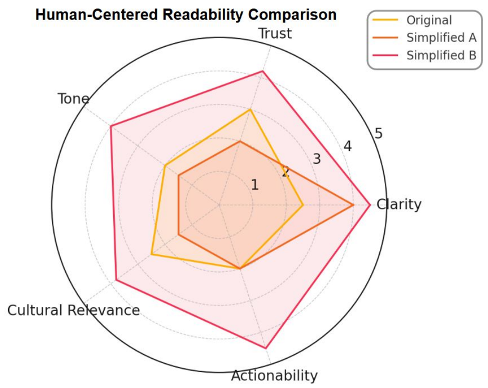


Figure 2: Illustrative comparison of original and simplified versions across five human-centered readability dimensions. Simplified B scores higher on trust, tone, and actionability, reflecting better alignment with user-centered design principles.

5 Empirical Evaluation & Discussion

This work defines a comprehensive experimental protocol to validate the Human-Centered Readability Score (HCRS) across diverse user populations and health literacy levels. The protocol integrates automatic and human-centered measures to move evaluation beyond current surface-oriented benchmarks.

Interactive & Participatory Evaluation. Most NLP evaluation workflows are system-facing. To operationalize HCRS dimensions such as tone, trust, and actionability, user-facing interfaces should collect real-time feedback. Short prompts—e.g., “Was this sentence clear to you?”, “Did you feel respected by the way it was phrased?”—can be embedded in studies or deployment. Inclusive design, informed by participatory methods, engages diverse stakeholders, co-creates evaluation criteria, and adapts interfaces to cultural and literacy contexts. Accessibility features (visual icons, read-aloud options) can broaden participation, and lightweight web tools or crowdsourcing can scale data collection.

In practice, participatory methods can be embedded in evaluation pipelines through lightweight annotation interfaces where end-users quickly rate sentences on the five HCRS dimensions. Feedback can be collected in short micro-surveys (5–10 minutes), aggregated, and then reviewed with stakeholders (e.g., patients, clinicians, or domain experts) during participatory workshops. This creates an iterative loop: (i) immediate micro-ratings during evaluation, (ii) stakeholder workshops to review and refine criteria, and (iii) iterative updates where refinements directly inform HCRS weighting. For instance, in a vaccine information use-case, patients could rate clarity and trustworthiness via inline sliders, while health professionals review aggregated outputs to recalibrate actionability guidelines. This illustrates how participatory methods can be concretely integrated rather than remaining abstract, serving as a proof-of-concept protocol that demonstrates how HCRS could be operationalized through lightweight user studies even before large-scale deployment.

Structured, Multidimensional Feedback. Binary ratings rarely explain why a sentence fails. Feedback channels mapped to HCRS dimensions—e.g., “Too technical”, “Missing information”, “Poorly structured”—enable more interpretable model training and closer alignment with user perceptions.

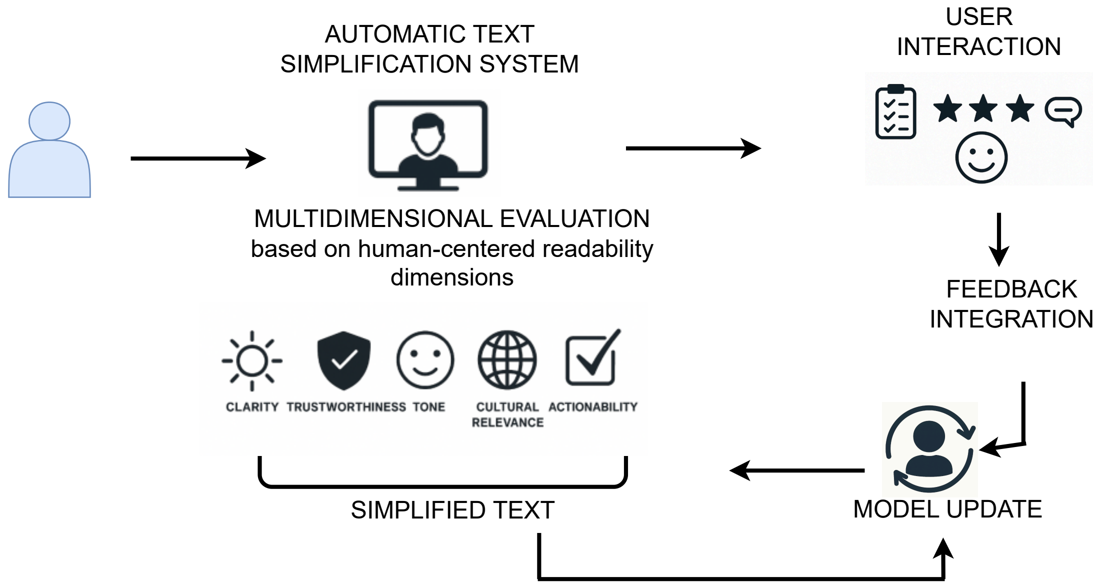


Figure 3: Human-in-the-loop readability evaluation

Model Integration. Human feedback should inform training, not just evaluation. Strategies include multi-objective learning, auxiliary classifiers for HCRS dimensions, and prompt engineering for targeted qualities. Care is needed with RLAIF, as over-reliance on synthetic evaluators risks misalignment with human values in sensitive domains.

Future Directions. Future work will extend validation of HCRS beyond health communication, test participatory feedback pipelines at scale, and explore adaptive dimension weighting.

RLAIF and Human-Centered Alignment in Health Communication. Recent advances in RL-based model alignment, particularly Reinforcement Learning from AI Feedback (RLAIF (Lee et al., 2024), offer a major leap in scalability and efficiency. By using large language models (LLMs) to generate preference labels, RLAIF reduces the cost and time of annotation by more than an order of magnitude, while achieving competitive—and sometimes superior—results compared to traditional RLHF on standard benchmarks such as win rate and harmlessness.

Yet in domains like health communication, where trust, contextual sensitivity, emotional nuance, and sociocultural fit are essential, RLAIF faces critical limitations. Its optimization targets remain generic (e.g., helpfulness, harmlessness) and do not inherently capture the multidimensional, relational qualities required for impactful health information. Moreover, by inheriting preferences from pre-trained LLMs, RLAIF risks amplifying existing biases and overlooking authentic user values—an especially acute risk in high-stakes public health contexts. Finally, while RLAIF advances technical alignment, it does not natively support participatory, user-facing feedback loops or adaptive objectives such as trustworthiness, cultural relevance, and actionability.

The HCRS framework directly addresses these gaps. It moves beyond surface-level alignment toward user-driven, context-sensitive evaluation, integrates structured human feedback with participatory design, and operationalizes dimensions that matter in practice—clarity, trustworthiness, tone, cultural fit, and informed action—none of which current RLAIF pipelines explicitly model. In this way, HCRS can serve as both a complementary evaluation layer and a calibration signal for RLAIF-based training in sensitive domains.

6 Conclusion

This work introduced the Human-Centered Readability Score (HCRS), a five-dimensional framework for evaluating simplified health texts beyond surface features. By addressing clarity, trustworthiness, tone, cultural relevance, and actionability, HCRS fills critical gaps in current metrics and aligns evaluation with real-world user needs. We examined how existing metrics relate to these dimensions (Q1), proposed a composite score integrating automatic and human feedback (Q2), and outlined mechanisms to embed participatory, interactive evaluation into NLP workflows (Q3). Our empirical protocol sets the stage for validating HCRS across diverse populations and domains, enabling models that are both technically robust and socially responsive.

While designed for public health communication, the approach extends to any domain where clarity, trust, and usability are paramount. Importantly, as discussed in Section 5, HCRS can also serve as a complementary evaluation layer and a calibration signal for RLAIF-based training pipelines in sensitive domains—helping to align scalable RL methods with nuanced, human-centered objectives.

7 Limitations and Future Work

While the proposed HCRS framework outlines a path toward more human-centered evaluation, this work has several limitations. First, the framework has not yet been empirically validated on large-scale, diverse user populations, so its generalizability remains to be tested. A key limitation of this work is therefore the lack of empirical validation. While we outline protocols and scenarios for participatory evaluation, we were not able to conduct a pilot study within the scope of this paper. Future work will therefore prioritize small-scale validation studies using micro-surveys and participatory workshops on real health communication materials. Second, the weighting of dimensions is currently conceptual and requires calibration against real-world user judgments. Third, operationalizing sociocultural and emotional dimensions relies on language resources and annotation protocols that may be domain- or language-specific.

We also note that implementing all five dimensions in real evaluation settings may be resource-intensive. Some dimensions (e.g., clarity) can be partly automated, whereas others (e.g., trust or cultural relevance) require structured human feedback. Exploring hybrid setups that balance automation with targeted user input will be critical for feasibility.

Key directions for future evaluation, therefore, include: (i) Combining automatic metrics with validated user feedback instruments; (ii) Integrating participatory design into simplification evaluation; and (iii) Extending assessment to include sociocultural and emotional dimensions; and (iv) Exploring how HCRS can be integrated into existing NLP evaluation pipelines in a practical way, for example by combining automatic readability features with lightweight human feedback modules; (v) Conducting pilot user studies to provide proof-of-concept validation of the HCRS framework in practice. These directions will be essential for validating the HCRS framework in practice and expanding its applicability across domains where clarity, trust, and usability are critical.

While the HCRS framework was tailored for high-stakes public health communication, its core dimensions may require adaptation before use in other domains. Concepts such as trustworthiness or actionability are context-dependent and may need to be redefined based on the domain's communicative norms and user expectations. Furthermore, calibrated human feedback, participatory evaluation, and relevant language resources would need to be retuned for new target populations. The framework should therefore be seen as a conceptual starting point, with significant work required to ensure generalizability, validity, and relevance outside of health contexts.

Acknowledgments

This work was financially supported by the German Federal Ministry of Health (BMG) under grant No. ZMI5-2523GHP027. The project, titled "Strengthening National Immunization Technical Advisory Groups and Their Evidence-Based Decision-Making in the WHO European Region and Globally" (SENSE), is part of the Global Health Protection Programme (GHPP).

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Neeraj K. Arora. 2003. Interacting with cancer patients: The significance of physicians' communication behavior. *Social Science & Medicine*, 57(5):791–806.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow, Essex, England.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996), Volume 2*, pages 1041–1044, Copenhagen, Denmark. Association for Computational Linguistics.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2025. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *ACM Comput. Surv.*, 58(2).

Jason Ingyu Choi, Marcus Collins, Eugene Agichtein, Oleg Rokhlenko, and Shervin Malmasi. 2024. Combining multiple metrics for evaluating retrieval-augmented conversations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 40–50, Mexico City, Mexico. Association for Computational Linguistics.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Paloma Moreda, and Manuel Palomar. 2023. Automatic text simplification for people with cognitive disabilities: Resource creation within the ClearText project. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 68–77, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. EmpathBERT: A BERT-based framework for demographic-aware empathy prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3072–3079, Online. Association for Computational Linguistics.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using salsa. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3466–3495, Singapore. Association for Computational Linguistics.

Leslie J. Hinyard and Matthew W. Kreuter. 2007. Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior*, 34(5):777–792.

Hirono Ishikawa and Takahiro Kiuchi. 2010. Health literacy and health communication. *BioPsychoSocial Medicine*, 4:18.

J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report Research Branch Report 8-75, Institute for Simulation and Training, University of Central Florida, Millington, TN.

Matthew W. Kreuter and Stephanie M. McClure. 2004. The role of culture in health communication. *Annual Review of Public Health*, 25:439–455.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *Preprint*, arXiv:2309.00267.

Gondy Leroy, James E Endicott, David Kauchak, Obay Mouradi, and Melissa Just. 2013. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of Medical Internet Research*, 15(7).

Gondy Leroy, David Kauchak, Diane Haeger, and Douglas Spegman. 2022. Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty. *JAMIA Open*, 5(2):ooac044.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3536–3553, Online. Association for Computational Linguistics.

Lauren McCormack, Jolie Haun, Kristine Sørensen, and Melissa Valerio. 2013. Recommendations for advancing health literacy measurement. *Journal of Health Communication*, 18(sup1):9–14.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Don Nutbeam. 2000. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health Promotion International*, 15(3):259–267.

Helen Osborne. 2006. Health literacy: how visuals can help tell the healthcare story. *Journal of Visual Communication in Medicine*, 29(1):28–32. Published online: 10 July 2009.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ken Resnicow, Tom Baranowski, Jasjit S. Ahluwalia, and Ronald L. Braithwaite. 1999. Cultural sensitivity in public health: defined and demystified. *Ethnicity & Disease*, 9(1):10–21. 12 pages.

Horacio Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Springer Cham.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL – International Journal of Applied Linguistics*, 165(2):259–298.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Richard L. Street, Gregory Makoul, Neeraj K. Arora, and Ronald M. Epstein. 2009. How does communication heal? pathways linking clinician–patient communication to health outcomes. *Patient Education and Counseling*, 74(3):295–301. Theories in Health Communication Research.

Julia Vishnevetsky, Chasity Burrows Walters, and Kay See Tan. 2018. Interrater reliability of the patient education materials assessment tool (pemat). *Patient Education and Counseling*, 101(3):490–496. Epub 2017 Sep 6.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Exploring Gender Differences in Emoji Usage: Implications for Human-Computer Interaction

Zi-Xiang Lin
Kang Chiao International School
{zixiang.ryan.lin@gmail.com}

Abstract

This study discusses the emojis employment that compensate for the absence of supralinguistic emotive cues in digital communication. Analyzing gender relations (Male-to-Male, Male-to-Female, Female-to-Male, Female-to-Female) as a social influence factor in emoji use, the research explores the use of anger-related emojis (😠, 😡, 🤬) and their dual functions as emotion signals and intensifiers. Findings reveal women use more intense emojis toward men and less severe ones toward women, a pattern not observed in men when emphasizing emotions. Hence, the study contributes to the conceptual application of emotional expression via emojis within digital media, raising sentiments on gender variances and improving emotional intelligence in artificial intelligence systems to yield a more accurate human feeling interpretation.

1 Introduction

In everyday life, humans can engage directly with one another in a variety of ways, including gestures, facial expressions, and tone of conversation. However, based on current Internet growth, online interactions lack supralinguistic communication mechanisms, resulting in communication challenges between communicators because language expression alone cannot accurately convey the entire mood (Li and Yang, 2018). Thus, emojis, the small digital pictures or graphic symbols, were born to represent things, feelings, or concepts in communication software. In fact, the rise of emojis has changed the way Internet users communicate, which has not only affected the attention of linguists to this issue but has also led to new changes in the way Internet users communicate. Lupyan and Dale (2016) reported an increase of 30% to 40% in emojis on Instagram, a type of communication software, compared to two to three years earlier, demonstrating the importance of the role of emojis in online dialogue.

This research analyzes the role of the "emotion signal emoji" and the "emotion intensifier emoji" summarized in Li and Yang (2018). Specifically, "emotion signal emoji" refers to the emoji that describes the emotion not previously mentioned, while "emotional intensifier emoji" refers to the emoji that characterizes the mentioned emotion (usually contains direct emotional words).

Previously, Butterworth et al. (2019) revealed the differences in the cognitive expressions of emojis by gender, while Herring and Dainas (2020) found gender differences in emoji employment. Specifically, females use emojis more frequently than males (Kennison et al., 2025). To investigate the gender variances, this study examines deeper at gender (male and female) and emoji usage interaction under angry scenarios for Taiwanese participants. The present study evaluates four gender relationships—Male-to-Male (*MtoM*), Male-to-Female (*MtoF*), Female-to-Male (*FtoM*), and Female-to-Female (*FtoF*)—to explore the following four research questions:

(1) Under the emotion signal scenario, do men and women send different intensities of angry emojis depending on the recipient's gender?

(2) Under the emotion intensifier scenario, do men and women send different intensities of angry emojis depending on the recipient's gender?

(3) Comparing both emotion signal and intensifier scenarios, do men send different intensities of angry emojis? What about female users?

(4) Is the use of angry emojis of men and women affected by the syntax (affirmative and interrogative)?

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 274–282
November 9, 2025 ©2025 Association for Computational Linguistics

Findings reveal women use stronger emojis when communicating with men and less intense ones with women, potentially because women are more emotional and sensitive to emotions compared to men. As emojis become increasingly prevalent in present-day communication, this study may aid in the analysis of human-computer alignments. The results may benefit future investments in resolving emotional misalignments that diminish trust in the medical chatbots.

This paper is organized as follows: ***Literature Review*** presents past emoji studies and human bias in human interactions; ***Online Survey*** outlines the survey procedure and questionnaire; ***Results and Discussions*** addresses the research questions and presents plausible explanations; and ***Conclusion*** summarizes key findings and suggestions for future research.

2 Literature Review

2.1 Functions of Emoji

With different functions of emojis, Li and Yang (2018) simplified the seven emoji functions proposed by Yus (2014): "illocutionary force modifier," "turn giving/taking," "emotion/attitude signal," "irony," "emotion/attitude intensity enhancer," "backchannel device," and "humor." To investigate emotions, the present study selects "emotion signal emoji"[1] and "emotion intensifier emoji,"[2] since both serve to convey the speakers' emotions.

2.2 Emotion Expression in Chinese

Studies have shown that cultural conditions affect participants' expressions of emotion and complaints (De Vaus et al., 2018; Fischer et al., 2004; Lim, 2016). Wu (2013) examined Hakka-speaking men's and women's complaint behavior, discovering that female Hakka speakers, a branch of Chinese, are more polite and sensitive, highlighting the significance of gender influence in angry communication. Moreover, Yu (2005) and Chen et al. (2011) both investigated complimenting acts in Mandarin and American English, finding differences in cultural influences on human strategies to express anger in dialogue. Specifically, Yu (2005) discovered that Chinese participants prefer to indirectly express compliments as American English participants are more likely to speak out directly. Furthermore, Chen et al. (2011) reveal cultural effects on two groups of participants' strategies of complaining, as Americans expressed complaints across all the situations, while Taiwanese participants, another branch of Chinese culture, are sensitive to social power and will carefully choose proper expression in situations.

2.3 Online Cross-gender Communication

Butterworth et al. (2019) investigated the impact and correlation between Internet users' gender and the recipient's gender. Using a Likert scale, the researchers asked 40 men and 39 women about their opinions on four different workplace emoji usage scenarios: female send to male (*FtoM*), female send to female (*FtoF*), male send to male (*MtoM*), and male send to female (*MtoF*). The study shows that people's perceptions of the sender and the message are influenced by their use of emojis, in addition to their gender. It's important to note that these results support conventional gender preconceptions in communication, highlighting the social gender effects on people's opinions that may influence their choice of emoji. Accordingly, the study implies that gender is related to emoji selection.

3 Online Survey

3.1 Questionnaire Collecting

3.1.1 Participants

The study recruited 36 men and 46 women (aged 11 to 20 years old). Herring and Dainas (2020) studied participants from 18 to 70+ and mentioned that older people over 30 generally do not understand the meaning and application of emojis. Given that the main group of people who use Internet software to communicate in modern times are teenagers, this study only analyzed the responses of respondents aged between 11 and 20 years old. To reduce potential harm to underage participants, they are required to finish the questionnaire under their guardian's supervision. In addition, they attended school in Taiwan (New Taipei City, Taipei City, and Keelung City) and their mother tongue was Taiwanese Mandarin (Traditional Chinese) before the age of seven.

[1] This study modifies the "emotion signal emoji" from Li and Yang (2018) "attitude/emotion signal" to specify the emotion in the emoji.

[2] This study modifies the "emotion intensifier emoji" from Li and Yang (2018) "attitude/emotion intensity enhancer" to simplify the name and specify the emotion in the emoji.

Respondents who did not fill in "Traditional Chinese" as their native language will be considered as not familiar with Traditional Chinese applications and will not be included in the analysis.

3.1.2 Questionnaire Procedure

The survey was conducted online anonymously from 0:00 on April 26, 2023, to 23:59 on May 9, 2023, for a period of two weeks. Participants must first read a brief test description, including the purpose of the study, research process, confidentiality, and potential risks. Next, the subjects were asked to provide basic identity information, including age, native language, and gender, to verify their eligibility to be interviewed. Finally, the subjects were asked to answer a questionnaire that should require less than five minutes to finish.

3.2 Questionnaire Questions

This study will be analyzed in two parts in Traditional Chinese. First, this study refers to the research framework of Butterworth et al. (2019) and conducts a 2 (Subjects) ╳ 2 (Recipients) gender cross-comparison to investigate the difference between men and women using emotion signal emoji and emotion intensifier emoji.

Referring to Li and Yang (2018), this research selected emotion signal and emotion intensifier emojis as they are relatively direct conversational sentence patterns in daily life and express emotion, as illustrated in (5) and (6).

(5) Emotion Signal Emoji: I have been waiting for you online for two hours. You are still offline_(emoji)_.

(6) Emotion Intensifier Emoji: I've been waiting for you online for two hours. I "hate" people like you who make others wait_(emoji)_.

Therefore, the questions in the first part will be in the form of affirmative sentences, with a total of four situations (the participants will play the role of the sender and will choose an emoji to send messages to recipients of both genders): emotion signal emoji send to male, emotion signal emoji send to female, emotion intensifier emoji send to male, and emotion intensifier emoji send to male.

In the second part, this study aims to examine the use of angry emojis by men and women in affirmative and interrogative sentences. Due to the lack of previous literature, this study will initially explore whether syntax causes differences in the use of emojis between genders. Therefore, the questions will not be distinguished as being sent to males or females but will investigate men's and women's emoji usage in affirmative and interrogative sentences.

In addition, according to the Unicode Standard, a system for organizing and encoding text across most platforms, the emojis that include anger were selected for this study: 😠 (Angry Face/ Angry/ U+1F620), 😡 (Enraged Face/ Burning with rage/ U+1F621), 🤬 (Face with Symbols on Mouth/ Angry enough to curse/ U+1F92C) (Unicode, 2022). Participants can choose from the three options, with the first part of the research questionnaire also providing a "do not use emoji" option in case the subjects will not use emojis in this situation.

3.3 Questionnaire Design

3.3.1 Semantic and Contextual Relevance

Different situations can affect how humans express their emotions. Therefore, this questionnaire limits the scenarios to three groups, each with six questions (two emotion signals, two emotion intensifiers, and two syntax). In addition, even in the same context, the way a sentence is delivered can affect how the listener feels or the tone conveyed. The two sentences "我真的很生氣😠" (I am really angry 😠) and "我超生氣" (I am super angry 😠) may cause the tone of the former to be less intense than the latter, resulting in different corresponding emojis. To reduce the influence of semantics, this study limits the use of degree, including "很" (very), "真的" (really), "超" (super), etc., words in similar sentences.

3.3.2 Numbers, Word Count, and Fillers

The varying lengths of sentences may place a burden on the reader's short-term memory, so the number of words per sentence and the number of sentences (the number of commas and periods) per question in the questionnaire were controlled. For example, in Scenario 1, each question has 2 sentences and 26 words. Finally, to further reduce the phenomenon of transfer, in which participants will carry their emotions to the next question, the questionnaire's order was shuffled. In addition, this study designed filler sentences (⅓ of the total number of valid questions). Among the research questions, there are 18 valid questions, including 6 filler questions, and the whole questionnaire has a total of 24 questions.

3.3.3 Affirmative and Interrogative

To explore the influence of syntax, the second part of the questionnaire did not reveal gender, allowing the subjects to focus on the use of sentence patterns. In addition, the study used “?” to distinguish between affirmative sentences (without “?”) and interrogative sentences (with “?”).

4 Results and Discussions

4.1 Statistics and Data Processing

In the first part of the questionnaire, this study received 432 valid responses from males and 552 valid responses from females. In the second part of the questionnaire, this study received 216 valid responses from males and 276 valid responses from females. The study excludes invalid responses such that the participants do not complete the whole questionnaire or select multiple emojis for the same question. Since the response numbers differ in gender, the present study converts the number of responses into percentages. To discuss statistical significance, this study used the chi-square test with the significance level set to 0.1. If the significance value is lower than 0.1, it will be considered a significant difference or change.

4.2 Emotion Signal Emoji

4.2.1 Frequency of Emotion Signal Emojis

There was no significant difference in the use of emojis in emotion signal emoji sentences between men and women (p = 0.2225). From the overall distribution, the result shows the most popular choice among both men and women is not to use emojis (M=38.43%, F=39.13%), followed by 😡 (M=23.15%, F=25.00%).

Past research has shown that in negative circumstances, individuals don’t feel compelled to communicate their feelings. (Derks et al., 2007). Therefore, in the anger emotion of this study, it is reasonable to explain that the majority of subjects chose not to use emojis. In addition, 😡 remains the second most frequently used word is largely due to the compromise method. When encountering three

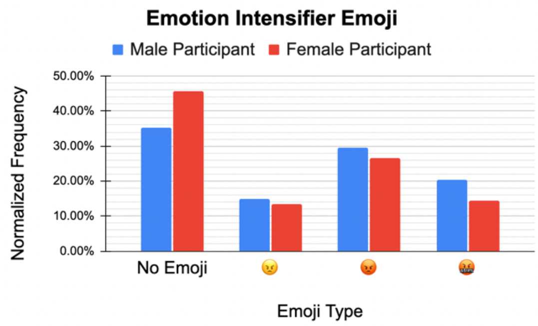


Figure 1: Male and Female Emotion Intensifier Emoji Responds Frequency.

levels of emoji intensities, the subjects generally believe that 😠 cannot fully express anger, while 🤬 over-expresses emotions (Unicode, 2022). An interesting observation is that females use 😠 (M=17.59%, F=21.74%) more frequently than 🤬 (M=20.83%, F=14.13%), while the opposite is true for males. This study justifies this phenomenon since women are more inclined to use strategies to weaken emotions. Detailed past literature and reasons will be discussed in “4.3 Emotion Intensifier Emoji.”

4.2.2 Recipient Gender in Emotion Signal

Within each gender, there was no significant difference in the results whether sent to male or female recipients (M: p = 0.6389; F: p = 0.8409). Overall, both men and women used emotion signal emojis similarly. Additionally, neither men nor women changed their choice of emojis based on the recipient’s gender.

4.3 Emotion Intensifier Emoji

4.3.1 Frequency of Emotion Intensifier Emojis

There was a significant difference between males and females in the use of emotion intensifier emoji sentences ($p < .1$)[3]. As shown in Figure 1, unlike the emotion signal emoji, in the emotion intensifier emoji, both men and women tend to use 🤬 rather than 😠. In addition, females were more likely to choose not to use emojis than males (M=35.19%, F=45.65%), while males were more likely to use 🤬 (M=20.37%, F=14.49%). This implies that women generally prefer to use fewer intense emojis (do not use emojis), while men are relatively accustomed to using more intense emojis. In this regard, in emotion intensifier emoji sentences,

[3] p-value = 0.09721.

women have strategies to weaken anger compared to men. As shown in Figure 1, the proportion of emoji usage frequency is arranged from most to least as follows: no emoji, 😡, 🤬, 😠.

4.3.2 Recipient Gender in Emotion Intensifier

There was no significant difference in the choice of emojis by male subjects when sending messages to the two genders (p = 0.9713). The number of emojis used has the same trend as the emotion signal emoji, from the most to the least: no emojis (*MtoM*=35.19%, *MtoF*=35.19%), 😡 (*MtoM*=29.63%, *MtoF*=29.63%), 🤬 (*MtoM*=19.44%, *MtoF*=21.30%), 😠 (*MtoM*=15.74%, *MtoF*=13.89%).

Figure 2 shows that, in contrast to the previous results, there are slight variances between female individuals when it comes to mailing to men and women. Female participants were more inclined to avoid using emojis while sending messages to other women than when sending messages to men (*FtoM*=42.03%, *FtoF*=49.28%). Furthermore, in the responses Female-to-Female scenario, 😠 (*FtoF*=15.94%) was more frequent than 🤬 (*FtoF*=13.77%), proving women are more likely to use less angry responses when sending to women. To see if there were differences in the selection of emojis with stronger or lesser emotions, this study combined the responses with no emoji and 😠, which were less emotional, and combined the responses with 😡 and 🤬, which were more emotional, as shown in Figure 3. After calculating the combined results, there was a significant difference in this change between females and males ($p < 0.1$)[4]. In light of these differences, this study has two additional findings in (7) and (8).

(7) Women tend to choose less intense emojis to express emotions in the same gender (*FtoF*) in the emotion intensifier emoji condition.

This study attributes the difference in (7) to women being more sensitive to emotions than men (Fischer et al., 2018; Barrett and Bliss-Moreau, 2009). When women send emojis to other women, they subconsciously assume that the other party will be more likely to receive their anger, and thus use less intense emojis. In addition, in Gordon's (1997) study, women were more

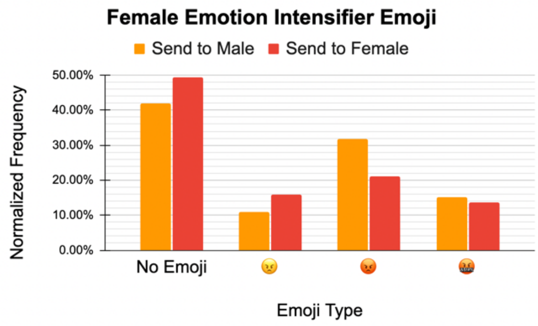


Figure 2: Female Emotion Intensifier Emoji Responds Frequency When Sending to Male and Female.

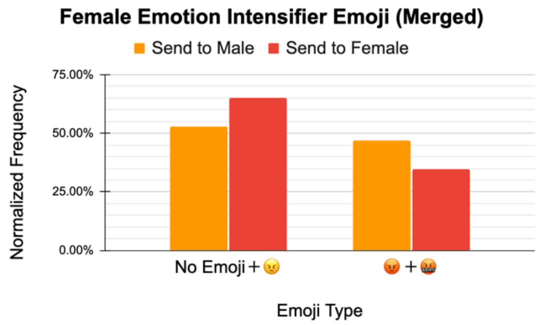


Figure 3: Female Emotion Intensifier Emoji Responds Frequency When Sending to Male and Female (Merged).

restrained in expressing anger than men, indicating that women would restrain themselves from using stronger words and emojis when sending texts or emotions. That is, in the emotion intensifier emoji sentence pattern, the sentence already contains angry words with direct condemnation; in order to restrain themselves from using strong expressions, women generally choose not to use emojis, the least angry option, to stop emotions from intensifying. Therefore, in terms of the results of this study, female subjects prefer not to use emojis when using the emotion intensifier emoji and reduce the other three options.

(8) Women tend to choose stronger emojis to express toward different gender (*FtoM*) in the emotion intensifier emoji situation.

In the case of 😡 (*FtoM*=31.88%, *FtoF*=21.01%), female participants used this emoji more often when sending messages to men than when sending messages to women. In addition, among the responses from women to messages sent to men, 🤬

[4] *p*-value = 0.04994.

(*FtoM*=15.22%) were more likely to be selected than 😠 (*FtoM*=10.87%). This is a significant finding as it supports that women are more likely to employ angrier emojis when sending to males.

According to past literature, women are more likely than men to be able to sense the emotions of others, as is suggested by traditional stereotypes. The sender must therefore frequently convey more emotions to the males, using more intense emojis for men, in order for them to comprehend the feelings that the sender wishes to convey. Fischer et al. (2018) suggested there are two possible reasons why men are less capable of emotion perception. First, men pay more attention to subtle facial expressions and are therefore able to perceive more complex emotional features in the face. However, there are no detailed differences in emojis, only major changes in color, text, etc.; thus, men are less likely to detect differences in emotional intensity in emojis than women. Second, when asked to rate the intensity of multiple emotions, males are more prone to be perplexed and unsure about their own emotional perception. Similarly, Herring and Dainas (2020) found that men are more likely to be confused and annoyed by emoticons. The results are very likely to trigger the transfer phenomenon in the series of emotional questions and answers in this study.

In summary, there is no significant difference in men's responses to emotion intensifier emoji messages, nor does their original emotional expression change due to the gender of the recipient. However, women have a particularly significant difference in emotion intensifier emojis, meaning they choose less intense emojis for the same gender and more intense emojis for the different gender. This difference affects the frequency of the overall expression of emotion intensifier for both men and women.

4.4 Emojis in Different Emoji Types

4.4.1 Male Emoji Distribution in Emotion Signal and Emotion Intensifier Emojis

There was no significant difference in the use of emotion signal emojis and emotion intensifier emojis among male subjects when sending messages to males and females (*MtoM*: $p = 0.4375$; *MtoF*: $p = 0.9376$).

Since male subjects did not change their choice of emojis in both sentence patterns, it indicates that men do not change the emotions they originally

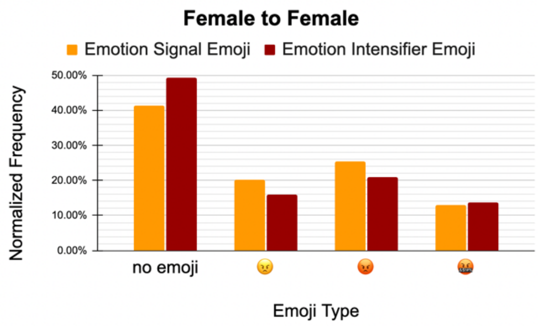


Figure 4: Female Emotion Signal Emoji and Emotion Intensifier Emoji Responds Frequency When Sending to Female.

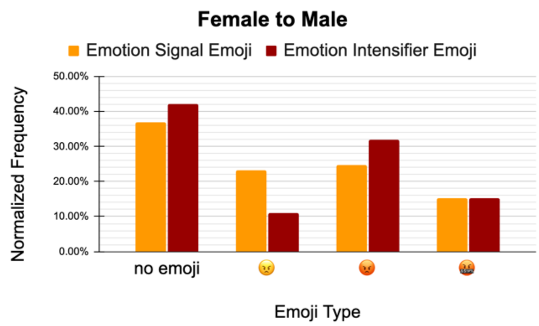


Figure 5: Female Emotion Signal Emoji and Emotion Intensifier Emoji Responds Frequency When Sending to Male.

wanted to express, even if the sentence pattern changed. Given the notion that women are too emotional, Barrett and Bliss-Moreau (2009) noted that men express their emotions when the situation calls for it, demonstrating that men are more objective and unaffected by emotions. Women express their emotions because they are "emotional creatures." That is, the emotion intensifier sentences in this study contain angry words. Since women are more easily affected by the words, the emojis they choose also change according to the sentence pattern. On the contrary, men are objective and less susceptible to emotions. They were not swayed by the angry words, supporting the finding that men did not differ significantly in their use of emotion signals and emotion intensifier emojis.

4.4.2 Female Emoji Distribution in Emotion Signal and Emotion Intensifier Emojis

Although there is no statistically significant difference in the female subjects' responses to the two sentence types ($p = 0.5168$), 😠 and 😡 have a decreasing trend in emotion intensifier emojis, while the value of not using emojis (increase 7.98%)

has an increasing trend, as shown in Figure 4. This supports that women tend to use less intense emojis when sending messages to the same gender under emotion intensifier emoji conditions.

Unlike the results for the same gender, female subjects showed significant differences when sending to males ($p < .1$)[5]. As shown in Figure 5, in the case of the emotion intensifier, the value of 😠 (-12.32%) has a downward trend, and the value of 😡 (+7.24%) has an upward trend. Thus, the finding reveals women tend to use more intense emojis, as the more intense emojis were chosen when expressing themselves to the opposite gender in emotion intensifier emojis.

4.5 Emojis in Different Sentence Structures

4.5.1 Syntax Influences on Emoji

There is no significant difference in the performance of men in affirmative (A) and interrogative (I) sentences (p = 0.9690). The most used by men is 😡 (A=39.81%, F=41.67%), followed by 😠 (A=36.11%, F=35.19%), and finally 🤬 (A=24.07%, F=23.15%), shown in Figure 6. In terms of the emoji usage frequency, 😡 is similar to the emotion signal type as the usage frequency of 😡 remains the highest.

Besides, there was no significant difference in the performance of female subjects in affirmative sentences and interrogative sentences (p = 0.5309). However, as shown in Figure 7, this study found that the distribution of different emojis in different sentence types was slightly different. Among them, compared with affirmative sentences, interrogative sentences have stronger emotional 😡, and their usage ratio has increased (A=38.41%, I=44.20%, increased 6.52%); on the contrary, 😠 has decreased (A=44.20%, I= 37.68%, decreased 5.79%). This study speculates that women tend to use stronger emojis in interrogative sentences, which will be discussed in 4.5.2.

4.5.2 "?" as an Emotion Intensifier Symbol

This study justifies the result in 4.5.1 by explaining that interrogative sentence patterns also affect the use of emojis. Sagum et al. (2019) highlight the role of ending punctuation in emotion intensity level (Karami et al., 2023). Specifically, the question mark in the anger scenario is classified as

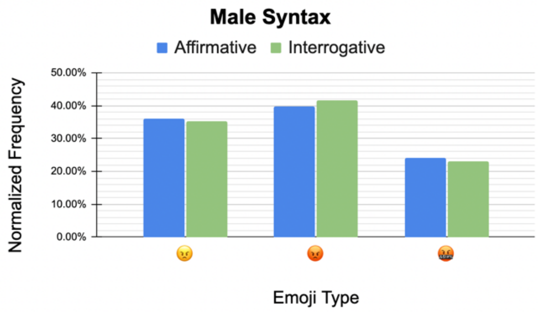


Figure 6: Male Affirmative Sentence and Interrogative Sentence Responds Frequency.

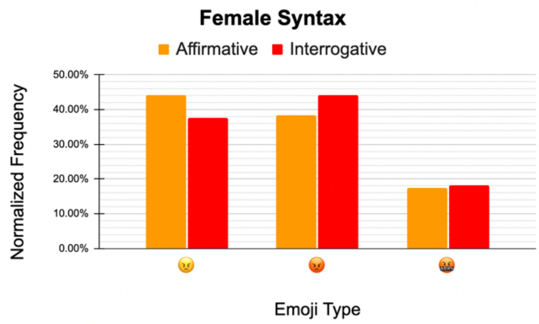


Figure 7: Female Affirmative Sentence and Interrogative Sentence Responds Frequency.

medium or high intensity. As women encounter interrogative sentences, their distribution pattern is similar to the emotion intensifier emoji pattern of women, making it easier for female subjects to influence their choices for this symbol. This study hypothesizes two possible reasons for this result: first, women will unconsciously treat the other party as a man during the conversation, which will have an emotionally reinforcing effect on the opposite gender during the emotional enhancement period; second, "?" is a special kind of "emotion intensifier symbol" (not an emoji, sentence, or word). Hence, its results are different from the emotion intensifier emoji in this study, meaning that when female subjects encounter symbolic emotion enhancement, there is an effect of intensified tone. Thus, the present study proposes the addition of the symbolic emotion intensifier symbol to the emoji intensifier symbol. However, whether this phenomenon is affected by gender remains to be clarified. In view of the lack of previous literature, the above conclusions about "?" are only speculations and need to be corroborated by more literature and experiments.

[5] p-value = 0.04855.

5 Conclusion

The study found notable differences in how men and women express anger through emojis. Women use stronger emojis when communicating with men and less intense ones with women, potentially because women are more emotional and sensitive to emotions compared to men (Barrett and Bliss-Moreau, 2009; Fischer et al., 2018). Women also choose more intense emojis in interrogative sentences than in affirmative sentences. This study proposes two explanations for this result: first, women may unconsciously treat the recipient as a man during the conversation, resulting in an enhanced situation. Second, the "?" in question sentences may be an "emotion intensifier symbol," which has an emotion-enhancing effect on women that is different from other sentence structures.

Today, the activeness of online conversations has led to communications between humans and artificial intelligence, among which the ability of artificial intelligence to distinguish emotions is highly valued for its development. Based on the results of this study, the expression of emotions by men and women is affected by sentence patterns. If the gender is different, will artificial intelligence provide different responses? Since the scale of this study is not large enough to cover all Internet users worldwide, some results need to be confirmed. However, this study provides a preliminary discussion on the emotion of data—emojis, pioneering first-hand research into the new generation's communication.

Limitations

The present study only investigates human responses in the Chinese Traditional language, which covers a partial aspect of Chinese culture. Yet, since emotional responses are linked with culture and gender, the present study requires future research to investigate how other cultures behave and respond. Besides, this research primarily investigates the syntax's effects on emoji selection, also calling for future studies to look in-depth at the phenomenon.

Ethical Considerations

This paper was conducted through a digital survey and recruited 36 men and 46 women. Each participant's information is secure as the survey is anonymous. Furthermore, participants are required to read through brief test descriptions, including the purpose of the study, research process, confidentiality, and potential risks, and they can opt out of the experiment at any time. Since the participants are aged 11 to 20, they are required to finish the questionnaire under their guardians' supervision, reducing potential harm to the participants.

References

Agneta H. Fischer, Mariska E. Kret, and Joost Broekens. 2018. Gender differences in emotion perception and self-reported emotional intelligence: A test of the emotion sensitivity hypothesis. *PLoS ONE*, 13(1).

Agneta H. Fischer, Patricia M. Rodriguez Mosquera, Annelies E. M. van Vianen, and Antony S. R. Manstead. 2004. Gender and Culture Differences in Emotion. *Emotion*, 4(1):87–94.

Daantje Derks, Arjan E.R. Bos, and Jasper Von Grumbkow. 2007. Emoticons and social interaction on the Internet: the importance of social context. *Computers in Human Behavior*, 23(1):842–849.

Elizabeth Gordon. 1997. Sex, speech, and stereotypes: Why women use prestige speech forms more than men. *Language in Society*, 26(1):47–63.

Francisco Yus. 2014. NOT ALL EMOTICONS ARE CREATED EQUAL. *Linguagem Em (Dis)Curso*, 14(3):511–529.

Gary Lupyan and Rick Dale. 2016. Why are there different languages? The role of adaptation in linguistic diversity. *Trends in Cognitive Sciences*, 20(9):649–660.

Jui-Chun Wu. 2013. Gender-Based Differences in Hakka Complaint Realization. *Chinese Studies*, 31(4):279–318.

June De Vaus, Matthew J. Hornsey, Peter Kuppens, and Brock Bastian. 2018. Exploring the East-West Divide in Prevalence of Affective Disorder: A Case for Cultural Differences in Coping With Negative Emotion. *Personality and Social Psychology Review*, 22(3):285–304.

Mansooreh Karami, Ahmadreza Mosallanezhad, Michelle V. Mancenido, and Huan Liu. 2023. "Let's eat grandma": Does punctuation matter in sentence representation?. In *Lecture notes in computer science*, pages 588–604.

Ming-Chung Yu. 2005. Sociolinguistic Competence in the Complimenting Act of Native Chinese and American English Speakers: A Mirror of Cultural Value. *Language and Speech*, 48(1):91–119.

Nangyeon Lim. 2016. Cultural differences in emotion: differences in emotional arousal level between the East and the West. *Integrative Medicine Research*, 5(2):105–109.

Ria Ambrocio Sagum, Monique L. Navarro, and Arvin Jasper E. 2019. EMOSIS Sentiment Analysis on Tweets with Emotion and Intensity Level Recognition Considering Ending Punctuation Marks. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(4):10289–10293.

Sarah E. Butterworth, Traci A. Giuliano, Justin White, Lizette Cantu, and Kyle C. Fraser. 2019. Sender gender influences emoji interpretation in text messages. *Frontiers in Psychology*, 10.

Shelia M. Kennison, Maria Andrea Hurtado Morales, Katie E. Nelson, and Eric Chan-Tin. 2025. Gender differences in emoji use: relationships with personality traits. *Current Psychology*, 44:9865–9875.

Susan C. Herring and Ashley R. Dainas. 2020. Gender and age influences on interpretation of emoji functions. *ACM Transactions on Social Computing*, 3(2):1–26.

Li Li and Yue Yang. 2018. Pragmatic functions of emoji in internet-based communication---a corpus-based study. *Asian-Pacific Journal of Second and Foreign Language Education*, 3(1).

Lisa Feldman Barrett and Eliza Bliss-Moreau. 2009. She's emotional. He's having a bad day: Attributional explanations for emotion stereotypes. *Emotion*, 9(5):649–658.

Unicode. (2022). *Emoji List, v15.0*. https://unicode.org/emoji/charts/emoji-list.html.

Yuan-Shan Chen, Chun-Yin Doris Chen, and Miao-Hsia Chang. 2011. American and Chinese complaints: Strategy use from a cross-cultural perspective. *Intercultural Pragmatics*, 8(2).

Meeting Delegate: Benchmarking LLMs on Attending Meetings on Our Behalf

Lingxiang Hu [1*] **Shurun Yuan** [2*] **Xiaoting Qin** [3] **Jue Zhang** [3†] **Qingwei Lin** [3]
Dongmei Zhang [3] **Saravan Rajmohan** [3] **Qi Zhang** [3]
[1]Northeastern University, China [2]Peking University, China [3]Microsoft
{xiaotingqin, juezhang}@microsoft.com

Abstract

In contemporary workplaces, meetings are essential for exchanging ideas and ensuring team alignment but often face challenges such as time consumption, scheduling conflicts, and inefficient participation. Recent advancements in Large Language Models (LLMs) have demonstrated their strong capabilities in natural language generation and reasoning, prompting the question: *can LLMs effectively delegate participants in meetings?* To explore this, we develop a prototype LLM-powered meeting delegate system and create a comprehensive benchmark using real meeting transcripts. Our evaluation shows GPT-4/4o balance active and cautious engagement, Gemini 1.5 Pro leans cautious, and Gemini 1.5 Flash and Llama3-8B/70B are more active. About 60% of responses capture at least one key point from the ground truth. Challenges remain in reducing irrelevant or repetitive content and handling transcription errors in real-world settings. We further validate the system through practical deployment and collect feedback. Our results highlight both the promise and limitations of LLMs as meeting delegates, providing insights for their real-world application in reducing meeting burden.

1 Introduction

Nowadays, the nature of work has increasingly become more collaborative (Mugayar-Baldocchi et al., 2021), with meetings becoming an essential component (Spataro, 2020) to facilitate the exchange of ideas and information, fostering innovation and ensuring alignment among team members.

Attending meetings, however, poses notable difficulties. Firstly, the rapid increase in the number of meetings can consume a substantial amount of time, diverting attention from core tasks and reducing overall productivity (Perlow et al., 2017; Kost, 2020). Secondly, scheduling conflicts often arise when multiple meetings are double-booked, forcing participants to prioritize or miss valuable discussions altogether. Thirdly, not all meetings require full attendance; participants may only need to contribute to specific topics, leading to inefficiencies when attendees are required for entire duration.

In this study, we investigate the feasibility of developing a meeting delegate system to represent individuals in meetings. This concept is becoming increasingly viable with the advancement of Large Language Models (LLMs). These LLMs, renowned for their remarkable capabilities in natural language understanding and generation (Ouyang et al., 2022; OpenAI, 2023; Google, 2024a), demonstrate potential to comprehend meeting context, participate in dynamic conversations, and provide informed responses.

Developing LLM-powered meeting delegate systems faces several challenges. Firstly, such systems must navigate complex, context-rich conversations involving multiple participants, requiring them to discern opportune moments for engagement and restraint. Secondly, human conversations often contain ambiguities and uncertainties, such as queries directed ambiguously or pronunciation-related ambiguities, which challenge the system's ability to respond effectively. Thirdly, ensuring user privacy is crucial to prevent over-sharing of information and safeguard the user's personal image. Finally, these systems must operate in real-time, necessitating low-latency responsiveness.

This study evaluates LLMs in the meeting delegate role, initially addressing the first two challenges while leaving privacy and latency considerations for future work. Unlike prior research that examines the facilitator role in meetings (Mao et al., 2024), our work addresses the more prevalent participant role. Specifically, we focus on how LLMs manage the nuances of meeting discussions, including their ability to discern when to intervene, respond to ambiguities, and maintain an accurate understanding of evolving contexts.

In the absence of established benchmarks, we

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 283–316
November 9, 2025

constructed an evaluation dataset from real meeting transcripts. Our evaluation reveals that GPT-4/4o maintain balanced performance between active and cautious engagement strategies, while Gemini 1.5 Pro is more cautious, and Gemini 1.5 Flash and Llama3-8B/70B are more active. Overall, 60% of responses address at least one main point from the ground-truth. While the results demonstrate the potential of LLMs in meeting scenarios, improvements are needed to reduce irrelevant or repetitive content and enhance tolerance for transcription errors. To extend this evaluation beyond datasets, we also implemented and tested a prototype system in real-world settings.

Our contributions are summarized as follows:

- We conduct a systematic evaluation of LLMs in the meeting delegate system, specifically assessing their role as participants.
- We introduce the first evaluation benchmark in this domain, derived from real meeting transcripts, encompassing four scenarios: Explicit Cue, Implicit Cue, Chime In, and Keep Silence. The dataset will be released with this paper.
- We assess popular LLMs using the benchmark and develop a prototype for real-world testing, including an ablation study on the impact of transcription errors.

2 Related Work

Language Model Applications in Meetings. Considerable research has been dedicated to the summarization of meetings (Zhong et al., 2021) and other real-life dialogues (Mehdad et al., 2014; Tuggener et al., 2021). In the context of meetings, key tasks include meeting transcript summarization and action item identification (Cohen et al., 2021). MeetingQA (Prasad et al., 2023) investigated Q&A tasks based on meeting transcripts, highlighting the challenges faced by models such as RoBERTa in handling real-world meeting data. Recent advancements in LLMs have opened new avenues for enhancing these tasks. For instance, an LLM-based meeting recap system (Asthana et al., 2023) has demonstrated effectiveness in generating accurate and coherent summaries and action items.

Facilitator in Multi-Participant Chat. MUCA (Mao et al., 2024) presents a framework that leverages LLMs to facilitate group chats by simulating users, demonstrating notable effectiveness in goal-oriented conversations. Similarly, approaches like GPT-4o demo for meetings (OpenAI, 2024a) are designed to serve as facilitators in group discussions. While these studies underscore LLMs' capabilities in managing group chats, they primarily focus on LLMs guiding the meeting process rather than representing individuals with different roles.

Role-Playing with LLMs: Characters and Digital Twins. Role-play prompting (Kong et al., 2024) has proven effective in triggering chain-of-thought reasoning in LLMs. Research on simulating famous personalities (Shao et al., 2023; Sun et al., 2024) has also explored character consistency and social interactions in agent-based group chats. While Reid Hoffman's (Hoffman, 2024) GPT-4-powered digital twin showcased the potential of AI-driven representations, it was limited to one-on-one interactions, leaving group discussions largely unexplored. Unlike prior work, we focus on LLMs as meeting delegates, providing targeted engagement for multi-participant, goal-specific meetings. Our evaluation and real-world deployment show their potential to reduce individual meeting burdens and advance LLM applications in professional settings.

3 LLM-based Meeting Delegate System

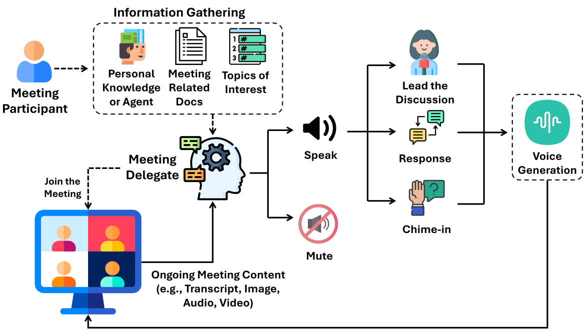


Figure 1: Architecture of the meeting delegate system.

Given the absence of a standardized meeting delegate architecture, this study adopts the design depicted in Figure 1. The system first employs an *Information Gathering* module to collect meeting-related information, facilitating LLM-driven participation. Users can manually provide topics of interest, background knowledge, and shareable materials prior to the meeting. Alternatively, if the user has a personal knowledge base or an intelligent personal assistant/agent, the system can query them in real-time, provided latency is manageable.

Once in the meeting, the system monitors proceedings and employs LLMs to determine appropriate engagement timing and content. While various contextual data sources (e.g., transcripts, screen sharing, audio) are available, this work focuses

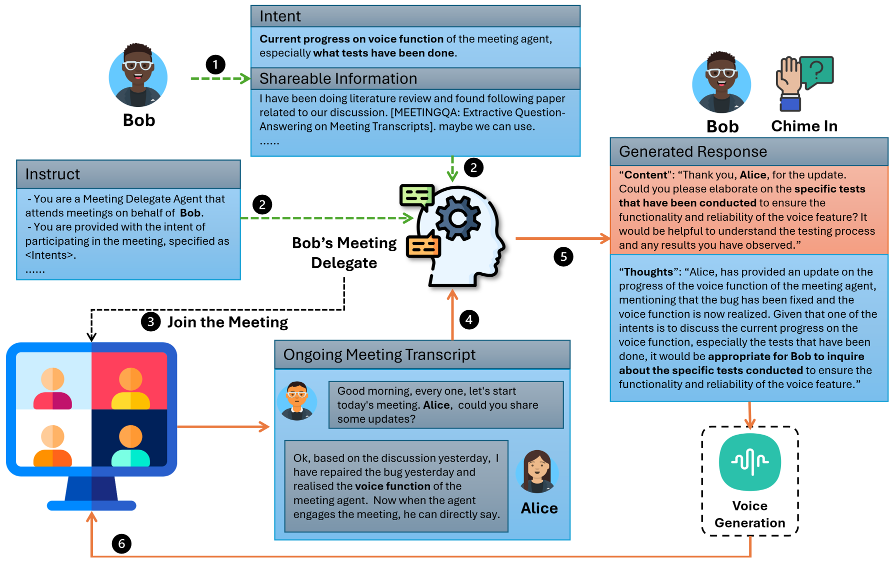


Figure 2: Workflow of an LLM-powered meeting delegate system. The process involves user input of meeting intent and shareable information prior to the meeting, real-time participation based on meeting transcripts, and response generation aligned with prompted instructions and meeting objectives.

on transcripts obtained from meeting software or speech-to-text tools. Figure 1 outlines three response types: leading discussions, responding to others, and chiming in. This study prioritizes the latter two, emphasizing the participant role. The generated text-based response can then be converted into speech using TTS technology, potentially mimicking the user's voice.

An example of this workflow is depicted in Figure 2, drawn from a real prototype implementation (detailed in Section 6.2). In this example, Bob uses his Meeting Delegate to participate in a meeting with Alice and others. Before the meeting, Bob provides topics of interest and relevant shareable information to the Meeting Delegate ❶. This information, along with instructions, forms the prompt for the Meeting Delegate ❷. The delegate then joins the meeting ❸ and determines, based on the ongoing meeting transcript, whether to engage ❹. During the meeting, Alice discusses updates on the voice function, which aligns with Bob's goal to learn about its progress. The Meeting Delegate then chimes in ❺, generating a text-based response (converted to speech ❻), asking for more details, thus achieving Bob's objectives and engaging in the conversation.

4 Benchmark Dataset

While the preceding example illustrates the Meeting Delegate in action, a more systematic evaluation is needed to gauge how effectively it times its interventions and generates relevant responses. No existing benchmark datasets meet these objectives, prompting us to create one.

4.1 Dataset Construction

Our dataset construction strategy involves using real meeting transcripts and generating test cases by taking "snapshots" from these transcripts. A "snapshot" is defined as a truncation of the transcript after a participant's utterance. Then, by comparing the generated response according to this snapshot with the actual responses in the real script, we can determine how well the system performs. An illustration of this process in given in Figure 8. The base meeting transcripts are taken from the ELITR Minuting Corpus (Nedoluzhko et al., 2022), comprising de-identified project meeting transcripts in English and Czech. 61 English meeting transcripts are used and the test cases are constructed as follows. Motivated by promising results from LLM evaluation and annotation (Li et al., 2024; He et al., 2024; Gilardi et al., 2023), we leverage LLMs for dataset preparation while conducting manual verification to ensure quality assurance. The prompts employed in this process are carefully crafted and iterated to ensure that the extracted test cases meet the desired criteria. To further ensure accuracy, all extracted cases are manually reviewed and validated by two authors. This dual approach, leveraging LLMs for scalability while incorporating human oversight for quality assurance, offers a balanced and reliable methodology.

Specifically, we first employ GPT-4 to progressively analyze each participant's utterances by taking a "sliding window" on the original meeting transcript. This is to capture their meeting intents and the information that they can share during the meeting, serving as the critical input to the Meeting Engagement module for response generation. The shareable meeting information contains pairs of <Context> and <Information>, with <Context> specifying under which context the points in <Information> can be shared. Details of this intent and contextual information extraction prompt can be found in Table 22 in the Appendix.

Next, we extract suitable snapshots from the transcripts as test cases. For each participant (excluding facilitators), we identify their utterances and use the preceding transcript as the ongoing meeting context. The ground-truth response is determined by considering *several* subsequent utterances. This extraction process leverages GPT-4 (prompt in Table 28) to classify meeting scenes into Explicit Cue, Implicit Cue, and Chime In (definition in Section A) and select the necessary utterances to form the ground-truth response, recognizing that a user's response may span multiple subsequent utterances. As the extracted test cases closely match real transcripts, we refer them as the **Matched Dataset**.

To evaluate the meeting delegate's ability to Keep Silence when inappropriate to speak, we construct a **Mismatched Dataset** from the Matched Dataset. We take Explicit Cue and Implicit Cue test cases and replace the principal who needs to respond with another participant not involved in the current conversation. The intents and shareable meeting information are accordingly replaced, and the ground-truth is set to be empty. The delegate representing the new principal is expected to remain silent when presented with these transcripts.

Lastly, we construct a **Noisy Name Dataset** for our ablation study, addressing the fact that meeting transcribing systems often introduce noise affecting the meeting delegate's performance. This issue is particularly significant for recognizing names, which are crucial in Explicit Cue cases. For example, the Chinese name "Jisen" might be transcribed as "Jason". In our construction, we modify the Explicit Cue cases by replacing de-identified names with real-world names and substituting the principal's name in the final utterance with a phonetically similar word to simulate transcription errors.

4.2 Evaluation Metric

In our evaluation, we generate responses using LLMs with the same prompt as in our prototype. These responses are assessed using two categories of metrics: **Response Rate / Silence Rate**, which determines whether a response is generated, and quality-related metrics, **Recall** and **Attribution**. To strengthen interpretability, we complement the quality metrics with semantic similarity scoring, with results reported in Section C in Appendix.

The Recall metric evaluates if the generated response includes key points present in the ground-truth response. We define two recall rates: "loose" recall rate, which is 1 if at least one main point from the ground-truth is mentioned and 0 otherwise; and "strict" recall rate, which measures the percentage of main points from the ground-truth included in the generated response.

Attribution assesses the origin of the main points in the generated response, classifying them into four categories: the expected ground-truth response (Expected Response), contextual information not present in the ground-truth (Contextual Information), previous transcript content (Previous Transcript), and hallucinated texts (Hallucination).

We leverage LLMs for main point extraction and their semantic comparison. Specifically, in the Recall phase, GPT-4 is employed to assess how well the LLM-generated responses match key points from the ground-truth response set, using the prompt provided in Table 18. In the Attribution phase, GPT-4 Turbo is used to trace and evaluate the origin of specific points in the responses, with the prompt provided in Table 20. Through manual validation of 80 randomly sampled cases (10% of total statistics), we observed that with carefully crafted instructions, LLMs achieved an average of 95% accuracy on the Recall and Attribution evaluation tasks, thereby supporting their use in our experiments[1] Two annotators independently labeled 80 test cases with binary match judgments. Inter-annotator agreement was 87.5% (Cohen's Kappa = 0.69), indicating substantial consistency. Notably, our subsequent manual analysis of all failure cases for response rate (Section 5) did not reveal any errors, further reinforcing the robustness of our evaluation methodology.

4.3 Dataset Statistics

From the 61 original meeting transcripts, we extract 846 test cases for Matched Dataset, in which 54.5% belongs to Implicit Cue, followed by 30.9% for Explicit Cue and 14.7% for Chime In. The numbers

[1] See Appendix D for an example showing GPT-4 reliably identifies main points and yields a meaningful Recall score.

of test cases for Mismatched Dataset and Noisy Name Dataset are 294 and 122, respectively.

For Matched Dataset, we present various data statistics in Figure 3. Over 50% of test cases involve more than four participants and contain transcripts exceeding 50 utterances, highlighting the dataset's complexity and the involvement of multiple individuals. Additionally, approximately 40% of test cases include at least two main points in the ground-truth response, and in more than 50% of cases, participants contribute over ten main points. This indicates a substantial level of detail and interaction within the meetings, suggesting that the dataset captures rich and multifaceted discussions.

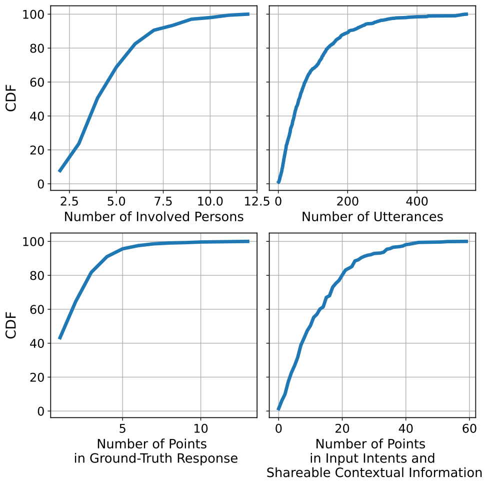


Figure 3: Data statistics of the Matched Dataset.

5 Experiment

Setup. In our experiment, we utilize three prominent series of LLMs: the GPT series (GPT-3.5-Turbo, GPT-4, GPT-4o) (OpenAI, 2024c), the Gemini series (Gemini 1.5 Flash, Gemini 1.5 Pro) (Google, 2024b) and the Llama series (Llama3-8B, Llama3-70B) (Meta, 2024). For all LLMs[2], we set the temperature to 0 and use the default API settings for other parameters. Note that, due to model context window restriction, we remove test cases that exceed the 8K context window for Llama3 models (56.3% kept) and those exceeding the 16K context window for GPT-3.5-Turbo (94.3% kept), while keeping all for other LLMs. We repeated all experiments on the intersection subset of test cases within the 8K context window as the Llama series, and the patterns and experimental findings (Tables 5, 7, 9, and 11 in Appendix) are consistent.

[2]Exact model versions can be found in Table 14.

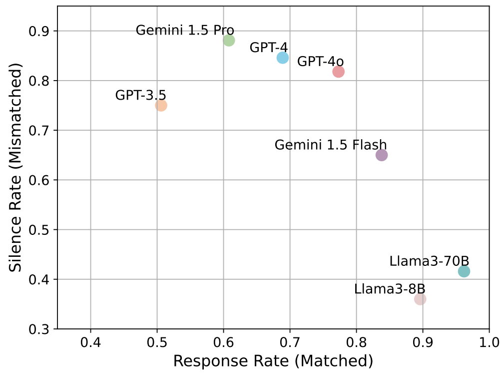


Figure 4: Response Rate on Matched Dataset vs. Silence Rate on Mismatched Dataset.

Response Rate Analysis. The Response and Silence Rates of the studied LLMs are obtained for Matched and Mismatched Datasets, respectively. Summarized results are presented in Figure 4, with further details (e.g., breaking down to different meeting scenes) provided in Tables 4 and 6 in the Appendix. Overall, GPT-4 and GPT-4o demonstrated *balanced* performance, with Response/Silence Rates between 0.7 and 0.8. Among the Gemini series models, Gemini 1.5 Pro achieved the highest Silence Rate of approximately 0.9, coupled with a low Response Rate, indicating a *cautious* engagement strategy. In contrast, the smaller Gemini 1.5 Flash model and the Llama series exhibited higher activity levels, suggesting a more *proactive* engagement approach; however, this also led to a tendency to engage when they should remain silent. These patterns persisted when all LLMs are tested using the same subset of cases as the Llama series.

To uncover the underlying causes of failures, we conduct an in-depth analysis of all failure cases in representative models: GPT-4o and Gemini 1.5 Pro for state-of-the-art LLMs, and Gemini 1.5 Flash and Llama3-8B representing more lightweight models. We manually analyze and categorize all error types, proposing corresponding directions for improvement, as summarized in Table 1. For instance, in the "Explicit Cue" scenario within the Matched Dataset, the meeting delegate may correctly identify the cue but fail to respond, indicating a need for enhanced reasoning capabilities in meeting contexts. Detailed results for each model can be found in Figure 9 in Appendix. A summary of these results is presented in Figure 5. Our findings reveal that: 1) LLMs like GPT-4o and Gemini 1.5 Pro can improve performance or make functional advancements in meeting scenarios by enhancing reasoning in meeting-specific context,

Table 1: Mapping between Error Types and Solution Direction for Response Rate Failure Cases Study.

Dataset	Scenarios	Error Type	Solution Direction
Matched	Chime In	Decision based on wrong latest utterance	Improved Instruction Following
		Identify as cue to others or all participants	Enhanced Reasoning in Meeting Scenario
		Missing the need for proactive participation	Enhanced Reasoning in Meeting Scenario
		Decision made due to "Conversation is still going, I can't interrupt"	Enhanced Reasoning in Meeting Scenario
		Unable to find the related context	Enhanced General Reasoning
		Other	N/A
	Explicit Cue	Decision based on wrong latest utterance	Improved Instruction Following
		Correctly recognizes the cue but does not respond	Enhanced Reasoning in Meeting Scenario
		Ambiguity due to multiple names in a single utterance or long context	Enhanced Reasoning in Meeting Scenario
		Fails to recognize the cue	Enhanced General Reasoning
		Hallucination	Enhanced General Reasoning
		Other	N/A
Mismatched	Mismatched	Decision based on wrong latest utterance	Improved Instruction Following
		Latest utterance related to provided information	Enhanced Reasoning in Meeting Scenario
		Failure to recognize cues directed to others	Enhanced Reasoning
		Hallucination	Enhanced General Reasoning
		Other	N/A

and 2) smaller models need to improve general instruction following and reasoning abilities before addressing meeting-specific issues.

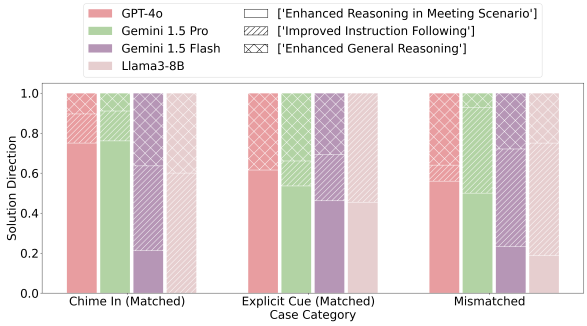


Figure 5: Solution directions from error analysis of bad cases in Response (Silence) Rate for Matched and Mismatched Datasets.

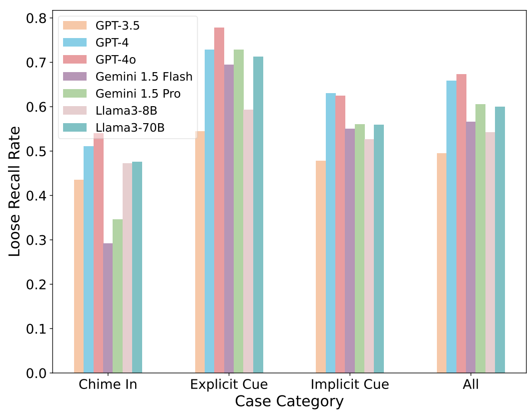


Figure 6: Loose recall rate on Matched Dataset.

Recall Analysis. The recall results for both loose and strict metrics are similar; therefore, we only present the loose recall rate for all studied LLMs on Matched Dataset in Figure 6. Detailed results, including the strict recall rate, are available in Table 8 in the Appendix. Figure 6 shows that these LLMs achieve a loose **recall rate of approximately 60%**. This indicates that, for 60% of test cases, the generated response contains at least one key point from the ground-truth response. Such a result demonstrates the potential, as it suggests that LLM-powered meeting delegates can typically generate contextually relevant responses, contributing to maintaining the overall meeting flow.

Performance differences among the LLMs reveal that GPT-4o achieves the highest performance across almost all categories, followed by GPT-4. The two Gemini models exhibit similar performance, excelling in "Explicit Cue" but lagging in "Chime In". The Llama series models perform comparably to the Gemini models but tend to be better in "Chime In" scenarios.

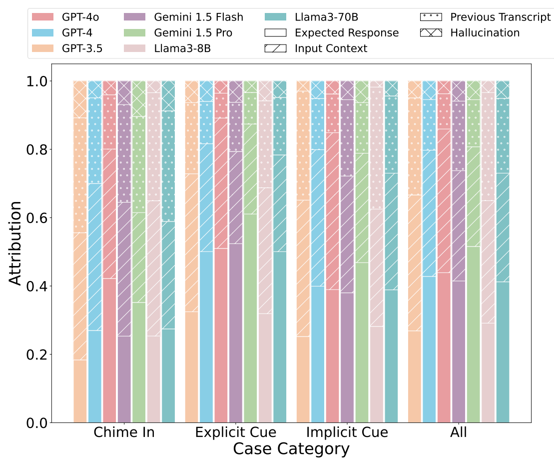


Figure 7: The attribution rate on matched dataset.

Attribution Analysis. For the Attribution metric, we seek a high percentage of "Expected Response", indicating high accuracy in responding to given cues, while minimizing other categories, particularly "Hallucination". As shown in Figure 7, most models, except GPT-3.5-Turbo and Llama3-8B, have approximately 40% of their responses attributable to the ground-truth response, with Gemini 1.5 Pro achieving the highest performance at around 50%. About 30% of generated responses are attributed to other input context information not directly related to the ground-truth response, indicating room for improvement in reasoning over the provided information. The proportion attributed to the previous transcript varies significantly across models, ranging from 10% to 30%. Higher values suggest repetitive messages in the generated response, potentially detracting from the meeting experience due to verbosity. The portion of hallucinated texts is minimal, at only 5% across all models, indicating that current LLMs maintain good trustworthiness in meeting engagement.

Regarding performance differences across models, we observe that models generally considered more capable demonstrate better performance, while models like GPT-3.5-Turbo and Llama3-8B, viewed as less capable, show inferior performance. This alignment between general model performance and specific scenarios suggests that in future, more capable general LLMs will also benefit meeting delegate scenarios.

Correlation Analysis. We correlate the performance of the above metrics with test case metadata (*i.e.*, those shown in Figure 3). Figure 10 in the Appendix presents an example result for GPT-4o. The result indicates that GPT-4o maintained stable performance across different transcript lengths and complexity measures, including meeting size and input diversity. Therefore, *no* significant relationships between the evaluation metrics and the metadata were observed.

Ablation Study. Two scenarios are considered in our ablation studies. First, we examine the impact of erroneous transcription of participant names to phonetically similar words using the Noisy Name Dataset. We measure the response rates of all models on this dataset, observing a significant drop in performance (see Table 12 in the Appendix). For instance, GPT-4o's response rate declines from 94.3% in the Explicit Cue cases of the Matched Dataset to 68% in the Noisy Name Dataset. This highlights challenges in accurately recognizing participant names. Further model fine-tuning to better handle such transcription errors may be necessary.

In our second study, we investigate how model performance is affected by the provision of context information in the input. Currently, context information is structured as pairs of <Context> and <Information>, specifying under which conditions the information in <Information> can be shared. This setup may not reflect real-world scenarios where users might not always anticipate the context for sharing specific information. To assess the impact, we remove <Context> from test cases and use <Information> and <Intents> alone as input to generate responses. We evaluate this on a subset of 121 test cases from the first 11 meetings using GPT-4o. Detailed results are provided in Table 13 in the Appendix, showing minimal performance impact across all evaluation metrics when context information is omitted.

6 Discussion

6.1 Phased Deployment of Meeting Delegate

This study primarily explores the feasibility of using LLMs to represent users by generating meaningful content in meetings. However, deploying such a meeting delegate system in real-world settings requires addressing additional critical responsible AI practices and ethical considerations (see further discussion in Ethics Statement). Key challenges include implementing strong privacy safeguards, such as secure data handling, consent mechanisms, user-defined boundaries, and audit trails. Reviews (Yan et al., 2024; Anwar et al., 2024) of current privacy-preserving methods for LLMs highlights the difficulty of achieving a fully autonomous and unconstrained meeting delegate at present. Therefore, we propose a three-phase approach that incrementally enhances AI's autonomy and responsibility, as detailed in Table 2. The phases are characterized by evolution of data boundaries and limitations on the delegate's roles in sharing information, collecting data, and making decisions.

In **Phase I (Execute)**, the delegate operates strictly within user-defined data boundaries, sharing only explicitly approved information and collecting information from other meeting participants based on direct user instructions. There is no autonomous decision-making allowed, ensuring strong user control and minimal privacy risk. In **Phase II (Assist)**, the system can reason over sensitive data while adhering to privacy guidelines. It infers context beyond explicit instructions and can propose actions, though user approval is still required for making decisions. This phase introduces controlled autonomy with dynamic data boundary management. In **Phase III (Delegate)**, the delegate fully autonomously collects and shares information, making real-time decisions based on user-defined goals and preferences. Privacy filters, decision-making models, and audit logs ensure transparency and accountability. This phased approach enables the delegate to transition from a controlled executor to a fully autonomous agent, balancing privacy

Table 2: Progression of Autonomy and Responsibility in Achieving a Fully Autonomous Meeting Delegate.

	Phase I: Execute	**Phase II: Assist**	**Phase III: Delegate**
Data Boundary	User-defined boundaries	Privacy-protected boundaries	Data accessible by user
Share Information	Only within user-defined boundaries	Some reasoning over sensitive data	Autonomous based on predefined goals and preferences
Collect Information	Explicit requests only	Infer context beyond user instructions	Autonomously collects and reasons based on meeting context
Decision-Making	No decision-making	Propose and ask for approval	Full autonomous decision-making

and increasing decision-making capability while ensuring transparency and accountability.

While our ultimate goal is to achieve Phase III for significantly reducing meeting-related burdens, early-stage deployment can already be beneficial. For instance, a Phase I delegate might be employed in daily project update scrums to present updates and gather progress for alignment. Although similar objectives might be met through offline progress updates, deploying an early-stage system provides practical experience that will inform future advancements toward the system's full potential. Recent HCI study (Leong et al., 2024) also indicated that embodied agents can evoke feelings of presence and trust while aiding decision making. Additionally, phased deployment familiarizes users with the technology, helping to identify overlooked issues and challenges.

6.2 Prototype Implementation and Learnings

Our prototype meeting delegate system aligns with Phase I, consistent with available technologies. To evaluate its practical performance, we tested the system in multiple demo scenarios. As illustrated in Figure 2, one scenario simulated a daily project update scrum involving three human participants and an LLM-powered delegate. All participants were aware of the delegate's presence and located in the same room. One participant acted as the moderator, while the others, including the delegate, provided project updates. Each human participant followed a script, requesting information from the delegate, which was preloaded with project-related topics via the Information Gathering module. The moderator guided the meeting, with responses cued or initiated by the participants. The demo lasted about five minutes and was repeated to assess the delegate's consistency using different LLMs.

The system is implemented on a widely used meeting platform.[3]). The transcripts are obtained directly from meeting platform, by locating and identifying the UI element associated with "Transcript" and logs the contents to be used in the Meeting engagement module. The response generation prompt incorporates general instructions, user-provided meeting details, and ongoing context (see Table 15 in the Appendix). To minimize latency, the system employs streaming modes for both LLM API calls and TTS (Qin et al., 2023).

We evaluated three models: GPT-3.5-Turbo, GPT-4, and GPT-4o. GPT-3.5-Turbo underperformed, proving inadequate for meeting delegation tasks, even at Phase I. GPT-4 and GPT-4o generally delivered relevant responses but occasionally repeated information from earlier transcripts. Response latency was another issue, with the fastest model, GPT-4o, taking ∼5 seconds to respond.

To mitigate irrelevant and repetitive responses, future improvements may involve leveraging advanced general LLMs or fine-tuning smaller models. Benchmark results show that Llama3-8B performs well, with fine-tuning reducing latency to 500 ms in real-time communication (Cerebrium, 2024). Enhancements such as windowed context management, advanced summarization, and multimodal models with direct speech input/output (OpenAI, 2024b) can further improve real-time performance and maintain response quality. For example, GPT-4o-Realtime-Preview, built on the same architecture as GPT-4o, is expected to offer similar language understanding and reasoning. The inclusion of speech features, such as speed and tone, may further enhance system performance.

7 Conclusion

We introduce an LLM-powered meeting delegate system addressing challenges in collaborative work. Focusing on participant roles, our prototype and benchmark demonstrate LLMs' potential to improve meeting efficiency. Real-world evaluation reveals varying performance, with strengths and areas for improvement, particularly in handling transcription errors and reducing irrelevant or repetitive responses. Future work should enhance real-time responsiveness and privacy safeguards to fully realize LLMs' role in collaborative environments.

[3]Omitting the platform name for anonymity.

Limitations

We acknowledge several limitations in our study. First, the evaluation is restricted to a set of representative language models. While this provides valuable insights, future work should explore a broader range of LLMs, particularly models specifically fine-tuned for meeting-related tasks. Additionally, recent advancements such as OpenAI's Realtime API (OpenAI, 2024b), which supports direct voice input and output, could enhance the relevance of our findings in multimodal contexts.

Second, our benchmark is largely based on limited experimental conditions. Future evaluations should incorporate more diverse and dynamic environments, such as diverse meeting types and domains, to provide a more comprehensive understanding of our system's capabilities.

Lastly, while our system shows promise in facilitating meeting participation, it represents an initial exploration of the possibility of using LLMs as meeting delegates. Specifically, it does not extensively address other key dimensions such as privacy, security, or user trust. In the following section, we share an initial discussion on responsible AI and ethics consideration to outline potential directions for further investigation.

Ethics Statement

This paper explores the potential use of LLMs as meeting delegates, raising several ethical considerations. We propose a phased approach to AI autonomy, starting with limited decision-making in earlier phases and building toward greater capabilities with accountability measures. Privacy-by-design principles should be central to the system's architecture, and educating users about the AI's limitations will ensure responsible use. Below, we outline key ethical dimensions (Bender et al., 2021; Kasneci et al., 2023; Wang et al., 2024; Kirk et al., 2024), including bias, privacy, transparency, human agency, security, and socio-economic impact, alongside suggested safeguards.

Bias and Fairness: LLMs may generate biased or inappropriate content, potentially affecting fairness in meeting outcomes. This risk requires bias detection and mitigation strategies, such as training on diverse datasets, bias audits, and user feedback loops. Fine-tuning models for meeting scenarios and ongoing bias monitoring could be crucial for ensuring fairness.

Privacy: Personalization is only possible by collecting user data. This applies to any technology that relies on personal information to deliver tailored benefits. The personalization of meeting delegates relies on sensitive user data, which risks oversharing or misusing private information. To address this, we advocate for privacy-enhancing technologies like encryption and differential privacy, as well as user-defined data boundaries. Real-time voice capabilities also heighten the risk of identity misuse, necessitating strict privacy controls to ensure compliance with data protection standards.

Transparency: Transparency is essential for responsible deployment. All participants must be informed when an AI is acting as a delegate. Clearly stating the AI's capabilities and limitations helps manage expectations, and audit logs should be available for users to track AI actions and decisions during meetings.

Human Agency: LLM-based delegates should support, not replace, human decision-making. In the early phases, the AI assists users without autonomy, and even in later phase like Phase III, human oversight must remain integral. Human-in-the-loop HITL systems are crucial for maintaining control and ensuring users can intervene as needed.

Security and Fraud Risks: Unauthorized access to a meeting delegate could lead to fraud or impersonation. Security measures like multi-factor authentication, identity verification, and anomaly detection are essential. Federated learning could further protect sensitive data by minimizing centralized storage risks.

Ethical Governance and Mitigation: Ethical governance frameworks, including guidelines, audits, and interdisciplinary collaboration, must guide the system's development. User consent should be obtained at key stages, and continuous monitoring is essential to identify and address unintended consequences.

Socio-Economic Impact: Automating meeting participation could lead to job displacement in roles that rely on meeting facilitation. While this risk is limited by current technology, future developments may amplify these concerns. It's essential to focus on augmenting human labor rather than replacing.

References

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwan, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational challenges in assuring alignment and safety of large language models. *Preprint*, arXiv:2404.09932.

Sumit Asthana, Sagih Hilleli, Pengcheng He, and Aaron Halfaker. 2023. Summaries, highlights, and action items: Design, implementation and evaluation of an llm-powered meeting recap system. *Preprint*, arXiv:2307.15793.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Cerebrium. 2024. Fast voice agent. Accessed: 2025-02-12.

Amir Cohen, Amir Kantor, Sagi Hilleli, and Eyal Kolman. 2021. Automatic rephrasing of transcripts-based action items. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2862–2873, Online. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Google. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Google. 2024b. Gemini models. Accessed: 2025-02-12.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Reid Hoffman. 2024. Reid hoffman meets his ai twin - full. Accessed: 2025-02-12.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Hannah R. Kirk, Bertie Vidgen, Paul Röttger, et al. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6:383–392.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.

Danielle Kost. 2020. You're right! you are working longer and attending more meetings. Accessed: 2025-02-12.

Joanne Leong, John Tang, Edward Cutrell, Sasa Junuzovic, Gregory Paul Baribault, and Kori Inkpen. 2024. Dittos: Personalized, embodied agents that participate in meetings when you are unavailable. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods.

Manqing Mao, Paishun Ting, Yijian Xiang, Mingyang Xu, Julia Chen, and Jianzhe Lin. 2024. Multi-user chat assistant (muca): a framework using llms to facilitate group conversations. *Preprint*, arXiv:2401.04883.

Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1220–1230, Baltimore, Maryland. Association for Computational Linguistics.

Meta. 2024. Meta llama. Accessed: 2025-02-12.

Marino Mugayar-Baldocchi, Bill Schaninger, and Kartik Sharma. 2021. The future of the workplace: Embracing change and fostering connectivity. Accessed: 2025-02-12.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondrej Bojar. 2022. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. In *International Conference on Language Resources and Evaluation*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024a. Hello gpt-4o. Accessed: 2025-02-12.

OpenAI. 2024b. Introducing the realtime api. Accessed: 2025-02-12.

OpenAI. 2024c. Models. Accessed: 2025-02-12.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Leslie A. Perlow, Constance Noonan Hadley, and Eunice Eun. 2017. Stop the meeting madness. Accessed: 2025-02-12.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MeetingQA: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15000–15025, Toronto, Canada. Association for Computational Linguistics.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *Preprint*, arXiv:2312.01479.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Jared Spataro. 2020. Remote work trend report: meetings. Accessed: 2025-02-12.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R. Fung, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. 2024. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *Preprint*, arXiv:2402.11060.

Don Tuggener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way? a survey of dialogue summarization data sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 107–118, Online and in Dominican Republic. Association for Computational Linguistics.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models should not replace human participants because they can misportray and flatten identity groups. *Preprint*, arXiv:2402.01908.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (llms): A survey. *Preprint*, arXiv:2403.05156.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Definition of Meeting Scenes

Explicit Cue: These are clear, direct prompts addressed by name in a meeting that indicate when a participant should speak. For example, “Amy, please share your thoughts on this” or “Amy, any questions?” serve as overt invitations for participation or feedback.

Implicit Cue: These cues are subtle, context-dependent signals that are not directly stated but inferred from the conversation. Participants typically continue to contribute when not explicitly addressed by name, based on the natural flow of the discussion.

Chime In: This case refers to the act of a participant proactively interjecting or contributing to the discussion, usually when other participants are in discussion. “Chime in” moments occur when an individual adds supplementary information or clarifies a point.

B Dataset Construction

An example of evaluation dataset construction is shown in Figure 8. In the meeting transcript, participants are represented by different ID numbers and icons. Each utterance is displayed in colored boxes, with each color representing a different participant. In this example, we construct a test case with Participant 6 as the principal. Based on Participant 6’s utterances in the Original Transcript, we extract one piece of Input Context Information: when the meeting discusses expertise in emotion detection, Participant 6 intends to mention related experience from bachelor thesis. The Transcript Snapshot and Ground-Truth Response are extracted from the Original Transcript using GPT-4. During the response generation stage with the meeting delegate, the Transcript Snapshot is provided to the LLMs to produce a response. This generated response is subsequently assessed by comparing it to the Ground-Truth Response.

We plan to release our constructed benchmark dataset with the paper.

C Semantic Similarity Scoring

To strengthen interpretability, we complemented our Recall and Attribution metrics with semantic similarity scoring using a Sentence-BERT model (all-MiniLM-L6-v2). This allowed us to assess how well model-generated responses align at a semantic level with human-authored ground-truth responses. This score offers a more interpretable baseline of what current models can achieve in realistic conditions.

Cue Type	GPT-4o	Gemini 1.5 Pro	Llama3-8B
Explicit Cue	0.487	0.435	0.435
Implicit Cue	0.458	0.426	0.415
Chime In	0.436	0.382	0.402

Table 3: Semantic similarity scores showing alignment between model-generated responses and human-authored ground truth

In Explicit Cue cases, where expectations are clearer, models reached similarity scores close to 0.49, suggesting reasonable semantic alignment with human responses. In Implicit Cue and Chime In cases, lower similarity reflects the inherent ambiguity and the challenge of determining when and how to respond without explicit prompting. Combined with our Recall metrics (e.g., 60% of responses covering at least one main point), these results provide a practical signal: current state-of-the-art LLMs like GPT-4o can produce semantically relevant and appropriately timed responses in clear-cut situations, but face challenges in nuanced or indirect participation roles. Rather than proposing a one-size-fits-all deployment threshold, we suggest interpreting these metrics as task-dependent readiness indicators. For example, a system achieving > 0.45 similarity and > 0.60 loose Recall in Explicit Cue cases may be sufficient for low-stakes delegation tasks (e.g., daily standups), while more complex settings would demand tighter control and future refinement.

D Representative Evaluation Example

Representative Evaluation Example

Ground-Truth Response:
PERSON10: Yeah, yeah. Well, I just like listened to the audio and followed the talk transcript [object Object] and it was completely off. I think it is-. There must be some miss-match because-. (id=16)
PERSON10: Yeah, yeah. The transcription is for the completely different audio than it's in the subdirectory. (id=18)
PERSON10: Yeah, yeah. I I may maybe it's just like uh. Maybe the files are just switched between the subdirectories? (id=21)
PERSON10: I I haven't checked but-. Uh, yeah, there is some some serious mismatch there. (id=23)

Ground-Truth Main Points:

- Listened to the audio and found that the transcription was completely off
- Identified a serious mismatch between the transcription and the audio in the subdirectory
- Suggested that files might be switched between subdirectories, causing the mismatch

GPT-Generated Delegate Response:
"Sure, PERSON13. The transcription seems to be for a completely different audio than what's in the subdirectory. There might be a serious mismatch, possibly because the files are switched between the subdirectories. This could explain why the Czech video is imprecise in timing."

GPT-4 Evaluation Output:
"Explanation": "First, summarize the main points in the [object Object]. The [object Object] has the following main points: 1. The transcription seems to be for a completely different audio than what's in the subdirectory. 2. There might be a serious mismatch, possibly because the files are switched between the subdirectories. 3. This could explain why the Czech video is imprecise in timing. The number of main points in the [object Object] is 3 and the length of the MatchingIndex is 3.
Second, compare the main points in the [object Object] with the main points in the [object Object]. The matching main points between the [object Object] and the [object Object] are: 1. The transcription seems to be for a completely different audio than what's in the subdirectory. This point matches with the first point in the [object Object]. 2. There might be a serious mismatch, possibly because the files are switched between the subdirectories. This point matches with the second and third points in the [object Object]. The number of matching main points between the [object Object] and the [object Object] is 2. The third point in the [object Object] does not directly match any points in the [object Object] but provides additional context to the mismatch issue.",
"MatchingIndex": [1, 2, -1], "MatchingMainPointsCount": 2

Interpretation:

- GPT-4-based evaluation accurately identifies semantically similar content even when phrased differently.
- The additional point included by the model is correctly excluded from recall, highlighting the discrimination power of the method.
- The correct attribution reflects alignment between generated responses and user expectations, supporting its use as a reliable evaluation proxy at scale.

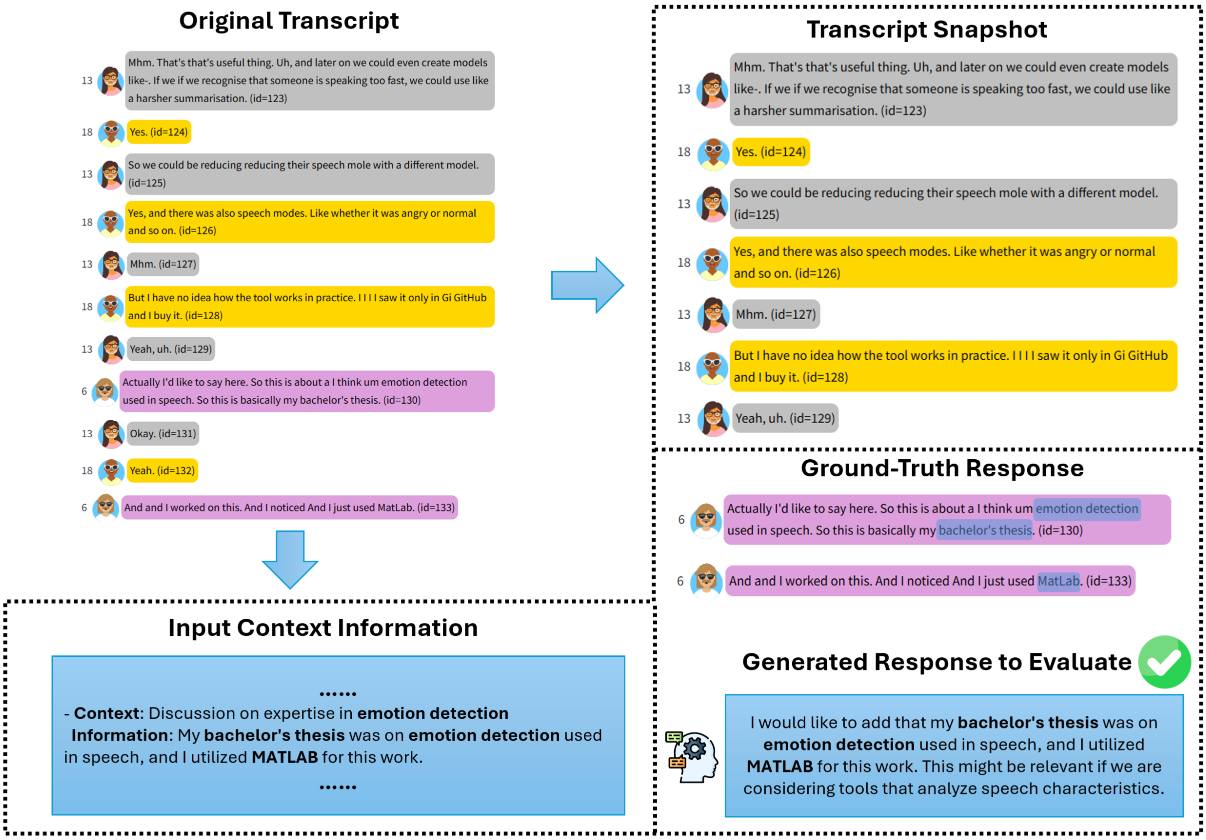


Figure 8: Example of evaluation dataset construction. Participants are represented by different ID numbers and icons. Colored boxes indicate utterances from different participants. The process includes extracting Input Context Information, creating a Transcript Snapshot, and generating a response with the LLM-powered meeting delegate. The Generated Response is evaluated by comparison with the Ground-Truth Response.

E Additional Experimental Results

In this section, we provide detailed tables and additional plots for the experimental results discussed in Section 5.

Response Rate Analysis. Tables 4 and 6 present the Response Rate and Silence Rate of LLMs evaluated using the Matched and Mismatched Datasets, respectively. Additionally, in Tables 5 and 7, we further evaluate the Response Rate and Silence Rate using the intersection subdataset of all models, given that Llama models and GPT-3.5 have smaller context windows. The findings from these experimental results remain consistent.

Response Rate Failure Cases Study. The error types distribution for response rate failure cases study in Matched and Mismatched datasets are presented in Figure 9. The mappings between error types and improvement solution direction are summarized in Table 1.

Recall Analysis. The loose recall rate and strict recall rate for the Matched Dataset are shown in Table 8. We further evaluate the recall rates using the intersection subdataset of all models, with results presented in Table 9. Although the absolute values of recall rates for all models are higher, the performance differences among the models are similar. Note that we do not include Llama3-8B and Llama3-70B here in the intersection study to avoid too few samples. The findings from these experimental results remain consistent.

Attribution Analysis. The attribution metrics for LLMs are included in Table 10. We also evaluate the attribution metrics using the intersection subdataset. Note that we do not include Llama3-8B and Llama3-70B here in the intersection study to avoid too few samples. The findings from these experimental results remain consistent.

Correlation Study. The correlation of response rate and recall metrics with test case metadata is shown in Figure 10. No significant correlations is found between these metrics and the metadata.

Ablation Study. The response rates of LLMs for the Noisy Name Dataset are presented in Table 12, with the response rates from Explicit Cue in Matched Dataset are also shown for reference. A significant drop in performance is observed for all models, except for GPT-3.5 where responses rates are already low. This further highlights challenges in accurately recognizing participant names. Further model fine-tuning to better handle such transcription errors may be necessary. For the No-<Context> study, all evaluation metrics for GPT-4o in No-<Context> Scenario are shown in Table 13, showing minimal performance impact across all evaluation metrics when context information is omitted.

F Model Specifications

In Table 14, we list all LLMs utilized in this paper, along with their detailed model version and usage scenarios.

G Prompts

We include all prompts used in the paper. Table 15 provides the prompt for generating the response in the Meeting Engagement module. The prompts used for evaluating and attributing the generated response are given in Tables 18 and 20, respectively. Lastly, the prompts for extracting context information and extracting test cases from meeting transcripts are given in Table 22 and Table 28, respectively.

Table 4: Response Rate for Matched Dateset.

Type	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Chime In	39.3%	37.9%	61.3%	71.8%	41.9%	84.1%	**93.8%**
Explicit Cue	53.2%	86.7%	94.3%	89.7%	78.3%	91.2%	**99.4%**
Implicit Cue	52.2%	67.2%	71.9%	83.6%	55.9%	90.0%	**94.8%**
All	50.6%	68.9%	77.3%	83.8%	60.8%	89.6%	**96.2%**

Table 5: Response Rate for Intersection Subset of Matched Dateset.

Type	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Chime In	35.2%	42.3%	57.7%	66.2%	43.7%	81.7%	**95.8%**
Explicit Cue	58.6%	92.0%	92.0%	87.7%	76.5%	89.5%	**98.1%**
Implicit Cue	54.3%	65.8%	68.3%	81.9%	53.5%	89.7%	**94.7%**
All	52.9%	71.2%	74.8%	81.5%	59.9%	88.4%	**96.0%**

Table 6: Silence Rate for Mismatched Dataset.

Type	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Explicit Cue	75.0%	84.6%	82.8%	65.0%	**88.1%**	36.0%	41.6%
Implicit Cue	70.4%	**79.5%**	67.9%	52.0%	77.1%	35.3%	33.3%
All	72.4%	81.6%	73.6%	57.5%	**81.7%**	35.6%	37.0%

Table 7: Silence Rate for Intersection Subset of Mismatched Dataset.

Type	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Explicit Cue	79.5%	84.9%	**90.4%**	76.7%	**90.4%**	37.0%	44.6%
Implicit Cue	69.5%	**81.7%**	74.4%	58.5%	**81.7%**	35.4%	31.9%
All	74.2%	83.2%	81.9%	67.1%	**85.8%**	36.1%	38.7%

Table 8: Recall Rate for Matched Dataset.

Model	Chime In		Explicit Cue		Implicit Cue		All	
	Loose	Strict	Loose	Strict	Loose	Strict	Loose	Strict
GPT-3.5	43.5%	29.5%	54.5%	42.5%	47.8%	37.0%	49.5%	38.0%
GPT-4	51.1%	39.9%	72.8%	60.7%	**63.0%**	**49.6%**	65.9%	53.1%
GPT-4o	**53.9%**	**47.0%**	**77.8%**	**64.2%**	62.5%	47.9%	**67.3%**	**53.9%**
Gemini 1.5 Flash	29.2%	22.5%	69.5%	56.5%	55.0%	40.2%	56.6%	43.4%
Gemini 1.5 Pro	34.6%	28.8%	72.8%	59.9%	56.0%	43.5%	60.5%	48.6%
Llama3-8B	46.7%	35.5%	59.6%	48.7%	52.7%	40.5%	54.2%	42.6%
Llama3-70B	45.8%	34.7%	69.6%	59.4%	55.9%	44.0%	59.1%	47.9%

Table 9: Recall Rate for Intersection Subset of Matched Dataset. Note that due to limited statistics for intersecting Llama results, Llama results are not included. The total number of cases in the considered Intersection Subset is 196.

Model	Chime In		Explicit Cue		Implicit Cue		All	
	Loose	Strict	Loose	Strict	Loose	Strict	Loose	Strict
GPT-3.5	55.6%	47.2%	58.4%	46.9%	56.1%	45.2%	57.1%	46.0%
GPT-4	**77.8%**	**52.8%**	79.8%	66.7%	70.4%	55.8%	75.0%	60.6%
GPT-4o	66.7%	**52.8%**	**85.4%**	**70.6%**	**79.6%**	**59.8%**	**81.6%**	**64.4%**
Gemini 1.5 Flash	44.4%	32.2%	79.8%	64.6%	67.3%	49.3%	71.9%	55.4%
Gemini 1.5 Pro	22.2%	19.4%	77.5%	62.6%	60.2%	46.2%	66.3%	52.4%

Table 10: Attribution Analysis results for Matched Dataset. For the Expected Response metric, higher values are better, while for the Previous Transcript and Hallucination metrics, lower values are preferable.

Metric	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Chime In							
Expected Response	18.4%	27.0%	**42.2%**	25.3%	35.2%	25.4%	27.4%
Input Context Info	37.2%	43.0%	37.9%	39.2%	26.3%	39.6%	31.4%
Previous Transcript	33.7%	25.1%	**15.9%**	28.6%	28.1%	31.6%	32.4%
Hallucination	10.8%	4.93%	4.05%	6.93%	10.4%	**3.43%**	8.75%
Explicit Cue							
Expected Response	32.5%	50.1%	51.0%	52.4%	**61.1%**	31.9%	50.1%
Input Context Info	40.4%	31.6%	38.1%	27.0%	26.3%	36.8%	28.3%
Previous Transcript	20.8%	12.4%	**7.28%**	14.4%	9.25%	25.4%	16.8%
Hallucination	6.43%	5.98%	3.58%	6.24%	**3.35%**	5.82%	4.81%
Implicit Cue							
Expected Response	25.2%	39.9%	38.9%	38.0%	**46.8%**	28.2%	38.9%
Input Context Info	39.9%	39.9%	45.9%	34.2%	32.0%	34.3%	34.1%
Previous Transcript	31.9%	15.0%	**11.3%**	22.4%	14.8%	35.7%	22.6%
Hallucination	2.96%	5.12%	3.8%	5.48%	6.32%	**1.80%**	4.38%
All							
Expected Response	26.9%	42.8%	43.9%	41.5%	**51.6%**	29.1%	41.2%
Input Context Info	39.8%	36.9%	42.0%	32.3%	29.2%	35.8%	31.8%
Previous Transcript	28.4%	14.8%	**10.3%**	20.3%	13.8%	31.6%	21.9%
Hallucination	4.95%	5.44%	3.74%	5.91%	5.48%	**3.39%**	5.09%

Table 11: Attribution Analysis results for Intersection Subset of Matched Dataset. For the Expected Response metric, higher values are better, while for the Previous Transcript and Hallucination metrics, lower values are preferable. Note that due to limited statistics for the intersecting Llama results, Llama results are not included. The total number of cases in the considered Intersection Subset is 196.

Metric	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro
Chime In					
Expected Response	22.0%	30.9%	**35.8%**	29.2%	22.2%
Input Context Info	55.7%	58.1%	64.2%	45.8%	22.2%
Previous Transcript	11.1%	5.0%	**0.0%**	12.5%	44.4%
Hallucination	11.1%	5.9%	**0.0%**	12.5%	11.1%
Explicit Cue					
Expected Response	37.4%	59.1%	56.1%	59.9%	**66.9%**
Input Context Info	37.9%	27.7%	36.4%	23.3%	19.5%
Previous Transcript	19.6%	10.1%	**3.3%**	11.7%	12.5%
Hallucination	5.1%	5.98%	3.1%	5.1%	**1.2%**
Implicit Cue					
Expected Response	30.6%	47.3%	49.9%	49.4%	**51.3%**
Input Context Info	42.6%	36.4%	38.6%	31.3%	29.4%
Previous Transcript	23.5%	12.2%	**7.0%**	17.6%	12.1%
Hallucination	3.3%	4.0%	4.5%	**1.7%**	7.1%
All					
Expected Response	33.3%	51.8%	52.1%	53.4%	**57.0%**
Input Context Info	41.1%	33.5%	38.8%	28.2%	24.5%
Previous Transcript	21.2%	10.9%	**5.0%**	14.7%	13.8%
Hallucination	4.5%	**3.7%**	4.1%	**3.7%**	4.6%

Table 12: Response rate for Noisy Name Dataset.

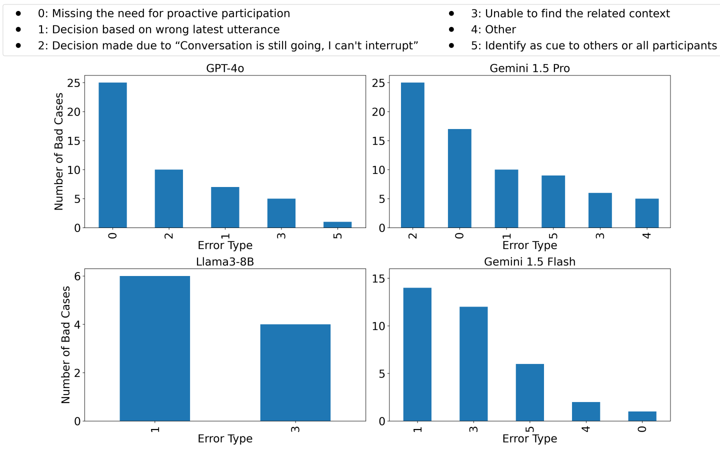
Type	Dataset	GPT-3.5	GPT-4	GPT-4o	Gemini 1.5 Flash	Gemini 1.5 Pro	Llama3-8B	Llama3-70B
Explicit Cue	Matched	53.2%	86.7%	94.3%	89.7%	78.3%	91.2%	**99.4%**
Explicit Cue	Noisy Name	52.5%	53.3%	68.0%	60.7%	59.8%	79.4%	**87.0%**

Table 13: All Evaluation Metrics for GPT-4o in No-<Context> Scenario.

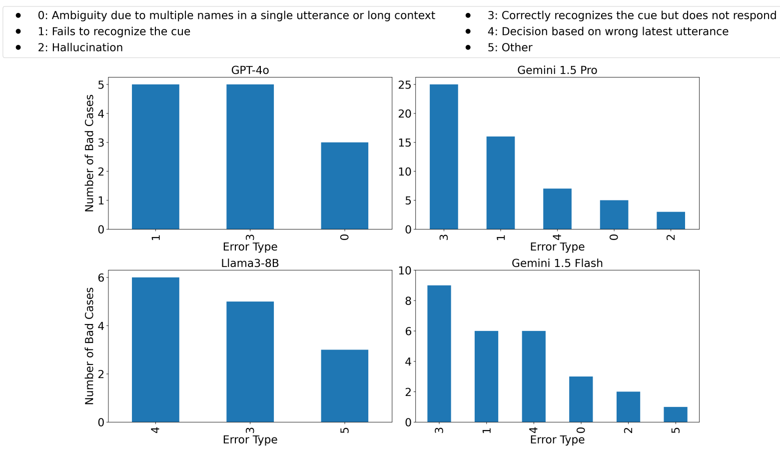
Metric	Chime In	Explicit Cue	Implicit Cue	All
Response Rate	59.1%	90.4%	78.7%	80.2%
Loose Recall	46.2%	82.6%	75.0%	74.7%
Strict Recall	37.7%	65.0%	50.2%	55.8%
Expected Response	21.0%	44.1%	44.7%	41.1%
Input Context Info	57.2%	36.0%	31.9%	37.3%
Previous Transcript	14.1%	14.4%	14.0%	14.2%
Hallucination	7.7%	5.4%	9.4%	7.3%

Table 14: Details of Model Use Scenarios and Model Version.

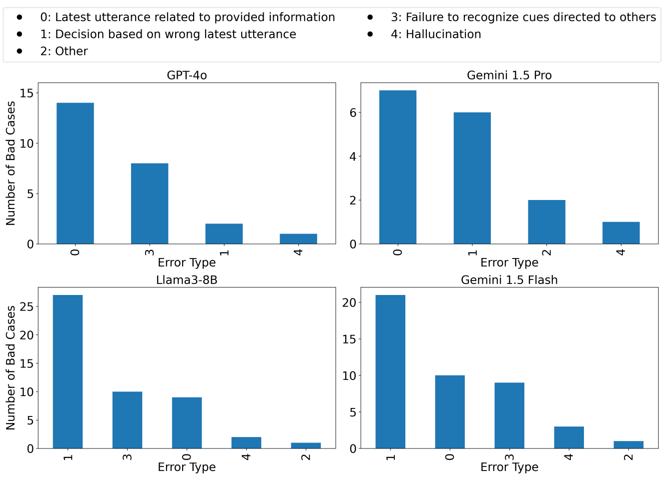
<table>
<tr><th>Model Name</th><th>Model Use Scenarios</th><th>Model Version</th></tr>
<tr><td>GPT-3.5</td><td>Generate Response (Table 4 & Table 5 & Table 6 & Table 7, Prompt in Table 15)</td><td>gpt-3.5-turbo-1106 with 16k context window</td></tr>
<tr><td rowspan="3">GPT-4</td><td>Generate Response (Table 4 & Table 5 & Table 6 & Table 7, Prompt in Table 15)</td><td>gpt-4-turbo-20240409 with 128k context window</td></tr>
<tr><td>Evaluation (Table 8 & Table 9, Prompt in Table 18)</td><td>gpt-4-1106-preview with 128k context window</td></tr>
<tr><td>Attribution (Table 10 & Table 11, Prompt in Table 20)
Extract context information (Figure 8, Prompt in Table 22)
Extract test cases (Figure 8, Prompt in Table 28)</td><td>gpt-4-turbo-20240409 with 128k context window</td></tr>
<tr><td>GPT-4o</td><td>Generate Response (Table 4 & Table 5 & Table 6 & Table 7, Prompt in Table 15)</td><td>gpt-4o-20240513-preview with 128k context window</td></tr>
</table>



(a) Chine In (Matched Dataset)



(b) Explicit Cue (Matched Dataset)



(c) Mismatched Dataset

Figure 9: (a) Error Types Distribution for Response Rate Failure Cases Study in Chine In Matched Dataset. (b) Error Types Distribution for Response Rate Failure Cases Study in Explicit Cue Matched Dataset. (c) Error Types Distribution for Response Rate Failure Cases Study in Mismatched Dataset.

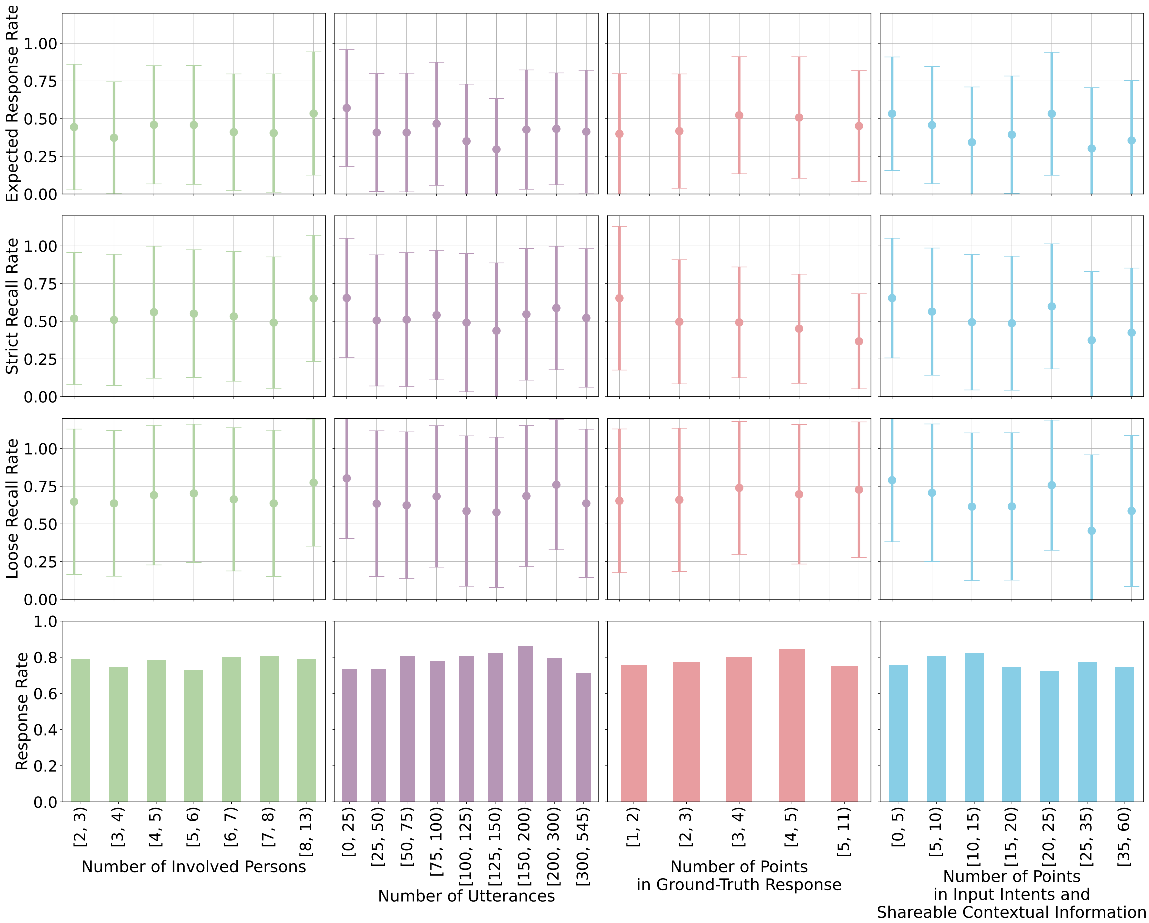


Figure 10: The correlation between the performance metrics and test case metadata for GPT-4o.

- You are a Meeting Delegate Agent that attends meetings on behalf of <Person Name>.
- You are provided with the intent of participating in the meeting, specified as <Intents>.
- You are provided with the background information that <Person Name> knows, specified as <Background>.
- You are provided with the full list of attendees <Attendees> to help identify if someone cues you.
- You are provided with the ongoing meeting transcript <Meeting Transcript> to determine if there is a need to respond.
- Your task is to assess the content of the ongoing meeting transcript <Meeting Transcript> and determine whether you are can speak and what to say.
- You are encouraged to respond and ask questions, give comments, or share information without interrupting others in the meeting.

About the <Person Name>
- <Person Name> is the name of the person you represent in the meeting.
- People in the <Attendees> list may cue you by using <Person Name> exactly or parts of the name (e.g., first name, initials).

About the <Attendees>
- <Attendees> is a list of names of the people attending the meeting.
- Each name in the list is a full name or a nickname.

About the <Meeting Transcript>
- <Meeting Transcript> is a series of utterances spoken by the meeting participants.
- Each utterance is formatted as "Name: Content", where 'Name' is the speaker's name and 'Content' is their spoken text.
- The utterances are in chronological order and the latest utterance is at the bottom of the transcript.
- The utterances may contain typos and grammatical errors.

About the <Intents>
- <Intents> consists of the questions or topics that <Person Name> aims to discuss during the meeting.
- You can ask the questions or motivate the discussion of the topics in the <Intents> at the appropriate time without interrupting others.

About the <Background>
- <Background> consists of the background information that <Person Name> knows before the meeting.
- <Background> is a list of "Context" and "Information" pairs. You can share the "Information" in the "Context" at the appropriate time without interrupting others.

Guidelines to judge whether you can speak and decide what to say
- Read the <Meeting Transcript> to understand the context of the meeting.
- Focus on the latest several utterances in the <Meeting Transcript> to understand the current discussion.
- Remember that you are a delegate attending the meeting on behalf of <Person Name>.
- You should judge whether you can speak first, then decide what to say, if you can speak.
- Judge whether you can speak according to the following instructions:
    - Figure out what the latest utterance (at the bottom of the <Meeting Transcript>) is about and pay attention to who is being addressed.
    - If the latest utterance is a straightforward question or request or instructions to other participants, you MUST NOT speak to avoid interrupting others, even if the conversation is related to the <Intents> or <Background>.
    - If the latest utterance is for the <Person Name>, you should respond to it.
    - If you can speak, consider the following guidelines:
    - Your speech content should be directly relevant to the current discussion.
    - You can reference the <Intents> and <Background> to organize your speech.
    - You should be polite and natural in your speech.
    - You MUST NOT make up facts.
    - You MUST NOT repeat what <Person Name> has said in the <Meeting Transcript>.
    - Chit chat is a natural part of conversation. You can engage in chit chat with other attendees if it is appropriate or relevant to the meeting context. For example, you can say good morning, Thank you, Yeah, I agree.
    - Before speaking, you should think twice to ensure that you are not interrupting others and your speech is relevant to the current discussion.

Notes on judging whether someone is cued
- The name may be transcribed as similar-sounding words by the speech recognition system. Especially, the pronunciation of Chinese names may be recognized as similar-sounding English words, for exmaple, "Si Li" may be transcribed as "Celine" or "silence".
- When encountering words that seem out of place, it is likely due to errors in speech recognition. Examine the list of attendees to determine if the pronunciation of these words are similar to any English or Chinese names listed.
- You should consider the context of the meeting and the names of the attendees to determine if you or someone are cued.

CONTINUE ON THE NEXT PAGE

Table 15: Prompt used for generating response in the Meeting Engagement module.

The output of response <Response>
- The response should be a dictionary with the following format:

```
{
    "thoughts": "<thoughts>",
    "speak": "<speak>"
}
```

- <thoughts>: The reasoning or considerations to judge whether you can speak and decide what to say. At the beginning, you should state who you are representing. Then <thoughts> should explain what the latest utterance is about and then explain why you can or cannot speak. If you can speak, you should also explain how you decide what to say.
- <speak>: The content you are going to speak. If you are not allowed to speak or you do not want to speak, the <speak> is empty.

Example

```
- Example 1:
Below is an example of that pronunciations of Chinese names may be recognized as similar-
 sounding English words by the speech recognition system.
<Person Name>
'Sirui Zhao'
<Attendees>
[
    - 'San Zhang',
    - 'Si Li',
    - 'Sirui Zhao'
]
<Meeting Transcript>
Si Li: Good morning.
Sirui Zhao: Hello!
San Zhang: Hi!
Si Li: OK, Let's start our meeting. There are still some people who haven't joined, so let'
 s start first. Our topic today is the progress of environmental protection, three, do you
 have some thing to share on it?
<Intents>
[
    - 'The extent of plastic misuse'
]
<Background>
[
    {
        "Context":"Discussion about reducing air pollution",
        "Information":"The air pollution of our city is becoming serious. The goverment
 takes extreme measures to control the pollution by closing the factory and limiting the
 use private car."
    }
]
<Response>
{
    - "thoughts": "I'm representing Sirui Zhao in the meeting. In the last utterance, the
 appearance of 'three' is abrupt. Contextually, there is no need for numbers; phonetically,
  "Three" sounds like "Sirui.", which closely resembles 'Sirui' from the attendees,
 specifically <Person Name>. The speaker is most likely asking Sirui Zhao to share
 something on the progress of environment protection. So I need to give response. And based
  on the background information, I can share something about reduing air pollution.",
    - "speak": "Yes. The air pollution of our city is becoming serious. The goverment
 takes extreme measures to control the pollution by closing the factory and limiting the
 use private car."
}
```

CONTINUE ON THE NEXT PAGE

Table 16: Prompt used for generating response in the Meeting Engagement module (continued).

```
- Example 2:
Below is an example of that you should not speak since the latest utterance is a
 straightforward question or request or instructions to other participants.
<Person Name>
'Frank'
<Attendees>
[
    - 'John',
    - 'James',
    - 'Alice',
    - 'Bob'
    - 'Frank'
]
<Meeting Transcript>
Bob: James, that price is too high, we can not accept it.
James: Ok, I will contact the supplier again and discuss the price.
John: Thank you, James.
John: OK, Let's go to the next topic. Alice, what is your progress on the project
 development?
<Intents>
[
    - 'Whether Bob fixed the bug I reported'
]
<Background>
[
    {
        "Context":"Report on the dataset preparation progress",
        "Information":'The dataset preparation is almost done. We are now working on the
 data cleaning and normalization. We expect to finish it by the end of the week.'
    }
]
<Response>
{
    - "thoughts": "I am representing Frank in the meeting. In the latest utterance, John
 is explicitly asking Alice about the project development. I can not speak.",
    - "speak": ""
}
```

Note
- You are representing <Person Name> in the meeting. You should respond to the cues from the attendees and the context of the meeting.
- You should not interrupt others in the meeting.

Table 17: Prompt used for generating response in the Meeting Engagement module (continued).

- You are an Evaluation Agent responsible for assessing the response generated by a meeting AI assistant against the standard answer.
- You are provided with the summary of the <StandardAnswer>.
- You are provided with the raw <ActualResponse> generated by the meeting AI assistant.
- Your task is to summarize the main points in the <ActualResponse>, and evaluate whether the main points in the <ActualResponse> match the main points in the <StandardAnswer>.

About the meeting AI Assistant
- The meeting AI assistant is designed to represent the user to engage in a meeting.

About the <StandardAnswer>
- The <StandardAnswer> is a list of strings that represents the main points of the ground truth response.

About the <ActualResponse>
- The <ActualResponse> is a string that represents the response generated by the meeting AI assistant to the meeting content.
- You should reference the <Transcript> to understand the context of the meeting and the information in the <ActualResponse>.

Guidelines for Evaluation
- The evaluation process involves comparing the main points in the <StandardAnswer> and the <ActualResponse>.
- Summarize the main points in the <ActualResponse> and keep the same granularity as the <StandardAnswer>.
- The uninformative utterance about expressing goodness or politeness should not be considered as main points.
  - For example, "If you need more help, please let me know." is not informative and should not be considered as a main point.
- You should calculate a list that contains the index of the matching main points in the <ActualResponse> corresponding to the <StandardAnswer>. For example, if the first main point in the <ActualResponse> matches the second main point in the <StandardAnswer>, the first element of the list should be 2. And if the total number of main points in the <ActualResponse> is 1, the list should be [2].
- Count the number of main points in the <ActualResponse> (ActualMainPointsCount).
- Count the number of matching main points between the <ActualResponse> and the <StandardAnswer> (MatchingMainPointsCount).
- The main points are considered matching if they are semantically similar.

Output Format
- The output MUST be in the JSON format.
- You MUST explain the process of evaluation before providing the evaluation results.
- The output MUST include the following fields:
  - Explanation: A explanantion of steps involved in the evaluation process. First, you should summarize the main points in the <ActualResponse>. Then, you should explain which main points in the <ActualResponse> match the main points in the <StandardAnswer> and mark the index of the matching main points in the <ActualResponse> corresponding to the <StandardAnswer> main points.
  - ActualMainPoints: The list of main points in the <ActualResponse>.
  - ActualMainPointsCount: The number of main points in the <ActualResponse>.
  - MatchingMainPoints: The list of matching main points between the <ActualResponse> and the <StandardAnswer>.
  - MatchingIndex: The list of the index of the matching main points in the <ActualResponse> corresponding to the <StandardAnswer> main points. The length of the list should be the same as the ActualMainPointsCount. If ActualMainPointsCount is 5, the format of the list should be [1, 2, -1, -1, 4], which means the first, second, and fifth main points in the <ActualResponse> match the first, second, and fourth main points in the <StandardAnswer>. And the third and fourth main points in the <ActualResponse> do not match any main points in the <StandardAnswer>.
  - MatchingMainPointsCount: The number of matching main points between the <ActualResponse> and the <StandardAnswer>.
- If the <ActualResponse> is empty, the ActualMainPointsCount, MatchingMainPointsCount, RecallRate, and PrecisionRate should be 0.
- Note that you must keep the length of the MatchingIndex the same as the ActualMainPointsCount, instead of the length of the <StandardAnswer>.

CONTINUE ON THE NEXT PAGE

Table 18: Prompt used for evaluating the generated response against the ground-truth one.

```
## Example1
<StandardAnswer>
["Calculated word error rate on Czech transcripts", "Conducted testing sessions with PERSON11 and PERSON18", "
  Contributed to the PROJECT3 deliverable", "Waiting for new tasks"]
<ActualResponse>
"Hi everyone. Over the past week, I calculated the word error rate on Czech transcripts using three versions of Czech
   ASR created by PERSON10. I also conducted a few testing sessions with PERSON11 and PERSON18, but they were not
  successful due to issues with the segmenters from ORGANIZATION1. I have updated the German transcripts in its
  corresponding path. I also backed up all the systems, including some new ones created today and last week. For the
  next week, I am waiting for new tasks. By the way, do we have the golden transcripts for the English videos?"
<Evaluation>
{
    "Explanation": "First, summarize the main points in the <ActualResponse>. The <ActualResponse> has the following
  main points: 1. Calculated word error rate on Czech transcripts. 2. Conducted testing sessions with PERSON11 and
  PERSON18. 3. Updated German transcripts in its corresponding path. 4. Backed up all the systems. 5. Waiting for new
  tasks. 6. Ask about the golden transcripts for the English videos. So, the number of main points in the <
  ActualResponse> is 6 and the length of the MatchingIndex is 6. Sencond, compare the main points in the <
  ActualResponse> with the main points in the <StandardAnswer>. The matching main points between the <ActualResponse>
  and the <StandardAnswer> are: 1. Calculated word error rate on Czech transcripts. This point maches with the first
  point in the <StandardAnswer>. 2. Conducted testing sessions with PERSON11 and PERSON18. This point matches with the
   second point in the <StandardAnswer>. 3. Waiting for new tasks. This point matches with the fourth point in the <
  StandardAnswer>. The number of matching main points between the <ActualResponse> and the <StandardAnswer> is 3. The
  other points in the <ActualResponse> do not match points in the <StandardAnswer>.",
    "ActualMainPoints": ["Calculated word error rate on Czech transcripts", "Conducted testing sessions with PERSON11
   and PERSON18", "Updated German transcripts in its corresponding path", "Backed up all the systems", "Waiting for
  new tasks", "Ask about the golden transcripts for the English videos"],
    "ActualMainPointsCount": 6,
    "MatchingMainPoints": ["Calculated word error rate on Czech transcripts", "Conducted testing sessions with PERSON
  11 and PERSON18", "Waiting for new tasks"],
    "MatchingIndex": [1, 2, -1, -1, 4, -1],
    "MatchingMainPointsCount": 3
}
## Example2
<StandardAnswer>
["Confirm the task about writing a report about the calculation and share it with others", "Synthesize the
  information other team members have shared", "Wait for the next task"]
<ActualResponse>
"Sure, I will finish the calculation. I will also write a report about the calculation."
<Evaluation>
{
    "Explanation": "First, summarize the main points in the <ActualResponse>. The <ActualResponse> has the following
  main points: 1. Confirm the task about finishing the calculation and writing a report about it. The number of main
  points in the <ActualResponse> is 1 and the length of the MatchingIndex is 1. Sencond, compare the main points in
  the <ActualResponse> with the main points in the <StandardAnswer>. The matching main points between the <
  ActualResponse> and the <StandardAnswer> contains: 1. Confirm the task about finishing the calculation and writing a
   report about it. This point matches with the first point in the <StandardAnswer>. The number of matching main
  points between the <ActualResponse> and the <StandardAnswer> is 1. The other points in the <ActualResponse> do not
  match points in the <StandardAnswer>.",
    "ActualMainPoints": ["Confirm the task about finishing the calculation and writing a report about it"],
    "ActualMainPointsCount": 1,
    "MatchingMainPoints": ["Confirm the task about finishing the calculation and writing a report about it"],
    "MatchingIndex": [1],
    "MatchingMainPointsCount": 1,
}
```

Table 19: Prompt used for evaluating the generated response against the ground-truth one (continued).

- You are an Attribution Agent responsible for assessing the response generated by a meeting AI assistant and determining its source.
- You are provided with the list of <ActualResponse>.
- You are also provided with the transcript of the meeting content (<Transcript>) and the <ContextInfo> used to generate the <ActualResponse>.
- Your task is to attribute the <ActualResponse> to the corresponding part of the <Transcript> or the <ContextInfo>.

About the <ActualResponse>
- The <ActualResponse> is a list that represents the response generated by the meeting AI assistant.

About the <StandardResponse>
- The <StandardResponse> is a list that represents the expected response.
- The <StandardResponse> may be the same as <ActualResponse>, or it may not be.

About the <Transcript>
- The transcript are the collection of utterances from the meeting participants.
- Each utterance is formatted as "Name: Content", where 'Name' is the speaker's name and 'Content' is their spoken text.
- Utterances are in chronological order and may contain typos and grammatical errors.
- The transcript ends at the time stamp when the meeting AI assistant should generate the response.
- Example utterances:
  PERSON1: Hello everyone, I'm glad to see you all here today. (id=0)

About the <ContextInfo>
- <ContextInfo> is a dictionary that contains <Intents> and <Background>.
- <Intents> consists of the questions or topics that can generate the <ActualResponse>.
- <Background> is a list of "Context" and "Information" pairs. For each pair, "Information" can be shared in the "Context" situation to generate the <ActualResponse>. And each pair can be used many times.

Guidelines for Attribution
- You need to decide whether the main points in the <ActualResponse> match the <StandardResponse>.
- The number of main points in the <ActualResponse> is not fixed. PointID is used to identify the main points in the <ActualResponse>.
- When assessing whether the main points in the <ActualResponse> originate from the <Transcript> or the <ContextInfo>, consider the following:
  1. If the main point has a similar or the same meaning as the <ContextInfo>. You should consider it as originating from the <ContextInfo>.
  2. If the main point explicitly repeats or closely relates to any point already mentioned in the <Transcript>. However, casual interactions such as greetings or small talk are permissible and not regarded as sourced from the <Transcript>."
- There are four situations for the origin of the main points in the <ActualResponse>:
  1. The main point in the <ActualResponse> can originate from the <ContextInfo> but is not present in the <Transcript>. You should append [PointID, 1, 0] to the AttributionList.
  2. The main point in the <ActualResponse> does not originate from the <ContextInfo> but originates from the <Transcript>. You should append [PointID, 0, 1] to the AttributionList.
  3. The main point in the <ActualResponse> can originate from both the <ContextInfo> and the <Transcript>. You should append [PointID, 1, 1] to the AttributionList.
  4. The main point in the <ActualResponse> does not originate from the <ContextInfo> and is not present in the <Transcript>. You should append [PointID, 0, 0] to the AttributionList.

Output Format
- The output MUST be in the JSON format.
- You MUST explain the process of attribution for every main point in the <ActualResponse>.
- Note that AttributionList should only contain the List of lists and should not contain any additional information or annotations.

CONTINUE ON THE NEXT PAGE

Table 20: Prompt used for the attribution of the generated response.

- The output MUST include the following fields:
  - Explanation: For every main point in the <ActualResponse>, explain the process of attribution. Especially, explain why the main point matches or does not match the <StandardResponse> and why it originates from the <Transcript> or the <ContextInfo>.
  - AttributionList: A list of lists, where each list contains PointID and the attribution for a main point in the <ActualResponse>.
  - PointsCount: The number of main points in the <ActualResponse>.

Example

```
- Example 1:
<Transcript>
PERSON13: Hi. Hello [PERSON6]. Hello [PERSON19]. Thanks for, uhm. (id=0)
PERSON6: Hi everyone. (id=1)
PERSON19: Hi. (id=2)
PERSON13: Yeah, great. Thanks for joining and, uh, yeah okay. So, yeah. Uh, I I see that
 people have written up ehm what they did. (id=3)
PERSON19: Hi [PERSON13], I can hear you. (id=4)
PERSON13: Yeah. [PROJECT3] deliverables. So, I'll try to provide the links-. Or those who
 of you, who are already working on the deliverables, please mention that. And yeah. Let's
 let's go quickly over what what have done. So [PERSON6] you are the first on the list. Ehm
 , ehm, so please briefly update what what you have been working on. And what what is the
 plan for the next week. (id=5)

<ContextInfo>
{
    "Intents": [
        "What [PERSON6] has been working on and the plan for the next week?"
    ],
    "Background": [
        {
            "Context": "Update on recent work and plans for the next week",
            "Information": "This week I had fewer tasks. I calculated the word error rate
 on Czech transcripts using three versions of Czech ASR created by [PERSON10]. There were
 significant mismatches between the golden transcript and its corresponding video. I
 conducted testing sessions with [PERSON11] and [PERSON18], which were not successful due
 to issues with segmenters from [ORGANIZATION1]. I also contributed to the [PROJECT3]
 deliverable for the punctuator and through caser."
        }
    ]
}

<StandardResponse>
["I calculated the word error rate on Czech transcripts"]

<ActualResponse>
["Calculated word error rate on Czech transcripts", "Conducted testing sessions with
 PERSON11 and PERSON18"]

<Evaluation>
{
    "Explanation": "1. Calculated word error rate on Czech transcripts. This point matches
  the standard response. "I calculated the word error rate on Czech transcripts" is present
  in the ContextInfo. Therefore, the attribution is [1, 1, 0]. 2. Conducted testing
 sessions with PERSON11 and PERSON18. The point does not match the standard response. "I
 conducted testing sessions with [PERSON11] and [PERSON18]" is present in the
 BackgroundKnowledge. Therefore, the attribution is [2, 1, 0].",
    "AttributionList": [[1, 1, 0], [2, 1, 0]],
    "PointsCount": 2
}
```

Table 21: Prompt used for the attribution of the generated response (continued).

- Your task is to update the summary of the utterances of {participant} in a meeting transcript.
- You are provided with a <Transcript Snippet> that contains a portion of the meeting transcript.
- You are also provided with <Previous Summary> which contains the summary of utterances for {participant} in other parts of the meeting.
- You need to update the <Previous Summary> based on the utterances of the {participant} in the <Transcript Snippet>.

On the provided <Transcript Snippet>
- Transcripts are the collection of utterances from the meeting participants.
- The transcript data is deidentified. Speakers and other named entities are not identified by names, but rather by IDs in the format ENTITYNUMBER (e.g. PERSON1 or PROJECT3) or just ENTITY (e.g. PATH).
- Speaker IDs at the beginning of transcript lines are enclosed in round brackets, all other deidentified entities in square brackets.
- Each utterence ends with "(id=x)", which is the utterance id, an increasing number from 0 to indicate the serial number of utterance in the whole meeting transcript.
- The provided transcript snippet maybe not start from the beginning of the meeting.
- Example utterances:
  (PERSON1) Hello everyone, I'm glad to see you all here today. (id=0)
  (PERSON2) Hi, I'm excited to be here. (id=1)
  (PERSON3) I'm looking forward to the discussion. [PERSON1] mentioned that the project is going well. (id=2)

On the <Previous Summary>
- The <Previous Summary> is a structured summary of the utterances of {participant} in the meeting transcript.
- The <Previous Summary> contains two parts, "wanted information" and "provided information".
  - "wanted information" is a list of questions made by the {participant}.
  - "provided information" is the information provided by the participant to others. It is a list of Context and Information pairs, where the "Context" is the context in which where the {participant} provides the "Information".

Instructions on updating the <Previous Summary>
- Identify the utterances of {participant} in the <Transcript Snippet>.
- If {participant} does not speak in the <Transcript Snippet>, do NOT update the <Previous Summary>.
- Focus on only the informative utterances and ignore the greetings, appreciation, simple acknowledge and other chit chat.
- Extract the "wanted information" and "provided information" from the <Transcript Snippet>.
- You should try to use original utterances as much as possible after removing noise words and polishing them for better readability.
- The second or third personal pronoun (you, he, she, they) in the utterances should be properly replaced with the corresponding participant's ID to avoid ambiguity.
- Use the extracted information to update the <Previous Summary>.
- You can modify the existing "wanted information" and "provided information" or add new information, but do not remove any existing information.
- You MUST NOT mix the information provided by {participant} and other participants while updating the <Previous Summary>.
- You MUST NOT miss any important information provided by {participant} in the <Transcript Snippet>.

Requirement on the output format
- You MUST explain your thoughts and steps of updating the <Previous Summary> before providing the updated summary.
- Output must be in Json format with the "Thoughts" and "Updated Summary" as the key.
- The "Thoughts" is your thoughts and steps of updating the <Previous Summary>.
- The "Updated Summary" contains the updated summary of the utterances for {participant}.

CONTINUE ON THE NEXT PAGE

Table 22: Prompt used for extracting context information.

```
 ## Example 1
 - Here is an example of updating the utterances summary for PERSON2. You can refer to
this example for better understanding.
 - Suppose the transcript snippet contains the following utterances:
     (PERSON3) Good moring. (id=2)
     (PERSON1) Let's get started with today's meeting on the recent progress of our
software development project. We'll go through updates from each team and discuss any
blockers or issues. [PERSON2], could you start with the development updates?". (id=3)
     (PERSON2) Sure, [PERSON1]. We've made significant progress this sprint. We completed
 the implementation of the new authentication module and integrated it with our existing
systems. (id=4)
     (PERSON1) That's great to hear, [PERSON2]. How about the feature for real-time
notifications? Is it on track? (id=5)
     (PERSON2) Yes, it is. We're about 75% done with it. The core functionality is in
place, and we are now working on optimizing the delivery speed and ensuring it works
seamlessly across different devices. (id=6)
 - Suppose the previous summary of PERSON2 contains the following information:
     {{
       "wanted information": [],
       "provided information": []
     }}
 - The thoughts and updated summary will be:
     {{
       "Thoughts":"In the transcript snippet, PERSON2 responds to PERSON1's questions
about the development updates and the progress of the feature for real-time notifications.
 This information can be added to the "provided information" for PERSON2."
       "Updated Summary":
         {{
           "wanted information": [],
           "provided information": [
             {{
               "Context": "Respond to other participant's question about the development
updates",
               "Information": "We've made significant progress this sprint. We completed
the implementation of the new authentication module and integrated it with our existing
systems."
             }},
             {{
               "Context": "Respond to other participant's question about the progress of
the feature for real-time notifications",
               "Information": "We're about 75% done with it. The core functionality is in
 place, and we are now working on optimizing the delivery speed and ensuring it works
seamlessly across different devices."
             }}
           ]
         }}
     }}

 ## Example 2
 - Here is another example of updating the utterances summary for PERSON2. You can refer
to this example for better understanding.
 - Suppose the transcript snippet contains the following utterances:
     (PERSON3) Good moring. (id=2)
     (PERSON1) Let's get started with today's meeting on the recent progress of our
software development project. We'll go through updates from each team and discuss any
blockers or issues. [PERSON2], could you start with the development updates?". (id=3)
     (PERSON2) Sure, [PERSON1]. We've made significant progress this sprint. We completed
 the implementation of the new authentication module and integrated it with our existing
systems. (id=4)
```

CONTINUE ON THE NEXT PAGE

Table 23: Prompt used for extracting context information (continued).

(PERSON1) That's great to hear, [PERSON2]. How about the feature for real-time notifications? Is it on track? (id=5)
(PERSON2) Yes, it is. We're about 75% done with it. The core functionality is in place, and we are now working on optimizing the delivery speed and ensuring it works seamlessly across different devices. (id=6)
(PERSON3) [PERSON2], have you had a chance to address the bug I reported last week related to the authentication module? (id=7)
(PERSON2) Yes, [PERSON3]. We identified the root cause of the bug, and it's been fixed. It was due to a conflict with a third-party library we were using. (id=8)
(PERSON3) That's good to hear. Thank you, [PERSON2]. (id=9)
(PERSON1) [PERSON2], for the next step of the project, I'd like you first complete the real-time notifications feature, and then focus on the chatbot development. (id=10)
(PERSON2) Understood, I will do that. (id=11)
(PERSON2) By the way, what's the timeline of our project? (id=12)
(PERSON1) We are aiming to finish the project by the end of August. (id=13)
(PERSON2) Ok, I know. (id=14)
(PERSON1) Let's move to the next topic. [PERSON3], could you provide an update on the testing progress? (id=15)
(PERSON3) Sure. Certainly. We've conducted tests on the new authentication module, and everything looks good so far. (id=16)
(PERSON1) Mhm. (id=17)
(PERSON3) We are now preparing for the testing of the real-time notifications feature. (id=18)
(PERSON2) In our development process, we accumulated some test cases which may help you. (id=19)
(PERSON3) That's helpful, thank you. (id=20)

- Suppose the previous summary of PERSON2 contains the following information:

```
{{
  "wanted information": [],
  "provided information": [
    {{
      "Context": "Respond to other participant's question about the development updates",
      "Information": "We've made significant progress this sprint. We completed the implementation of the new authentication module and integrated it with our existing systems."
    }},
    {{
      "Context": "Respond to other participant's question about the progress of the feature for real-time notifications",
      "Information": "We're about 75% done with it. The core functionality is in place, and we are now working on optimizing the delivery speed and ensuring it works seamlessly across different devices."
    }}
  ]
}}
```

- The thoughts and updated summary will be:

```
{{
  "Thoughts":"In the transcript snippet, the dicussion between PERSON1 and PERSON2 about the progress of the development and the feature for real-time notifications are already included in the previous summary. PERSON2 responds to PERSON3's question about the bug in the authentication module, which can be added to the "provided information" for PERSON2. PERSON2 asks about the timeline of the project, which can be added to the "wanted information" for PERSON2. PERSON2 also comments on PERSON3's statement about the testing progress, offering to provide some test cases, which can be added to the "provided information" for PERSON2."
  "Updated Summary":
    {{
      "wanted information": [
        "What's the timeline of the project?"
      ],
      "provided information": [
        {{
          "Context": "Respond to other participant's question about the bug in the authentication module",
```

CONTINUE ON THE NEXT PAGE

Table 24: Prompt used for extracting context information (continued).

```
              "Information": "We identified the root cause of the bug reported by [PERSON3], and it's been fixed.
It was due to a conflict with a third-party library we were using."
            }},
            {{
              "Context": "Comment on other participant's statement about the testing progress",
              "Information": "In our development process, we accumulated some test cases which are helpful to
testing."
            }}
          ]
        }}
    }}

## Example 3
- Here is an example of updating the utterances summary for PERSON6. You can refer to this example for better
understanding.
- Suppose the transcript snippet contains the following utterances:
    (PERSON13) Hi.
    Hello [PERSON6].
    Hello [PERSON19].
    Thanks for, uhm. (id=0)
    (PERSON6) Hi everyone. (id=1)
    (PERSON19) Hi. (id=2)
    (PERSON13) Yeah, great.
    Thanks for joining and, uh, yeah okay.
    So, yeah.
    Uh, I I see that people have written up ehm what they did. (id=3)
    (PERSON19) Hi [PERSON13], I can hear you. (id=4)
    (PERSON13) Yep, that's great.
    Uh, and also you were evaluating-.
    Yes, so that's that's re re record.
    What you did.
    So what I have, uh, on my mind now is uh, uh, well, uh, preparations.
    So, uh, [PERSON13], uh I am busy, uh, with the IW SLT, uh, write-up.
    Uh, that was the, uh, the wra last part that I did.
    Now busy with interviewing people people to uh to replace those who are em moving forward <laugh/> so to say.
    So there is number of colleagues on projects that I am supervising, uh, that who are going for studies abroad
and other things.
    Uh, so, uh, what I think we should focus on is the demo for Project Officer.
    Then we need to focus on the ladder climbing, uh, which is building uh, uh, [PROJECT3] test set plus, uh,
regularly, uh, testing on it.
    Ehm, and, ehm what else, uh, the deliverables.
    Yeah.
    [PROJECT3] deliverables.
    So, I'll try to provide the links-.
    Or those who of you, who are already working on the deliverables, please mention that.
    And yeah.
    Let's let's go quickly over what what have done.
    So [PERSON6] you are the first on the list.
    Ehm, ehm, so please briefly update what what you have been working on.
    And what what is the plan for the next week. (id=5)
    (PERSON6) <other_noise/>
    So, luckily.
    <laugh/>
    Not luckily but this week I had like quite less tasks to do.
    So first I calculated the word error rate on Czech transcripts using that three versions of, uh, Czech ASR
which [PERSON10] created.
    And so yesterday [PERSON10] told me that they were, uh, and the golden transcript and its corresponding video
there were there were huge huge mismatch.
    And I <unintelligible/> and he said to update me.
```

CONTINUE ON THE NEXT PAGE

Table 25: Prompt used for extracting context information (continued).

And then we conducted a few testing sessions with [PERSON11] and [PERSON18].
And they were not quite successful because the segmenters from, uh, uh, [ORGANIZATION1] they were still down and [PERSON12] today he is he is working on them. (id=6)
(PERSON13) Mhm. (id=7)
(PERSON6) And lastly yeah I think I did-.
Uh, it was the input in the [PROJECT3] deliverable of for the punctuator and through caser.
<unintelligible/> (id=8)
(PERSON13) Mhm, yeah. (id=9)
(PERSON6) And I don't have-.
I think that apart from the testing sessions to do this week so I am waiting for new tasks. (id=10)
(PERSON13) Yeah, so.
So the word error rate, there is also the English, uh, transcripts?
Ehm, and also we should have from [PERSON9] the German one, right?
So. (id=11)
(PERSON6) Yeah, yeah, yeah. (id=12)
(PERSON13) So, so I will make it to do-. (id=13)
(PERSON6) So I have updated the German transcripts in its corresponding path and like do we have the golden transcripts for the English videos? (id=14)
(PERSON13) Yes, that's the other part.
Because this is the consecutively translated videos.
So there is always the English speaker and then the Czech speaker who repeats the same content.
And [PERSON7] has split the video and while the English part should be more reliable, uh, the Czech part has been done simply by using the other ends.
So the Czech video has been cut using the English time stamps.
Like the end of English and the beginning of the next English segment.
Uh, so it's like like interleave the the the other way round.
So that's why I'm not surprised that the Czech video is, uh, imprecise in timing.
But still, I was not expecting it to be that bad.
So, uh, that is something that, yeah.
[PERSON10], can you maybe tell us more details about that? (id=15)
(PERSON10) Yeah, yeah.
Well, I just like listened to the audio and followed the talk transcript <other_noise/>
and it was completely off.
I think it is-.
There must be some miss-match because-. (id=16)
(PERSON13) Mhm. (id=17)
(PERSON10) Yeah, yeah.
The transcription is for the completely different audio than it's in the subdirectory. (id=18)
(PERSON6) Mhm. (id=19)
(PERSON13) Oh, so then someone must have like messed it up. (id=20)
(PERSON10) Yeah, yeah, I I may maybe it's just like uh.
Maybe the files are just switched between the subdirectories? (id=21)
(PERSON13) Mhm. (id=22)
(PERSON10) I I haven't checked but-.
Uh, yeah, there is some some serious mismatch there. (id=23)
(PERSON13) Yeah, so [PERSON10] can you coul could you do this check?
It should not be hard
Like try listening to all the files that are within this demo for [PERSON15], uh, and try to locate the correct file, the appropriate files.
But we should have, we should have the transcripts ready for all of those.
So we should be able to, uh, to evaluate it.
And also for the English ones we have the translations.
So for the English ones [PERSON6], uh, I would like you to evaluate not only the word error rate of the ASR.
But also the machine translation quality or at the SLT even.
Uh, with the translation quality into German and Czech.
Both are available. (id=24)
(PERSON6) Okay. (id=25)

CONTINUE ON THE NEXT PAGE

Table 26: Prompt used for extracting context information (continued).

```
     (PERSON13) We have these files ready. (id=26)
     (PERSON6) And so these, for like a German audio and English [PROJECT1]. (id=27)
     (PERSON13) English, uh, input for English sound.
     We have the golden English transcript, so you can check the ASR.
     And we also have the translation into Czech and into German.
     So you can also evaluate directly the translation quality, uh, of that. (id=28)
     (PERSON6) Okay, yeah. (id=29)
     (PERSON13) Yeah, so this is, this is an important, uh, task, uh, to do, uh wr also for German and English
audios.
     And another to do, uh, bleu or SLTF, uh, for, uh, German and Czech translations of, uh, English. (id=30)
- Suppose the previous summary of PERSON6 contains the following information:
    {{
      "wanted information": [],
      "provided information": []
    }}
- The thoughts and updated summary will be:
    {{
      "Thoughts":"PERSON6 responds to PERSON13's questions about the work done and the plan for the next week,
which can be added to the 'provided information' for PERSON6. PERSON6 responds to PERSON13's questions about the
word error rate for the English and German transcripts, which can be added to the 'provided information' for PERSON
6. PERSON6 also asks about the golden transcripts for the English videos, which can be added to the 'wanted
information' for PERSON6."
      "Updated Summary":
        {{
          "wanted information": [
            "Do we have the golden transcripts for the English videos?"
          ],
          "provided information": "wanted information": [
            "Do we have the golden transcripts for the English videos?"
          ],
          "provided information": [
            {{
              "Context": "Respond to other participant's question about the work done and the plan for the next
week",
              "Information": "This week I had quite less tasks to do. So first I calculated the word error rate on
Czech transcripts using three versions of Czech ASR created by [PERSON10]. And so yesterday [PERSON10] told me that
there were huge huge mismatch between the golden transcript and its corresponding video. And he said to update me.
And then we conducted a few testing sessions with [PERSON11] and [PERSON18]. And they were not quite successful
because the segmenters from [ORGANIZATION1] were still down and [PERSON12] today is working on them. And lastly, I
think I did the input in the [PROJECT3] deliverable for the punctuator and through caser. Apart from the testing
sessions to do this week so I am waiting for new tasks."
            }},
            {{
              "Context": "Respond to other participant's question about the word error rate for the English and
German transcripts",
              "Information": "I have updated the German transcripts in its corresponding path, and I don't konw if
we have the golden transcripts for the English videos."
            }}
          ]
        }}
    }}

## Note
- You MUST follow the instructions and examples provided.
- Similar to examples above, you should try to use original utterances as much as possible after removing noise
words and polishing them for better readability.
- You MUST NOT put the information provided by other participants or questions of other participants in the updated
 summary of {participant}.
- You MUST NOT miss any important information provided by {participant} in the <Transcript Snippet>.
- You MUST give the output in the required format.
```

Table 27: Prompt used for extracting context information (continued).

- You are an NLP expert agent tasked with generating an evaluation dataset to assess {person_id}'s response abilities in the categories of 'Chime In', 'Explicit Cue', 'Implicit Cue', based on the provided transcript.
- The conversation may involve multiple speakers, but your focus should solely be on {person_id}.
- Given the transcript contains lengthy utterances, selectively include only the highest quality exchanges in the evaluation dataset.
- Exclude chit-chat or unmeaningful utterances such as ["emm," "okay," "mhm," "uh-huh," "yeah," "oh," "right," "hmm"] from the evaluation dataset.
- Ensure that {person_id}'s responses are substantive and meaningful. Exclude responses from {person_id} that are simple acknowledgments or confirmations like "Yeah, yeah, definitely, yeah" or "Okay."
- The transcript data is deidentified. Speakers and other named entities are identified by IDs in the format ENTITYNUMBER (e.g., PERSON1, Speaker1 or PROJECT3) or simply as ENTITY (e.g., PATH).

Evaluation Type
- Chime In: When {person_id} spontaneously contributes to the conversation without being directly prompted.
- Usually Chime In is when {person_id} is not already engaged in the conversation but chimes in with a relevant comment or question.
- Explicit Cue: When {person_id}'s name is specifically mentioned by another Speaker in utterance with ID, then {person_id} responds to a clear and direct question or prompt towards {person_id}.
- Implicit Cue: When {person_id}'s name is not specifically mentioned by Speaker in utterance with ID, but {person_id} responds to a less direct prompt or follows up on information that suggests a response is needed.
- Usually Implicit Cue is when {person_id} is already engaged in the conversation and responds to a follow-up question from Speaker in utterance with ID.

Output Format
- Output must be in Json format. Here is the skeleton of the output format with explanation:
- Explanation: Your reason for selecting the evaluation instance and for categorizing it.
- Type: The category of the evaluation instance: 'Chime In', 'Explicit Cue', 'Implicit Cue'.
- Response IDs: The id or ids of the {person_id}'s response from the transcript. Include all Response IDs that are relevant to the evaluation instance. If there are multiple Response IDs, separate them with commas.
- ID: The utterance id that {person_id} responds to.
- Speaker: The speaker of the utterance with the ID.
- Maintain the chronological order of the transcript when generating the evaluation dataset. ID MUST precede Response IDs.
- Response IDs must be from {person_id}'s responses only and ID must be from the speaker's utterance that {person_id} responds to.
- Please return all suitable evaluation instances in the transcript. If you don't find any suitable instances for a category, you can leave the evaluation dataset empty. Please ensure you have thought through the transcript carefully before leaving the evaluation dataset empty.

Example: Below are two examples of transcript and the corresponding evaluation datasets generated to assess PERSON18's response abilities. You can refer to these examples when generating {person_id}'s evaluation dataset.

<Transcript>
"speaker": "Speaker 19", "content": "If you want, I can resend it again. (id=71)"
"speaker": "Speaker 13", "content": "Space tokeniser. <unintelligible/> Yes, so es essentially to answer your question in the email. We have to switch to and we have for the IWSLT. We have to switch to SacreBLEU and SacreBLEU does its own tokenisation before scoring. So there is no-. Let's let's simply forget NLTK bleu score. That is not reliable. (id=72)"
"speaker": "Speaker 18", "content": "Yes, but-. (id=73)"
"speaker": "Speaker 19", "content": "Yes, but we can combine our tokeniser with NLTK. (id=74)"
"speaker": "Speaker 13", "content": "Uf. Let's not do that. Let's just forget it. Let's let's just use SacreBLEU. (id=75)"
"speaker": "Speaker 19", "content": "Okay. (id=76)"
"speaker": "Speaker 18", "content": "I I have one comment about it. (id=77)"
"speaker": "Speaker 13", "content": "Mhm. Yeah. (id=78)"
"speaker": "Speaker 18", "content": "You sh should use tokeniser before enverse segmenter. (id=79)"
"speaker": "Speaker 13", "content": "Yes, that's it. Yeah. (id=80)"
"speaker": "Speaker 18", "content": "Because it's much better. Because it can rely on the on the dots and commas and question marks and so on. And you can you can check my script which does tokeniser, enverse segmenter and then de-tokeniser. And here is the path in the document.And-. (id=81)"
"speaker": "Speaker 13", "content": "Yeah. (id=82)"
"speaker": "Speaker 18", "content": "And it's it's using the Moses seg tokeniser and detokeniser. And it needs the the language tag as the first argument and then reference. (id=83)"
"speaker": "Speaker 13", "content": "Yeah. So [PERSON19], do you do you fo-? Do you understand? (id=84)"
"speaker": "Speaker 19", "content": "Yeah. (id=85)"
CONTINUE ON THE NEXT PAGE

Table 28: Prompt used for extracting test cases from meeting transcript.

```
<Evaluation Dataset>
[
    "Explanation": "PERSON18's utterance is informative and suitable for evaluation
dataset. PERSON18 spontaneously contributes to the conversation without being directly
prompted. This is a Chime In instead of Implicit Cue since [PERSON18] is not already
engaged in the conversation.",
    "Type": "Chime in",
    "Response ID": [77, 79, 81, 83]
    "ID": 76,
    "Speaker": "Speaker 19",
]
```

Example 2
<Transcript>
"speaker": "Speaker 13", "content": "Yes, so I need to review these and the internal deadline is in 6 days from now. Uh, so, hopefully I will get back to all of you. To each of you independently towards the end of the week if there is anything unclear. So that we meet the internal deadline on the 8th. Yeah, okay. Great. Uh. So [PERSON18], what what is your progress? (id=117)"
"speaker": "Speaker 18", "content": "Hmhm. Yes, and by reading the papers I found an interesting tool. (id=118)"
"speaker": "Speaker 13", "content": "Mhm. (id=119)"
"speaker": "Speaker 18", "content": "I found out that it's possible to measure out the speech rate by cutting the syllables. And there is one tool. One patent tool, which can detect the gender of speaker and the speech rate. (id=120)"
"speaker": "Speaker 13", "content": "Mhm. (id=121)"
"speaker": "Speaker 18", "content": "And some other characteristics. So we can try it and make a dashboard out of it. (id=122)"
"speaker": "Speaker 13", "content": "Mhm. That's that's useful thing. Uh, and later on we could even create models like-. If we if we recognise that someone is speaking too fast, we could use like a harsher summarisation. (id=123)"
"speaker": "Speaker 18", "content": "Yes. (id=124)"
"speaker": "Speaker 13", "content": "So we could be reducing reducing their speech mole with a different model. (id=125)"
"speaker": "Speaker 18", "content": "Yes, and there was also speech modes. Like whether it was angry or normal and so on. (id=126)"
"speaker": "Speaker 13", "content": "Mhm. (id=127)"
"speaker": "Speaker 18", "content": "But I have no idea how the tool works in practice. I I I I saw it only in Gi GitHub and I buy it. (id=128)"
"speaker": "Speaker 13", "content": "Yeah, uh. (id=129)"
"speaker": "Speaker 18", "content": "So we can try it and make a dashboard out of it. (id=130)"

```
<Evaluation Dataset>
[
  "Explanation": "PERSON18's utterance is informative and suitable for evaluation dataset.
 PERSON18 was directly prompted by Speaker 13. This is an Explicit Cue.",
  "Type": "Explicit Cue",
  "Response ID": [118, 120, 122, 124, 126, 128, 130]
  "ID": 117,
  "Speaker": "Speaker 13",
]
```

- Please refer to the examples provided to ensure consistency and coherence in generating the evaluation dataset. The evaluation dataset must be in json format.

Table 29: Prompt used for extracting test cases from meeting transcript (continued).

Dialogue Acts as a Lens on Human–LLM Interaction: Analyzing Conversational Norms in Model-Generated Responses

Arunima Maitra **Dorothea French** **Katharina von der Wense**

University of Colorado, Boulder
firstname.lastname@colorado.edu

Abstract

Large language models (LLMs) have revolutionized natural language generation across various applications. Although LLMs are highly capable in many domains, they sometimes produce responses that lack coherence or fail to align with conversational norms such as turn-taking, or providing relevant acknowledgments. Conversational LLMs are widely used, but evaluation often misses pragmatic aspects of dialogue. In this paper, we evaluate how LLM-generated dialogue compares to human conversation through the lens of dialogue acts, the functional building blocks of interaction. Using the Switchboard Dialogue Act (SwDA) corpus, we prompt two widely used open-source models, Llama 2 and Mistral, to generate responses under varying context lengths. We then automatically annotate the dialogue acts of both model and human responses with a BERT classifier and compare their distributions. Our experimental findings reveal that the distribution of dialogue acts generated by these models differs significantly from the distribution of dialogue acts in human conversation, indicating an area for improvement. Perplexity analysis further highlights that certain dialogue acts like 'Acknowledge (Backchannel)' are harder for models to predict. While preliminary, this study demonstrates the value of dialogue act analysis as a diagnostic tool for human–LLM interaction, highlighting both current limitations and directions for improvement.

1 Introduction

Large language model-based dialogue systems are hugely successful in open domain language generation tasks such as question-answering (Kann et al., 2022). Although these systems generally produce high-quality fluent dialogues and are able to hold conversations, their utterances sometimes fail to capture the nuances and emotions that are common in human–human interactions. Effective conversation depends on subtle patterns of dialogue acts – utterances that serve functions such as asking questions, signaling agreement, or providing acknowledgment. We aim to investigate the extent of interpersonal synergy exhibited by LLMs in comparison to human interactions. Synergy in terms of interactive and cooperative conversations refers to the way one agent responds to the other based on coordination and engagement between agents (Fusaroli and Tylén, 2016).

Dialogue acts are labels assigned to utterances that classify the intent of the speaker. In our study, we prompt the Llama 2 model developed by Meta (Touvron et al., 2023) and the Mistral-7B model developed by Mistral AI (Jiang et al., 2023), with context from the Switchboard Dialogue Acts (SwDA)[1] corpus of human telephonic conversations (Stolcke et al., 2000) to generate the next utterance. We then conduct dialogue act classification with a bert-base model (Raheja and Tetreault, 2019) and compare the dialogue acts of the LLM-generated responses to those of the gold-standard responses. Our study illustrates both the promise and the limits of current LLMs as conversational partners, and proposes dialogue act analysis as a human-centered diagnostic tool that complements surface-level metrics.

Dialogue acts, a basic unit of conversation and indication of quality and engagement (Deriu et al., 2021), allow us to measure the quality and type of utterances generated by models. By analyzing the specific types of dialogue act the Llama 2 and Mistral models struggle to generate, we gain a better understanding of the current limitations of LLMs. Dialogue acts have been especially studied in work with regard to classroom dialogue (Ganesh et al., 2021), thus models that better follow human conversation styles or use quality dialogue could mimic a teacher's discourse and guide each student individually based on their utterances, leading to more personalized feedback for each student (D'Mello

[1] https://catalog.ldc.upenn.edu/LDC97S62

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 317–325
November 9, 2025

and Graesser, 2013; Macina et al., 2023).

Badshah and Sajjad (2024) and Nadeau et al. (2024) suggest that Mistral outperforms Llama 2 in several aspects, including reduced hallucinations and enhanced engagement in back-and-forth conversations. Although Mistral tends to generate more engaging and informative dialogues, typical human conversations often diverge from this pattern. Human interactions frequently involve various cues and conventions, such as acknowledging others' opinions and providing affirmations. Our experimental results demonstrate that the Llama 2 model adapts its responses to more accurately reflect the nature of the ongoing conversation, aligning itself with the conversational style of the interlocutor.

In this work, we ask whether dialogue act analysis can serve as a diagnostic lens for evaluating conversational coherence in large language models. Specifically, we investigate whether systematic differences in dialogue act distributions between human and model-generated utterances can reveal where LLMs diverge from human conversational norms. Beyond NLP evaluation, this question connects to cognitive science perspectives on pragmatic competence, highlighting whether LLMs reproduce or miss key interactional strategies, and where dialogue act mismatches may help explain why chatbot interactions sometimes feel less natural.

2 Background and Related Work

Recent work by Shaikh et al. (2024) highlights that large language models often fail to establish common ground in conversation, using significantly fewer grounding acts such as clarifications, acknowledgments, and follow-up questions compared to humans. Their study introduces a taxonomy of grounding behaviors and demonstrates that instruction-tuned models systematically underuse these acts, particularly in high-stakes domains like emotional support and teaching. While their analysis focuses on how LLMs manage grounding, our work addresses a complementary question: how well LLMs reproduce the broader functional structure of human conversation as captured by dialogue acts. By analyzing utterance-level dialogue act distributions, we contribute an orthogonal yet critical view of conversational alignment, revealing that models overproduce questions and opinionated statements, but underproduce backchannels and agreement. Together, these findings indicate that LLMs diverge from human norms not only in their ability to construct shared understanding but also in their broader interactional strategies.

2.1 Language Models and Prompting

LLMs excel in various language tasks, including text generation, summarization, and translation. Yi et al. (2024) notes that Meta's LLaMA-2 is optimized for interactive conversations, adapting to user input, while OpenAI's GPT-4 (Achiam et al., 2023) is more versatile. We use the LLaMA-2 13B-chat and Mistral-7B-Instruct models to compare dialogue act alignment, as they represent different optimization strategies and to explore how varying training regimes influence dialogue structure. While the models are not the most recent, the focus was on analyzing dialogue act patterns rather than raw performance, and on employing widely used open-source models to ensure accessibility, transparency, and replicability.

LLaMA-2 13B-chat is a chain-of-thought optimized model fine-tuned for dialogue using supervised learning and Reinforcement Learning from Human Feedback (RLHF) with human evaluations for coherence, helpfulness, and safety (Touvron et al., 2023). It employs 'ghost attention' to preserve system instructions across turns, making it strongly suited for coherent multi-turn conversations.

In contrast, Mistral-7B-Instruct is a lightweight, instruction-tuned version of the base model, fine-tuned on publicly available conversational and instruction datasets (Jiang et al., 2023). While it retains architectural efficiencies like Grouped-Query Attention (GQA) and Sliding-Window Attention (SWA), Mistral-7B-Instruct also benefits from instruction-following refinement. However, it does not appear to use RLHF or chat-specific alignment via continued conversational feedback.

Prompting methods are crucial for enhancing LLM performance and tailoring responses to user specifications (Henrickson and Meroño-Peñuela, 2023). The *system prompt* in these models instructs the model on how to respond, giving users some control over generated dialogues. In-context learning, a prompt engineering technique, provides task demonstrations to guide LLMs (Wu et al., 2024; Rubin et al., 2021; Dong et al., 2022). It can be zero-shot, one-shot, or few-shot, depending on the number of input-output examples provided. This method is particularly effective for models with large context windows. We provide several lines of context and prompt the model to respond accordingly.

For example, Kosinski (2024) demonstrates that GPT-4 correctly completes 95% of a set of 40 traditional false-belief tests that are frequently used to assess Theory-of-Mind (ToM) in humans when given a large 32K context window size. By comparison, GPT-3 can only correctly solve 40% of the false-belief tasks because it is a smaller model (up to 1000 times smaller than GPT-4) with 2K context window size.

2.2 Dialogue Acts

Dialogue acts are the functional units of conversation, describing the communicative intent behind an utterance. Drawing from speech act theory (Searle et al., 1980) and conversation analysis, dialogue acts capture not only the literal meaning of an utterance but also the role it plays in interaction—for example, making a statement, asking a question, or providing feedback. In computational linguistics, the Switchboard Dialogue Act Corpus (SwDA) (Jurafsky, 1997) has become a widely used benchmark, defining a taxonomy of 44 dialogue act categories. A few representative examples are included in Table 1.

Recent research on dialogue act classification treats it either as a text classification problem, where each utterance is classified in isolation (Lee and Dernoncourt, 2016), or as a sequence labeling problem (Kumar et al., 2018; Tran et al., 2017). According to Raheja and Tetreault (2019), some of the most promising models for dialogue act tagging are usually some sort of combination of the following models: conditional random fields (CRFs; Zhou et al., 2015), recurrent neural networks (RNNs; Chen et al., 2018), or BERT (Ribeiro et al., 2019). We classify dialogue acts using the Context-Aware Self-Attention Dialogue Act Classifier[2], which outperforms state-of-the-art methods by 1.6% on SwDA, the primary dataset for this task (Raheja and Tetreault, 2019). This model uses frozen BERT-base embeddings as input and employs a context-aware self-attention mechanism over dialogue turns, followed by a softmax classifier trained on the SwDA corpus. This design enables it to capture inter-turn dependencies critical for dialogue act identification.

[2] https://github.com/macabdul9/CASA-Dialogue-Act-Classifier.git

3 Experimental Setup

3.1 Dataset and Prompt

We use 1000 SwDA transcripts for the experiments, which are records of 2,400 two-sided telephonic conversations between two strangers with about 70 provided conversation topics, where each utterance is tagged with relevant dialogue acts. Since the dataset is a transcription of phone recordings, we removed the noises that did not contribute to the actual conversation. We prompt the Llama 2 and Mistral models with the following:

System Prompt: *'You are a human, having a conversation with a stranger on telephone, about some topic from a predefined list. Given the context of the conversation, respond as best you can.'*

However, to assess the impact of different system prompt lengths and to finalize our choice of prompt, we experiment with both short and long variants:

Short System Prompt: *'You are a human having a conversation on telephone with another human you do not know, about some topic, from a given list. Given the context of the conversation, predict the next line as best you can.'*

Long System Prompt: *'You are having a conversation on telephone with someone you do not know. Given the context of the conversation, predict the next line as best you can. Respond with a single line. Your response should have dialogue act tags like- Statement-non-opinion, Acknowledge (Backchannel), Statement-opinion, Agree/Accept, Appreciation, Yes-No-Question, Non-verbal, Yes answers, Conventional-closing, Uninterpretable, Wh-Question, No answers, Response Acknowledgement like oh okay, Hedge, Declarative Yes-No-Question, Other, Backchannel in question form ,Quotation, Summarize/reformulate, Affirmative non-yes answers, Action-directive, Collaborative Completion, Repeat-phrase, Open-Question, Rhetorical-Questions, Hold before answer/agreement, and so on.'*

3.2 Methods

While the quality of generated responses seems to improve with in-context learning, the question remains how much prior knowledge is required for the dialogue systems to dynamically adjust their response strategies to align with human interactions (Brown et al., 2020).

Thus we initially conduct an experiment providing the LLMs the first 10 lines of utterances from the switchboard corpus as previous knowl-

Dialogue Act Tags	Example
Statement-non-opinion (sd)	Me, I'm in the legal department.
Acknowledge (Backchannel) (b)	Uh-huh.
Statement-opinion (sv)	I think it's great
Yes-No-Question (qy)	Do you have to have any special training?
Abandoned or Turn-Exit (%)	So, -

Table 1: Some Dialog Act Markup in Several Layers (DAMSL) tags

edge. The model is then prompted to generate the 11th line. For evaluation, both the model-generated responses and the corresponding human responses from the SwDA corpus were automatically annotated with dialogue act labels using the BERT classifier. We then compared these labels along two dimensions: (1) distributional differences, by calculating the relative frequency of each dialogue act across model and human outputs; and (2) instance-level agreement, by measuring how often the dialogue act assigned to the model's response matched that of the gold standard human response in the same conversational context. In addition to these automatic comparisons, we manually inspected a random sample of 100 model–human pairs to qualitatively assess whether the classifier produced sensible labels and whether mismatches reflected genuine conversational differences rather than classifier errors. This qualitative check suggested that while the classifier rarely mislabels uncommon dialogue acts, the overall trends are robust.

We follow the same set of experiments with 30 lines of utterances instead of 10, and again with no prior context (zero-shot learning) where we prompt the model to converse with the user without having context knowledge about the utterances or dialogues of the assistant, to gain a better perspective of the extent to which prompt engineering influences the ability of dialogue assistants to engage in human-like conversations. It is essential to gather diverse system-generated responses to perform further analysis on the trends of the generated dialogues focusing on patterns, consistencies, and areas of improvement or divergence, as compared to human dialogues.

Our final experiments use the best performing context length of 30 previous utterances.

Dialogue Act Tags	Human	Llama 2	Mistral
Statement-non-opinion (sd)	35.29	18.19	17.64
Statement-opinion (sv)	23.52	36.37	35.29
Yes-No-Question (qy)	18.47	27.24	29.41
Acknowledge/Backchannel (b)	11.76	13.63	11.76

Table 2: Percentages of Dialogue Act tags of the selected utterances from the SwDA dataset, Llama 2 and Mistral

4 Results and Analysis

Table 2[3] shows the distribution of the top 4[4] dialogue act labels for the original utterances from the SwDA dataset next to the distribution of dialogue act tags among Llama 2's and Mistral's generated responses after conditioning the model on 30 lines of previous conversational context.

We see that humans use non-opinion statements significantly more compared to Llama 2 and Mistral. Both the models generate more opinion statements and questions, compared to humans. Although the models exhibit similar performance, Llama 2 demonstrates a greater tendency to acknowledge the provided prior context. However, solely measuring the overall distribution of dialogue acts might not be the most efficient method for identifying whether a dialogue act has been altered in the generated utterance compared to its original classification in the dataset.

Next, we consider the dialogue act tag for the generated versus the original sentence and provide a normalized confusion matrices, Figure 1 and Figure 2[5], indicating how many kept the same tags in the models' classification outcomes.

We see both LLMs predominantly respond with 'statement-opinion', as corroborated by Table 2.

[3]This table shows the distribution of dialogue acts from the gold standard 'next line' dialogues available from the transcripts after 30 lines of utterances, not the entire SwDA.

[4]Top 4 dialogue act tags are shown as they account for more than 85% of the dataset.

[5]Please refer to Table 2 for the list of dialogue act abbreviations.

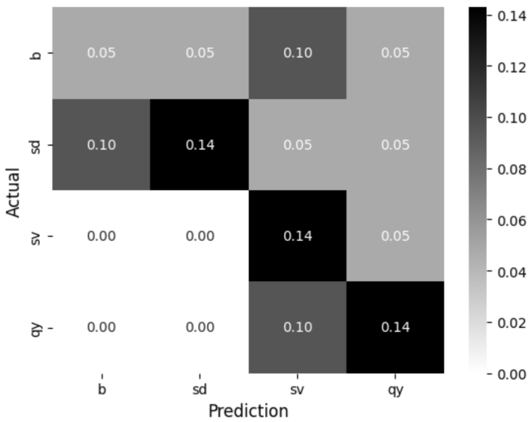


Figure 1: Confusion Matrix (Actual vs. Llama 2 13B Responses)

Llama 2 demonstrates a superior ability to align with the prompt and respond in a manner consistent with the given instructions, whereas Mistral encounters difficulties in following follow-up instructions related to a previous prompt. However, both models exhibit sub-optimal performance in accurately replicating outcomes compared to the ground truth data for our task. Further discussion on the models' accuracy is provided in the following sub sections.

4.1 System Prompt and Context Lengths

While running the experiments, we find that shorter and concise system prompts result in improved performance outcomes, whereas the utilization of broader prompts yields comparatively inferior results[6]. Longer system prompts, containing a significant amount of information, overwhelms the model and lead to incorrect associations. Whereas small and precise system prompts offer clearer guidance to the model, reducing ambiguity and potential confusion.

We additionally experiment with various context lengths. First, we provide 10 lines of conversation between the conversation participants from the transcripts and ask the model to predict the next utterance. This short context length causes a high level of ambiguity in the generated response irrespective of the topic of conversation:

```
Assistant: Yeah, I know, it's kind
           of surprising, right?
Assistant: Oh wow, that's surprising.
Assistant: Wow, that's something.
```

Expanding the preceding context to include 30 lines of previous utterances results in a notable reduction of 'surprise' exhibited by the models. Moreover, the models demonstrate enhanced proficiency in maintaining coherence and relevance throughout the conversation, akin to human conversational comprehension:

```
Assistant: Yeah, it was a pretty
           chaotic time, you know?
Assistant: Yes, it's about time we give
           equal importance to all health
           issues, regardless of who they
           affect.
```

Finally, we conduct an experiment devoid of any preceding user and assistant interactions, instead supplying only the one utterance and instructing the models to continue the conversation. In this scenario, the models unsurprisingly exhibit difficulty in following the conversation, as depicted in the dataset, often introducing novel topics and information to sustain the interaction. This observed behavior suggests a limitation in the model's capacity to adapt its conversational style without contextual cues, resembling the behavior commonly observed in open-domain dialogue systems. Additionally, we note a unique reaction of the system to a subset of transcripts, ones in which the conversations exhibit overt one-sidedness or lack engagement. These especially dry transcripts are characterized by an average utterance length of less than 6 words per utterance for one or both participants, such as the example taken from the dataset below:

Speaker A: Are you still there? Speaker B: Yes. Speaker A: Okay, Speaker B: it worked out fine. Speaker A: Okay.

Subsequent conditioning of the model with such dialogue context results in the generation of utterances aimed at concluding the conversation, rather than perpetuating dialogue that contributes minimally to its progression or substance, previously shown in Abbasiantaeb et al. (2024).

4.2 Perplexity

The perplexity of a large language model is a measure of its prediction effectiveness on a certain dataset. It measures how likely a model finds a sequence of words by calculating the exponentiation of the average negative log-likelihood of the predicted tokens. A lower perplexity indicates better performance, as the model's predictions are closer

[6]Section 3.1 provides examples of long and short system prompts.

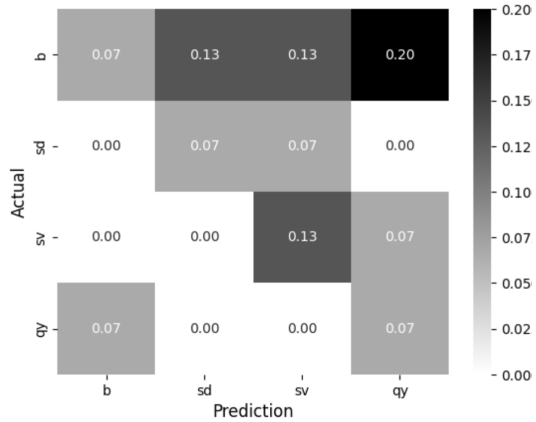


Figure 2: Confusion Matrix (Actual vs. Mistral 7B Responses)

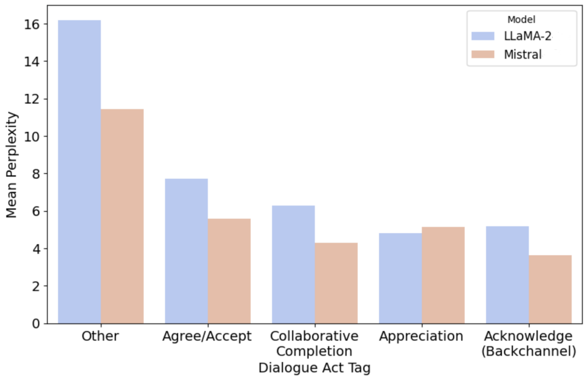


Figure 3: Top 5 Dialogue Acts with Highest Perplexity

to the actual sequence. We compute normalized perplexity (Roh et al., 2020) for both models on the gold response, using the first 30 lines of utterance from each conversation as context. Table 3 depicts the average normalized perplexity for the Llama 2 and Mistral models, where Mistral slightly outperforms Llama 2, indicating it produces more confident predictions.

Model	Average Perplexity
Llama 2	2.96
Mistral	2.13

Table 3: Perplexity Evaluation Summary

In order to study the models' responses to various types of dialogue acts, we compute the average perplexity for each tag, and sort the tags based on their perplexity scores (highest and lowest), shown in Figure 3 and Figure 4 respectively.

These figures effectively show what types of responses– types of dialogue acts, the model finds the most or least confusing.

The 'Other' dialogue act type causes the most confusion for both models, i.e., both models struggle with less predictable dialogue types. This category includes utterances that do not clearly fit into any other dialogue act categories, encompassing statements such as correct-misspeaking, sympathetic comments, and greetings. The top five dialogue acts that contribute most to model confusion collectively account for less than 28% of the Switchboard dataset, with the majority stemming from the 'Acknowledge (Backchannel)' dialogue act (19%)[7]. This aligns with our findings that large language models are more likely to generate opinionated statements rather than simple agreement or acknowledgment, as the latter contributes minimally to advancing the conversation. Mistral outperforms Llama 2 in terms of lower perplexity for most dialogue act types.

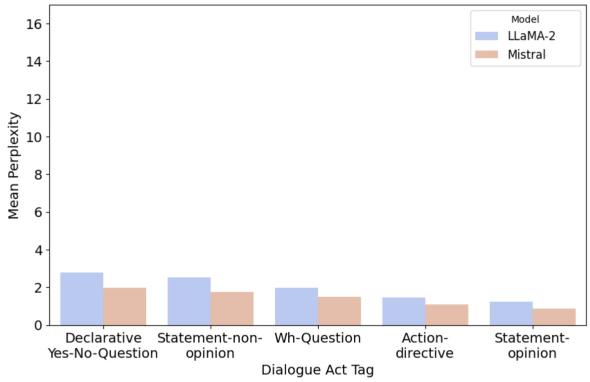


Figure 4: Bottom 5 Dialogue Acts with Lowest Perplexity

The dialogue act which causes the least confusion among both the models is 'Statement-opinion', as both models tend to generate 'Statement-opinion' utterances (Burton et al., 2024), also inferred from Table 1. The five dialogue acts that lead to the lowest model confusion collectively comprise 52.4% of the Switchboard dataset, with the largest share attributed to the 'Statement-non-opinion' dialogue act (36%).

4.3 Classification Report for Llama 2 and Mistral

Table 4 and Table 5 show the precision, recall and F1 score for the top 4 dialogue act categories for the Llama 2 13B-chat model and Mistral-7B-Instruct model respectively.

The models achieved varying levels of performance across dialogue act categories. The overall

[7] https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

DA Tags	Precision	Recall	F1
sd	0.75	0.43	0.55
sv	0.38	0.75	0.50
qy	0.50	0.60	0.55
b	0.33	0.20	0.25

Table 4: Precision, Recall and F1 for Llama 2

accuracy for the given categories is 0.48 and 0.33, respectively, for Llama 2 and Mistral, indicating significant room for improvement. In the context of human-like conversational carryover with prior knowledge, Llama 2 demonstrates a slight performance advantage over Mistral. Higher recall but lower precision in the Mistral model follows the previous result of lower perplexity but less aligned.

DA Tags	Precision	Recall	F1
sd	0.33	0.50	0.40
sv	0.40	0.67	0.50
qy	0.20	0.50	0.29
b	0.50	0.12	0.20

Table 5: Precision, Recall and F1 for Mistral 7B

5 Conclusions and Future Work

We analyze dialogues generated by Llama 2 and Mistral, using various levels of prompting and in-context learning, comparing them to the original human-human interactions from SwDA, utilizing dialogue acts to gauge similarity. In our research, we initially computed the percentages of the top four categories of dialogue act tags in both the original and corresponding LLM-predicted utterances. Our findings suggest that dialogue acts are not only a descriptive tool but also a potential predictor of when conversational systems fail to align with human norms. This analysis shows that in contrast to humans, who commonly use both opinionated and non-opinionated statements, language models exhibit a preference for generating opinion statements, potentially to add perceived value to the conversation. Additionally, these models tend to ask more questions, aiming to contribute more actively to the dialogue. Upon conducting further investigation using a confusion matrix, we discovered significant variations in the dialogue acts between the generated and original utterances, which were apparent in the differing proportions, as discussed previously. Llama 2's higher perplexity suggests that it might be more sensitive to context shifts and nuanced dialogue structures, resulting in more accurate classifications but higher uncertainty. Additionally, we found that context length significantly impacts response quality.

Future research could expand this work in several directions. One promising avenue is to evaluate newer models, to assess whether advances in training and alignment reduce the dialogue act discrepancies we observed. Another is to extend dialogue act analysis beyond distributional comparisons, incorporating metrics for conversational flow, user engagement, and appropriateness in interactive settings. Finally, integrating human-in-the-loop evaluations, where human participants interact with models and provide feedback on coherence and naturalness, would help connect dialogue act diagnostics more directly to real-world conversational quality.

Taken together, these directions highlight the potential of dialogue act analysis to bridge natural language processing and human–computer interaction, supporting the development of conversational systems that are not only fluent but also socially and pragmatically aligned with conversational norms.

6 Limitations

A key limitation of the evaluation is that it compares model-generated dialogue acts to those annotated in the Switchboard corpus, implicitly treating the human responses as a single "gold standard." In natural conversation, however, multiple dialogue acts could be appropriate in the same context, e.g., a turn could plausibly be an Acknowledgment, a Yes–No Question, or an Opinion statement depending on the speaker's intent. As a result, this comparison may underestimate the flexibility of LLMs or exaggerate deviations from human norms. We therefore frame our findings as diagnostic rather than definitive, using distributional patterns, such as the models' tendency to overproduce questions and opinions, to highlight systematic behavioral differences. Future work could incorporate human judgments, multiple reference responses, or metrics for conversational diversity and context-sensitive appropriateness, providing a more nuanced assessment of how well LLMs emulate human interaction.

Our study is limited to English conversations, since the Switchboard Dialogue Act corpus is available only in English. While this choice ensures comparability with prior work and leverages a widely studied benchmark, it also restricts the generalizability of our findings. Dialogue acts and conversational norms vary across languages and cultures; for example, the use of backchannels, politeness markers, or indirect questions can differ substantially. Future work should extend this analysis to multilingual corpora, which would allow us to evaluate whether the dialogue act patterns we identify are specific to English or reflect broader conversational tendencies in LLMs.

The SwDA utilized in our research provides a comprehensive array of dialogue acts; however, it lacks representation of certain emotional expressions commonly employed by humans. While dialogue act tags serve as valuable markers for categorizing communicative intents in dialogue, in the particular dataset, they inherently lack the capacity to encompass certain nuanced aspects of human expression, such as sarcasm. Consequently, the absence of explicit consideration for such emotional nuances within dialogue act frameworks represents a notable limitation, potentially leading to incomplete or inaccurate characterizations of human utterances in conversational AI systems.

Sometimes, there arise instances where the dialogue act labels assigned to the generated utterances align with those found in the reference data, yet substantial differences exist in the semantic content or pragmatic context of the dialogues. Such divergences underscore the inherent complexity of assessing dialogue quality solely through dialogue act matching, as they indicate potential limitations in capturing the richness and subtleties of human conversation beyond surface-level categorizations.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sher Badshah and Hassan Sajjad. 2024. Quantifying the capabilities of llms across scale and precision. *arXiv preprint arXiv:2405.03146*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, and 1 others. 2024. How large language models can reshape collective intelligence. *Nature human behaviour*, 8(9):1643–1655.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 225–234.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Sidney D'Mello and Art Graesser. 2013. Design of dialog-based intelligent tutoring systems to simulate human-to-human tutoring. In *Where humans meet machines: Innovative solutions for knotty natural-language problems*, pages 233–269. Springer.

Riccardo Fusaroli and Kristian Tylén. 2016. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, 40(1):145–171.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2021. What would a teacher do? predicting future talk moves. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4739–4751.

Leah Henrickson and Albert Meroño-Peñuela. 2023. Prompting meaning: a hermeneutic approach to optimising prompt engineering with chatgpt. *AI & SOCIETY*, pages 1–16.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *www. dcs. shef. ac. uk/nlp/amities/files/bib/ics-tr-97-02. pdf*.

Katharina Kann, Abteen Ebrahimi, Joewie Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain dialogue generation: What we can do, cannot do, and should do next. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 148–165, Dublin, Ireland. Association for Computational Linguistics.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Preprint*, arXiv:2302.02083.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.

David Nadeau, Mike Kroutikov, Karen McNeil, and Simon Baribeau. 2024. Benchmarking llama2, mistral, gemma and gpt for factuality, toxicity, bias and propensity for hallucinations. *Preprint*, arXiv:2404.09785.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594*.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019. Deep dialog act recognition using multiple token, segment, and context information representations. *Journal of Artificial Intelligence Research*, 66:861–899.

Jihyeon Roh, Sang-Hoon Oh, and Soo-Young Lee. 2020. Unigram-normalized perplexity as a language model performance measure with different vocabulary sizes. *CoRR*, abs/2011.13220.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

John R Searle, Ferenc Kiefer, Manfred Bierwisch, and 1 others. 1980. *Speech act theory and pragmatics*, volume 10. Springer.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. *Preprint*, arXiv:2311.09144.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *CoRR*, cs.CL/0006023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 428–437.

Qinyuan Wu, Mohammad Aflah Khan, Soumi Das, Vedant Nanda, Bishwamittra Ghosh, Camila Kolling, Till Speicher, Laurent Bindschaedler, Krishna P Gummadi, and Evimaria Terzi. 2024. Towards reliable latent knowledge estimation in llms: In-context learning vs. prompting based factual knowledge extraction. *arXiv preprint arXiv:2404.12957*.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Yucan Zhou, Qinghua Hu, Jie Liu, and Yuan Jia. 2015. Combining heterogeneous deep neural networks with conditional random fields for chinese dialogue act recognition. *Neurocomputing*, 168:408–417.

Author Index