HCI+NLP 2024

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

Proceedings of the Workshop

June 21, 2024

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

Sponsored by

Google

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 979-8-89176-111-7

Introduction

Welcome to the Third 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 two previous editions of this HCI+NLP workshop at EACL 2021 and NAACL 2022 and bring the third edition to NAACL 2024. In this workshop we present eleven papers, of which nine are archival papers, and two 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: Lydia Chilton and Sherry Tongshuang Wu.

We hope you find this workshop enjoyable! — Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dev, Michael Madaio, Ani Nenkova, Diyi Yang, and Ziang Xiao

Organizing Committee

Organizing Committee

Su Lin Blodgett, Microsoft Research, Canada
Amanda Cercas Curry, Bocconi University, Italy
Sunipa Dev, Google Research, USA
Michael Madaio, Google Research, USA
Ani Nenkova, Adobe Research, USA
Diyi Yang, Stanford University, USA
Ziang Xiao, Johns Hopkins University, USA

Program Committee

Program Committee

Özge Alacam, Bielefeld University
Maurice Jakesch, Cornell University
Mahmood Jasim, University of Massachusetts Amherst
Geza Kovacs, Google
Nina Markl, University of Essex
Marianna J. Martindale, University of Maryland, College Park and U.S. DoD
David Mimno, Cornell University
Tatiana Passali, CERTH/ITI and Aristotle University of Thessaloniki
Indira Sen, Rheinisch Westfälische Technische Hochschule Aachen
Qinlan Shen, Oracle
Zijie J. Wang, Georgia Institute of Technology
Austin P Wright, Georgia Institute of Technology

Invited Speakers

Lydia B. Chilton, Columbia University, USA
Sherry Tongshuang Wu, Carnegie Mellon University, USA

Table of Contents

Examining Prosody in Spoken Navigation Instructions for People with Disabilities

Cathy Jiao[♠] **Aaron Steinfeld**[◇] **Maxine Eskenazi**[♠]
[♠]Language Technologies Institute, Carnegie Mellon University
[◇]Robotics Institute, Carnegie Mellon University
{cljiao, steinfeld, max}cs.cmu.edu

Abstract

The introduction of conversational systems have made synthesized speech technologies common tools for daily activities. However, not all synthetic speech systems are designed with the needs of people with disabilities in mind. This paper describes a study in which 198 people – 80 participants with self-reported disabilities and 118 participants without – were recruited to listen to navigation instructions from a spoken dialogue system with different prosodic features. Results showed that slowing down speech rate aids in participants' number recall, but not in noun recall. From our results, we provide suggestions for developers for building accessible synthetic speech systems.

1 Introduction

The introduction of conversational systems such as Apple's Siri and Amazon's Alexa have made synthesized speech technologies common tools for daily activities. However, people with disabilities still struggle to interact with synthetic speech systems (Vieira et al., 2022). On the NLP side, current work in accessibility focuses on enhancing system or model functionality. Examples include finding appropriate data to train deep learning models that can be tailored to people with disabilities (Yaneva et al., 2017), or examining negative language model biases against disabilities (Venkit et al., 2022). Yet, user evaluation of these models deployed in systems are limited. On the other hand, past work in HCI and speech systems have noted the importance of using appropriate prosody to achieve better user understanding and recall of the system outputs. (Duffy and Pisoni, 1992; Mirenda and Beukelman, 1987; Paris et al., 2000; Wester et al., 2016; Rodero, 2017). To tie together perspectives from both NLP and HCI, we conducted a study in which user recall of instructions from an existing spoken dialogue system was evaluated in order to (1) determine what speech features are most beneficial for user recall of information, and (2) decide on future features to implement in the system. We recruited 198 people (with and without disabilities) and asked them to recall information under speech conditions which had either (1) *slowed down* speech or (2) *pauses before keywords* in the instruction. Furthermore, we grounded our study to the realm of understanding navigation instructions, which is a challenging setting because it requires users to remember exact *numeric entities* (e.g., departure times, building numbers) and *noun entities* (e.g., unfamiliar street names) in order to navigate effectively.

Our results showed that across all participants (with and without disabilities), slowing down speech rate aided in recalling numeric entities, even as the number of numeric entities in an instruction increased, but was less effective for noun entity recall. Furthermore, in speech conditions where breaks were inserted before numeric and noun entities, participants in general had lower performance in information recall. Our findings suggest that developers may be able to make adjustments that promote overall accessibly. In addition, appropriate cues for different types of information (i.e., numeric vs. noun) should be considered.

2 Related Work

Appropriate prosody in synthetic speech systems is important since synthetic speech often lacks natural cues and pacing found in human voices. Previous works have shown that inappropriate or absent prosodic cues led to large performance gaps in recall tasks from natural and synthetic speech (Paris et al., 2000), and that human voices were preferred

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 1–12
June 21, 2024

over synthetic voices as task complexity increases (Rodero, 2017). On the other hand, appropriate prosodic cues can improve peoples' abilities to disambiguate information – for example, distinguishing between similar spoken mathematical expressions (Gellenbeck and Stefik, 2009). Furthermore, adding appropriate prosodic cues can decrease performance gaps between different age groups in understanding speech (Stine and Wingfield, 1987; Wingfield et al., 1992; Langner and Black, 2005; Wolters et al., 2007; Roring et al., 2007). Some existing NLP systems have taken this information into account. For instance, CMU GetGoing (Mehri et al., 2019) is a trip-planning dialogue that introduced attention-grabbing prefixes and allowed "barge-in" to provide easier interaction for seniors. However, many modern spoken dialogue systems still fail to capture nuances of interaction for people with disabilities (Vieira et al., 2022). While previous work such as Koul (2003) showed that synthetic speech comprehension among people with disabilities were influenced by the complexity of task and acoustic-phonetic features, our study differs from previous work in that we grounded our study in a trip-planning task. This task is challenging to users (i.e., recalling unfamiliar street names), yet is important for independent daily travel, especially for people who are dependent on public transit.

3 Overview

To analyze the effects of speech prosody on information retention, we recruited participants with 108 and without self-reported disabilities to listen to 109 audio clips containing navigation instructions. The outline of our paper is as follows: in Section 4 we describe the audio clip collection process for the study. In Section 5 we explain the study design and participant recruitment. Finally, in Sections 6, 7 and 8 we provide study results and further discussion.

4 Audio Clips Collection

Audio clips were collected from GetGoing, a spoken dialog system that provides step-by-step navigation instructions to aid senior users in trip planning. GetGoing was previously deployed in Southwestern Pennsylvania and could be accessed by users through telephone. It follows a traditional pipeline dialog system architecture with natural language understanding and speech modules. Notably, the system uses Google Maps API[1] to curate sets of directions between two locations. Furthermore, the system uses Vonage's text-to-speech API [2] to output synthesized speech to the user.

The audio clips collection process involved three stages. In the *first stage*, GetGoing was used to generate navigation instructions using random start and end destination points around Southwestern Pennsylvania. In the *second stage*, a subset of the generated instructions was selected according to two information parameters defined in Section 4.1 to ensure our instruction set was well-balanced. This resulted in 48 unique instructions. In the *third stage*, speech conditions (described in Section 4.2) were created by either slowing down the speech or inserting breaks before keywords using the Vonage API. We recorded the speech output of GetGoing for each of the 48 instructions under each condition, and ended up with a total of 192 audio clips.

4.1 Navigation Instruction Parameters

To ensure the curation of a balanced set of navigation instructions from GetGoing, we defined each instruction by two information parameters. The first parameter is the *number of nouns entities* in an instruction. Noun entities included street and building names (e.g., *"Frew St"*, *"UPMC Presby"*). The second parameter is the *number of numeric entities* in an instruction. Numeric entities included any word that contained a number (e.g., *"9:30pm"*, *"7th Street"*). Certain items in an instruction were counted as both a noun entity and numeric entity. For instance, in the instruction *"Take the bus to Liberty avenue and 7th Street."*, the name *"7th Street"* was counted as a noun entity and numeric entity. We used noun and numeric entities as information parameters for two reasons. First, when navigating it is important to remember names, streets, and times. Second, our instructions are not multi-step and do not contain any other major pieces of information.

Next, we defined eight unique parameter combinations, or groups, which are listed in Table 1. Each parameter group contained 1-2 noun enti-

[1] `https://developers.google.com/maps/documentation/`

[2] `https://www.vonage.com/developer-center/`

Group	Numeric Entities	Noun Entities	Words *mean*	*sd*
1	3	2	17.3	3.6
2	2	2	18.7	3.6
3	1	2	19.2	3.9
4	0	2	14.2	3.6
5	3	1	16.2	3.4
6	2	1	15.8	3.0
7	1	1	11.8	2.5
8	0	1	11.3	2.4

Table 1: Navigation instruction parameter groups. Each group contains a set number of noun and numeric entities per instruction. The average number of words per instruction in each group is listed on the right column.

ties and 0-3 numeric entities. For each parameter group, six instructions were randomly selected from the inital set of instructions collected from GetGoing, resulting 48 unique instructions. Table 7 in the Appendix lists all instructions and their parameter values.

4.2 Audio Clip Conditions

Using Vonage's text-to-speech API, we created four speech conditions:

1. **Default**: The default Vonage API voice.
2. **Default-slow**: The Vonage API voice with the speech rate set to "slow".
3. **Break-short**: The default Vonage API voice with a 5ms break before every noun and numeric entity.
4. **Break-long**: The default Vonage API voice with a 15ms break before every noun and numeric entity.

We selected these conditions since previous work has shown that inserting pauses can aid in speech understanding, especially for seniors (Langner and Black, 2005; Wolters et al., 2007). While other prosodic cues and its effects on information retention can be explored, we decided to focus on these four conditions for the scope of our experiments.

We applied each of the four conditions to each of the 48 instructions to create 192 audio clips. Table 2 shows the duration of clips according to each condition. Clips in the *Default* condition had the

Condition	Clip Duration (s) *mean*	*sd*
Default	7.6	1.9
Default-slow	8.8	2.0
Break-short	7.8	1.9
Break-long	8.0	2.0

Table 2: Average audio clip lengths within the four speech conditions.

shortest average duration, while clips in *Default-slow* had the longest average duration.

5 Study Design

In this section, we describe the study design, procedure, and participants. All participants were recruited on Prolific.co. From there, they were directed to a Qualtrics page to consent to the study. After consenting, participants were redirected to a website to do the tasks. An example of the task interface is provided in Figure 4 in the Appendix. Each participant listened to 24 audio clips and entered what they heard in a text box on the website. Participants listened to six clips from each of the four conditions. Each of these six clips were randomly sampled from instructions from different parameter groups. All clips were presented to the participant in random order, and the website prevented participants from listening to the clip more than once in order to test recall. Once the participants finished the tasks, they were redirected back the the Qualtrics page where they answered a short questionnaire, which is included in Table 6 in the Appendix.

5.1 Participants

We recruited 198 adult participants from the U.S. through Prolific.co (two additional participants did not complete the study). The participants ages ranged from 18 to 78 years (*mean = 37.3, sd = 14.3*), and the age breakdown is shown Table 4. The participants reported their genders as follows: 105 participants self-identified as female and 82 as male, 10 as other genders, and 1 did not disclose their gender. The majority of participants reported English as their native language *(n = 192)*. Furthermore, the majority of participants reported that they used a computer daily on a scale from "1 (Never)" to "7 (Often - daily)" *(mean = 6.85, sd =*

Self-Reported Disabilities	Participants (n=198)
cognitive disability	12
communicative disability	3
dexterity disability	9
hearing impairment	13
mental disability	39
mobility disability	17
vision impairment	23
other	8
none	118

Table 3: Participants' self-reported disabilities. Some participants have multiple self-reported disabilities

Age Group	Participants (n=198)
18-24	35
25-34	70
35-44	41
45-54	21
55-64	19
65+	12

Table 4: Participants' ages.

0.51, min = 4, max = 7) and occasionally used a voice assistant *(mean = 3.58, sd = 2.09, min = 1, max =7)*.

We used filters from Prolific.co to recruit participants with and without disabilities. We first recruited 98 participants without any selection criteria. We then used five different Prolific filters to recruit people with varying disabilities. More specifically, we recruited participants who indicated they have (1) *vision*, (2) *hearing*, (3) *mobility*, (4) *chronic conditions*, and (5) *cognitive* disabilities on Prolific. Twenty participants were recruited using each filter. In addition, we asked participants to self-report any disabilities during the post-study questionnaire (shown in Table 6 in the Appendix) to handle discrepancies and ambiguity from the Prolific filter categories. The breakdown of participant self-reported disabilities is reported in Table 3.

6 Results

We grouped participants in two categories: those who had self-reported disabilities and those who did not, and analyzed recall of noun and numeric entities across the four speech conditions. While it is possible to examine the subgroups of self-reported disabilities individually (e.g., all participants with a mobility disability), we avoided this since the subgroups are imbalanced with respect to the number of participants. In addition many participants reported more than one disability.

6.1 Noun Retention Accuracy

Annotations for nouns were done manually, and each noun was assigned as "correct" or "incorrect" by phonetic similarity since participants transcribed text from audio. For example, *"Knight St"* was considered to be the same as *"Nite Street"*. Figure 1 shows the results.

In clips with *one noun*, little difference in noun accuracy (averaged across all conditions) was observed among people with self-reported disabilities (*mean = 94.27*) and those without (*mean = 90.61*. In clips with *two nouns*, the noun accuracy dropped in both groups, but the difference still remained small between people with self-reported disabilities (*mean = 46.35*) and those without (*mean = 43.08*). Large changes were not observed in noun accuracy across different conditions among both groups.

6.2 Number Retention Accuracy

In clips with *one number*, little difference in number accuracy (averaged across all conditions) was observed between people with self-reported disabilities (*mean = 79.38*) and those without (*mean = 78.25*). In clips with *two numbers*, the number accuracy dropped in both groups. However, there was greater decrease in performance for people with self-reported disabilities (*mean = 36.46*) than those without (*mean = 43.08*). In clips with *three numbers*, little difference was observed in the number accuracy between people with self-reported disabilities (*mean = 34.17*) and those without (*mean = 35.03*).

Unlike noun accuracy, differences in number accuracy were noticeable across conditions. For instructions with ≥ 2 numbers, participants (those with and without self-reported disabilities) in general performed the highest in number recall the *default-slow*. Meanwhile participants performed the worst in *break-short*, followed by *break-long*.

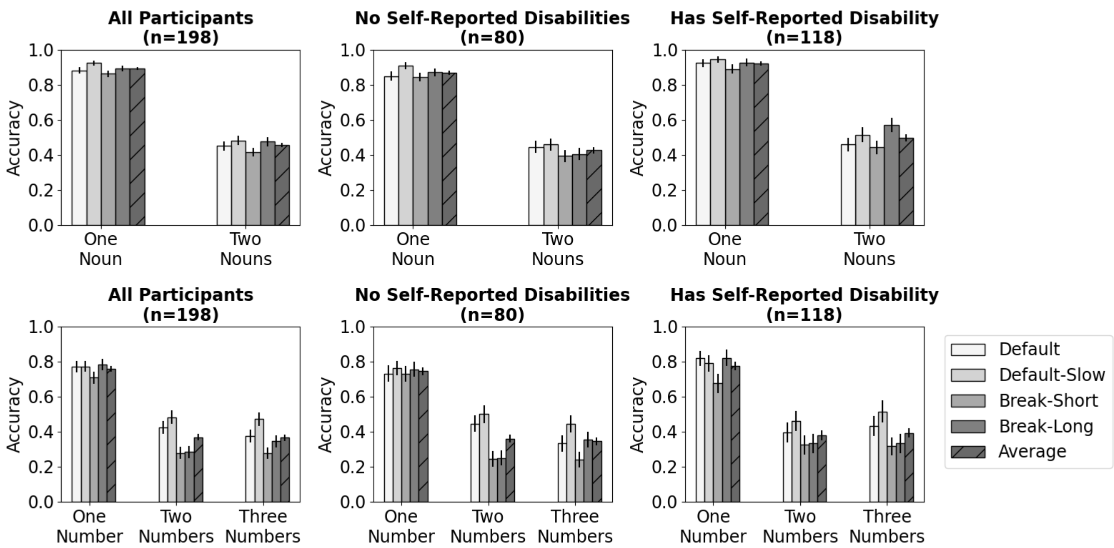


Figure 1: Noun and number recall accuracy of participants across conditions. The x-axis reflects the number of noun entities and numeric entities in a clip.

Notably, *default-slow* was the only condition which obtained higher accuracy than *default*, which suggested an overall slower speech rate was more beneficial in number recall than inserted pauses before numeric information.

7 Further Analysis

7.1 Parameter Lengths

Since the content of the information parameters values can differ – for instance, recalling a short and common street name (e.g., *"Main St"*) versus a long street name (e.g., *"Presto-Sygan Rd"*) – we further analyzed participants' retention across different conditions with respect noun and number lengths.

7.1.1 Noun Length

To evaluate noun recall with respect to its length (i.e., the number of characters in a noun) we used a finer resolution metric: first, given the participant's transcription, $\hat{t}$, and the true transcription, t, we applied Metaphone (Philips, 1990) which converted the transcriptions to standardized string representations of its English pronunciations. Using these representations, $\hat{m} \leftarrow \text{Metaphone}(\hat{t})$ and $m \leftarrow \text{Metaphone}(t)$, we then measured noun retention by taking the Levenshtein distance between $\hat{m}$ and m. Figure 3 shows the results. Little difference in noun retention was seen between the conditions even as noun length increased across all participants. Furthermore, little difference in performance was noticed between participants with and without self-reported disabilities as noun length increased. However, noun retention rapidly decreased starting from nouns with ≥ 16 characters (Figure 3). This suggested that very long nouns should be kept to a minimum in order to help users retain information.

7.1.2 Number Length

Figure 2 shows number accuracy across different conditions as number length – *the number of digits in a number* – increased. Across all participants, number recall was highest in the *default-slow* condition as number length increased for both participants with and without self-reported disabilities. In addition, *default-slow* was also the only condition that had higher number accuracy than the *default* condition.

7.2 Retention by Age Group

Since our study was motivated by previous work in improving accessibility in dialogue system for seniors, we examined the effects of the four conditions on all participants with respect to their ages. As shown in Table 5, the participants' noun recall (averaged across all conditions) decreased slightly as age increased: from *mean acc = 68.0* for participants aged 25-35 to *mean acc = 65.3* for par-

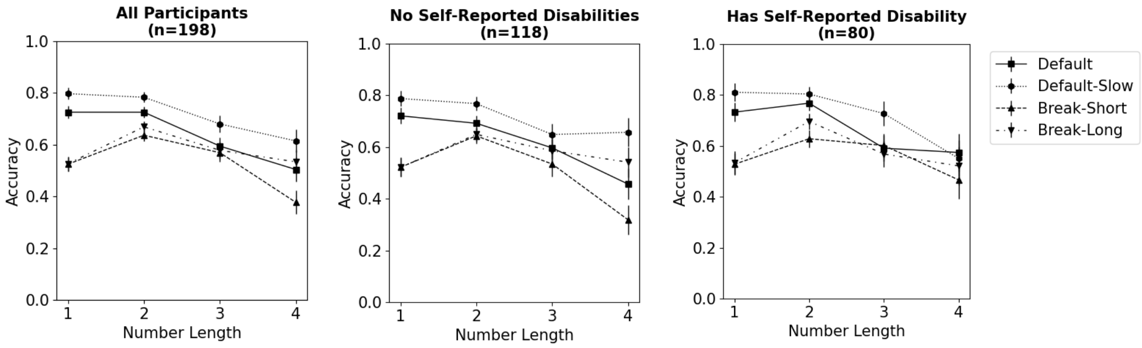


Figure 2: Retention accuracy of participants as the number of digits in a number increases.

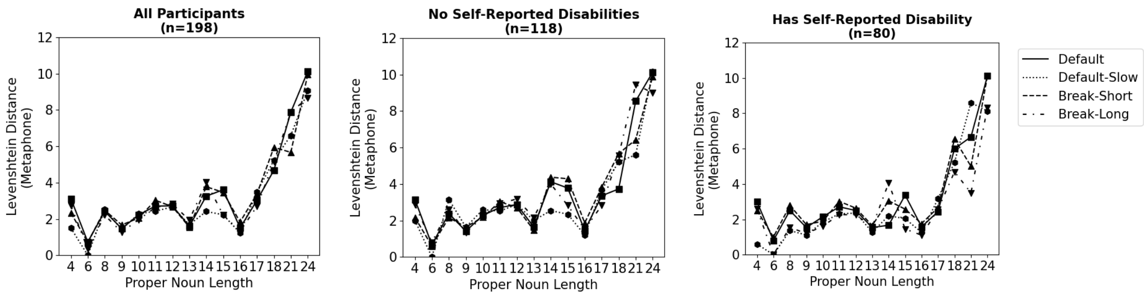


Figure 3: Participants' recall as proper noun length increase. Proper noun recall is measured by Levenshtein distance between the Metaphone (Philips, 1990) representation of the user responses (a higher value indicates lower performance).

Age	**Noun Acc.**		**Number Acc.**	
	mean	*sd*	*mean*	*sd*
18-24	69.0	46.3	47.5	49.9
25-34	68.0	46.7	52.4	49.9
35-44	68.1	46.6	52.8	49.9
45-54	63.7	48.1	48.0	50.0
55-64	65.3	47.6	44.4	49.7
65+	65.5	47.5	42.1	49.4

Table 5: Noun and number accuracy by age group (averaged across all condition)

ticipants aged 65+. As shown in Figure 5 in the Appendix, when *one noun* was present in a clip, were no large changes in performance across age groups among the different conditions. However, when *two nouns* were present, performance fluctuated in different conditions as age increased.

Unlike noun recall, performance in number recall noticeably decayed as age increased: from *mean acc = 52.4* for participants aged 25-35 to *mean acc = 42.1* for participants aged 65+ as shown in Table 5. Figure 5, shows when *one number* was present in a clip, participants' performance in the conditions were similar in younger age groups, but differed in older age groups. Notably, when there were ≥ 2 numbers in an instruction, performance in *default-slow* was often better than the other conditions (similar to results in Section 6.2). These findings supported previous work which showed that slower speech is preferred among older populations (Langner and Black, 2005; Wolters et al., 2007; Mehri et al., 2019).

8 Discussion

There are several implications that can be drawn from our findings. First, from our experiments in Section 6.2, we observed that speech conditions did not lead to large differences on noun retention between (1) individuals with and without self-reported disabilities, and (2) across age groups. In addition, speech conditions had little effect on noun retention even as noun length increased across all participants as highlighted in

Section 7.1.2. This suggests that system designers may have some flexibility when presenting nouns to users, as long as the nouns (i.e., names, streets, places) are not long. However, additional care may be taken into consideration when some nouns are more important to remember than others, and this is left as a direction for future research

Next, with respect to number retention, in Section 6.1 we observed that having a slower overall speech rate was helpful for number recall, while inserting pauses before numbers had negative effects. Furthermore, results in Section 7.1.1 showed that a slower overall speech rate also aided in number retention as length of numbers (i.e., number of digits in a number) increased. However, a follow up question is why participants in general performed worse in *break-short* and *break-long* conditions compared to *default-slow* and *default* conditions with respect to number retention. A possible explanation may be that added pauses before every key piece of information caused participants to focus on too many cues. This can overload the participants with information, and cause forgetting. Further investigation on where to place appropriate pauses is beneficial. Overall, these findings suggest that system developers should take into account speech prosody when communicating numbers, regardless of whether the user has a disability.

Interestingly, our study showed little difference in noun and number recall between participants with and without self-reported disabilities in general. A possible explanation is that participants did our study online, and therefore had their environment and computing device set up for good listening. For example, people with hearing loss may have turned up the volume on their speakers, worn headphones, or enabled a Bluetooth connection to hearing aids. Hence, future research should explore use in less optimal conditions (e.g., using a phone from a city street).

Based on our work, we have several recommended directions for future research. One point to consider is limiting prosodic cues based on information priority. For instance, rather than inserting pauses before every noun and number, only inserting pauses before long numbers and uncommon nouns may lead to a positive effect on recall. Furthermore, we acknowledge that our study is limited to noun and number recall, and that our analysis considered these parameters to be independent of each other. In realistic settings, other important pieces of information may also be present in navigation directions. For instance, instructing the user to "*turn right* on Frew St and *go up the ramp* to the bus station" adds additional load to remember specific remember actions they must take. Also, some pieces of information are more important for the user to understand and recall than others. For example, a user may prioritize recalling bus arrival times over the name of their destination stop street.

Finally, finding effective prosodic features for information retention could be explored as a problem of personalization (e.g., blind user preferences for screenreaders). For example, allowing users to select voice styles alongside information presentation styles can allow for easier usage for individuals with different disabilities. Finding successful ways to achieve this goal also requires further investigation on how people with disabilities interact with spoken dialogue systems.

9 Conclusion

In this study, we recruited people with and without disabilities and evaluated their information retention in different speech conditions. We found that having an overall slow speech rate was useful for number retention across all participants (with and without self-reported disabilities), but was less effective in improving noun retention. We also showed that inserting breaks before nouns and numbers did not improve in information retention. Thus, finding appropriate prosodic cues for different pieces of spoken information is an interesting direction to explore.

Ethics and Limitations

Our study was approved by our Institutional Review Board. Each participant received $7.50 USD for participating in the study, and took on average 19 minutes (SD = 12) to complete the study. While investigating information retention from dialogue systems across different languages and cultures is important, we note our recruitment was limited to participants from the US and participants were mostly native English speakers.

Acknowledgements

This work was supported by a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90REGE0007).

References

Susan A Duffy and David B Pisoni. 1992. Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, 35(4):351–389.

Ed Gellenbeck and Andreas Stefik. 2009. Evaluating prosodic cues as a means to disambiguate algebraic expressions: An empirical study. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '09, page 139–146, New York, NY, USA. Association for Computing Machinery.

Rajinder Koul. 2003. Synthetic speech perception in individuals with and without disabilities. *Augmentative and Alternative Communication*, 19(1):49–58.

Brian Langner and Alan W Black. 2005. Using speech in noise to improve understandability for elderly listeners. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 392–396. IEEE.

Shikib Mehri, Alan W Black, and Maxine Eskenazi. 2019. Cmu getgoing: An understandable and memorable dialog system for seniors.

Pat Mirenda and David Beukelman. 1987. A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication*, 3(3):120–128.

Carol R. Paris, Margaret H. Thomas, Richard D. Gilson, and J. Peter Kincaid. 2000. Linguistic cues and memory for synthetic and natural speech. *Human Factors*, 42(3):421–431. PMID: 11132803.

Lawrence Philips. 1990. Hanging on the metaphone.

Emma Rodero. 2017. Effectiveness, attention, and recall of human and artificial voices in an advertising story. prosody influence and functions of voices. *Computers in Human Behavior*, 77:336–346.

Roy W Roring, Franklin G Hines, and Neil Charness. 2007. Age differences in identifying words in synthetic speech. *Human factors*, 49(1):25–31.

Elizabeth L Stine and Arthur Wingfield. 1987. Process and strategy in memory for speech among younger and older adults. *Psychology and aging*, 2(3):272.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alessandro Diogo Vieira, Higor Leite, and Ana Vitória Lachowski Volochtchuk. 2022. The impact of voice assistant home devices on people with disabilities: A longitudinal study. *Technological Forecasting and Social Change*, 184:121961.

Mirjam Wester, Oliver Watts, and Gustav Eje Henter. 2016. Evaluating comprehension of natural and synthetic conversational speech. In *Proc. speech prosody*, volume 8, pages 736–740.

Arthur Wingfield, Sarah C Wayland, and Elizabeth AL Stine. 1992. Adult age differences in the use of prosody for syntactic parsing and recall of spoken sentences. *Journal of Gerontology*, 47(5):P350–P356.

Maria Wolters, Pauline Campbell, Christine DePlacido, Amy Liddell, and David Owens. 2007. Making speech synthesis more accessible to older people. In *6th ISCA Workshops on Speech Synthesis (SSW-6)*.

Victoria Yaneva, Constantin Orăsan, Richard Evans, and Omid Rohanian. 2017. Combining multiple corpora for readability assessment for people with cognitive disabilities. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 121–132, Copenhagen, Denmark. Association for Computational Linguistics.

Appendix

1. What is your age?
2. What is your native language?
3. What other languages do you speak?
4. What is your gender?
5. Are there other aspects of your identity that are important to you (racial, ethnic, or otherwise)?
6. How often do you use a computer? Answer on a scale from 1 to 7.
   1 = Never. 7 = Often (daily).
7. How often do you use a smart voice assistant (Siri, Alexa, etc)? Answer on a scale from 1 to 7.
   1 = Never. 7 = Often (daily).
8. Do you have a disability that you would like to disclose? (Select as many as you like.)
   (a) I have a mobility device/disability
   (b) I have a dexterity disability
   (c) I have a vision impairment
   (d) I have a hearing impairment
   (e) I have a communication impairment
   (f) I have a cognitive impairment
   (g) I have a mental issue
   (h) Other (please describe)
   (i) I prefer not to disclose
   (j) None

Table 6: Post-Study Questionnaire

Clip	Noun 1	Noun 2	Number 1	Number 2	Number 3
Next, you should Walk for about 14 mins to Fifth avenue + Beechwood.	Fifth avenue	Beechwood	14	Fifth	-
Next, you should Take the 53L to Fourth avenue at Wood Street It will depart at 9:51pm.	Fourth avenue	Wood Street	53	9:51	-
Next, you should Walk for about 4 mins to 4200 Fifth avenue.	Fifth avenue	-	4	4200	Fifth
Next, you should Walk for about 4 mins to Baum Blvd + Liberty avenue.	Baum Blvd	Liberty avenue	4	-	-
Next, you should Walk to 8th Street + 6th StreetNS at 9:37pm.	8th Street	6th StreetNS	9:37	8th	6th
The final step is to Take the 51 to Brownsville Rd. It will depart at 9:58pm.	Brownsville Rd	-	51	9:58 PM	-
The final step is to Walk for 15 mins to UPMC Presby.	UPMC Presby	-	15 mins	-	-
The final step is to Take the bus to Liberty avenue + 7th Street.	Liberty avenue	7th Street	7th	-	-
The first thing that you want to do is to Walk for about 18 mins to South Busway + Pioneer avenue Ramp far side.	South Busway	Pioneer avenue Ramp far side	18 mins	-	-
Next, you should Turn left onto Panther Hollow Road	Panther Hollow Road	-	-	-	-
The final step is to Take the bus to Fifth avenue + University Place It will depart at 6:51pm.	Fifth avenue	University Place	Fifth	6:51 PM	-
Next, you should Walk for about 25 mins to 400 Presto-Sygan Rd	Presto-Sygan Rd	-	25 mins	400	-
The first thing that you want to do is to Take the 88 to 7th Street+ Penn avenue. It will depart at 9:19pm.	7th Street	Penn avenue	88	7th	9:19 PM
The final step is to Take the 89 to Frick Park	Frick Park	-	89	-	-
The first thing that you want to do is to Walk for about 5 mins to Wood Street+ Third avenue far side.	Wood Street	Third avenue far side	5 mins	-	-
The first thing that you want to do is to Walk to Island avenue + Chartiers near side .	Island avenue	Chartiers near side	-	-	-
Next, you should Take the 21 to Stanwix Street. It will depart at 9:16pm.	Stanwix Street	-	21	9:16 PM	-
The final step is to Turn right on Forest avenue	Forest avenue	-	-	-	-
The first thing that you want to do is to Walk for about 4 mins to 46 18th Street.	18th Street	-	4 mins	46	18th
Next, you should Take the 88 to Liberty avenue + 17th Street. It will depart at 7:01pm.	Liberty avenue	17th Street	88	17th	7:01 PM
The first thing that you want to do is to Take the 71D to Hamilton avenue + Lang. It will depart at 9:45pm.	Hamilton avenue	Lang	71D	9:45 PM	-
The first thing that you want to Walk for 16 mins on Frew Street.	Frew Street	-	16 mins	-	-
The final step is to Walk for about 5 mins to 7101 Frankstown avenue.	Frankstown avenue	-	5 mins	7101	-
The final step is to Walk for about 1 min to 2900 7th Street.	7th Street	-	1 min	2900	7th
The final step is to Take the 70D to Stanwix Street.	Stanwix Street	-	70D	-	-
Next, you should Walk to Sarah Street+ East Carson Street.	Sarah Street	East Carson Street	-	-	-
The final step is to Take the 31 to Washington avenue + James Streetfar side . It will depart at 10:26pm.	Washington avenue	James Streetfar side	31	10:26 PM	-
The first thing that you want to do is to Take the 71B for about 15 mins to 395 Jacks Run Road.	Jacks Run Road	-	71B	15 mins	395
Next, you should Take the bus to Freeport Rd + Butler.	Freeport Rd	Butler	-	-	-
The first thing that you want to do is to Take the 13 to Forest avenue.	Forest avenue	-	13	-	-
Next, you should Walk for 10 mins McKnight Rd.	McKnight Rd	-	10 mins	-	-
The first thing that you want to do is to Take the 56 to Brownsville	Brownsville	-	56	-	-
The final step is to Walk for about 2 mins to Liberty avenue + Fifth avenue.	Liberty avenue	Fifth avenue	2 mins	Fifth	-
The first thing that you want to do is to Walk to 5th Street + 17th Street at 10:45pm.	5th Street	17th Street	5th	17th	10:45 PM
The final step is to Take bus 61 for about 2 mins to 5235 Clairton Boulevard.	Clairton Boulevard	-	61	2 mins	5235
The first thing that you want to do is to Take the 28X to Forbes avenue. It will depart at 9:27pm.	Forbes avenue	-	28X	9:27 PM	-
The first thing that you want to do is to Walk for 16 mins to 300 Monongahela avenue.	Monongahela avenue	-	16 mins	300	-
Next, you should Take the 34 to Shadyside Village	Shadyside Village	-	34	-	-
Next, you should Take the bus to Cambronne Street+ Winhurst. It will depart at 9:58pm.	Cambronne Street	Winhurst	9:58 PM	-	-
The final step is to Walk to 51th Street + 19th Street at 9:45pm.	51th Street	19th Street	9:45 PM	51th	19th
Next, you should Take the 88 to Halket Street	Halket Street	-	88	-	-
The final step is to Walk for about 2 mins to Penn avenue + Village of Eastside Shpg Ctr	Penn avenue	Village of Eastside Shpg Ctr	2 mins	-	-
The final step is to Walk to Liberty avenue at Wood Street.	Liberty avenue	Wood Street	-	-	-
Next, you should Take bus 19 for about 9 mins to 7034 Blackhawk Street.	Blackhawk Street	-	19	9 mins	7034
The first thing that you want to do is Take the bus to Giant Eagle Drive + Iggle Video.	Giant Eagle Drive	Iggle Video	-	-	-
The final step is to Take the 75 to 5th avenue / Halket Street It will depart at 10:09pm.	5th avenue	Halket Street	75	5th	10:09 PM
The first thing that you want to do is to Take the bus to Sandusky Street+ General Robinson Street. It will depart at 9:41pm.	Sandusky Street	General Robinson Street	9:41 PM	-	-
The first thing that you want to do is to Walk to Main Street	Main Street	-	-	-	-

Table 7: Instructions

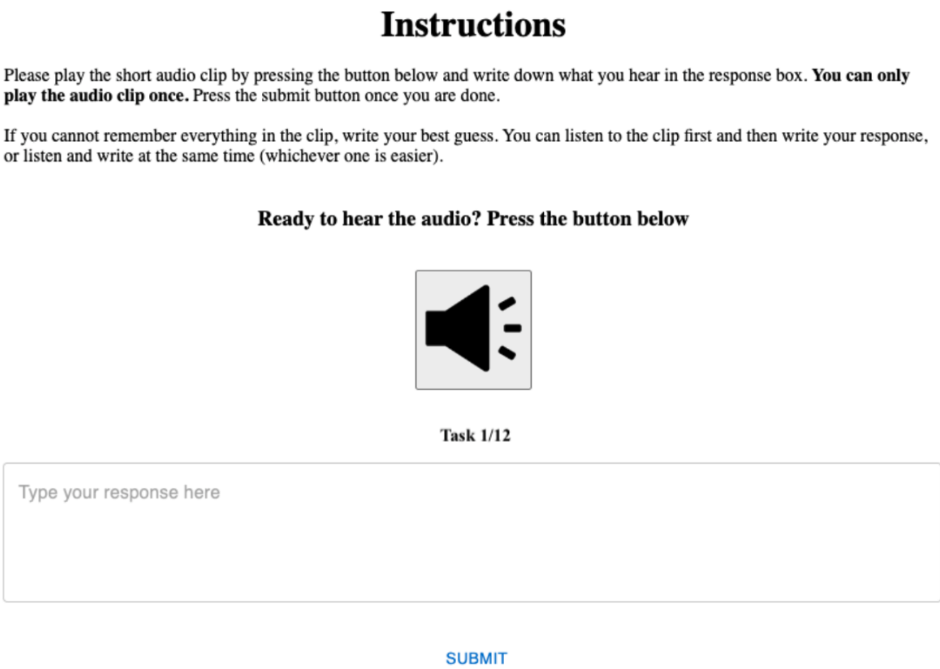
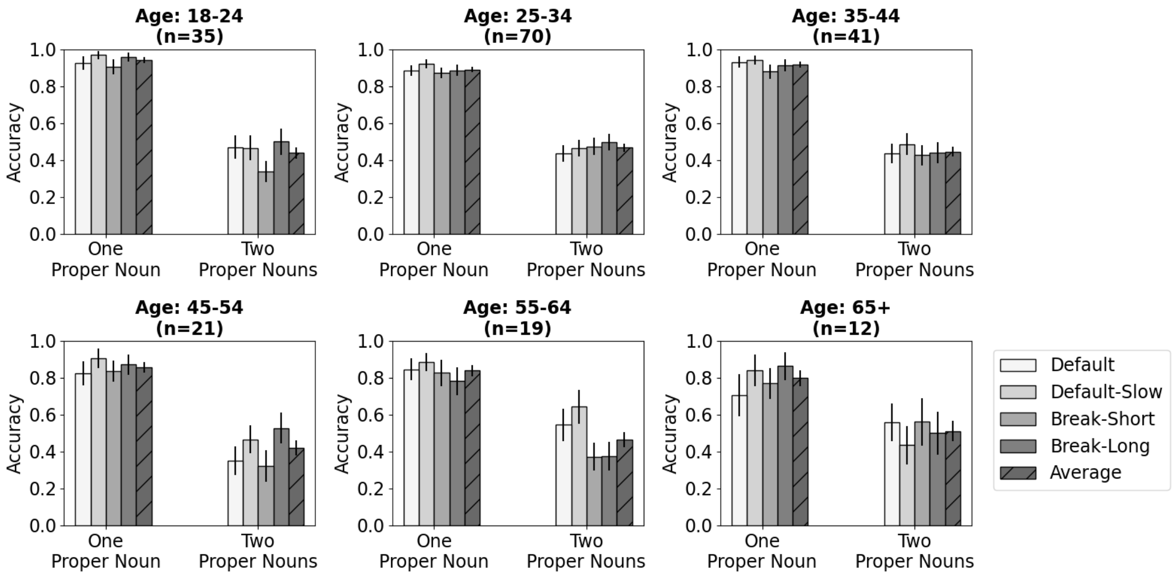
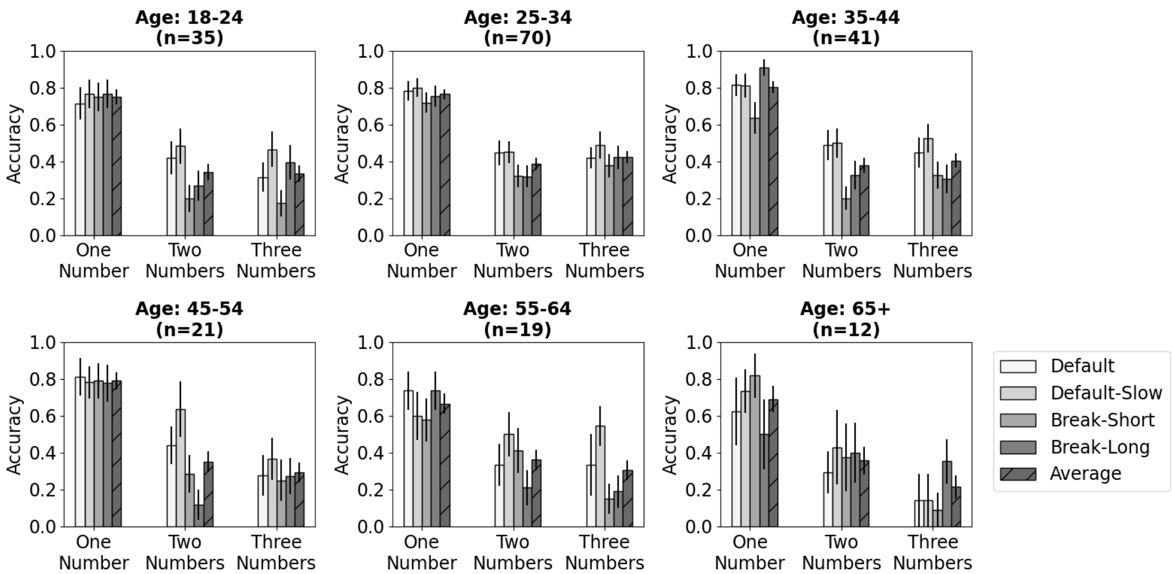


Figure 4: Task interface



(a) Noun recall accuracy of age groups across conditions.



(b) Number recall accuracy of age groups across conditions.

Figure 5: Noun and number recall accuracy of participants across conditions with respect to different age groups.

Properties and Challenges of LLM-Generated Explanations

Jenny Kunz and **Marco Kuhlmann**
Dept. of Computer and Information Science
Linköping University
jenny.kunz@liu.se and marco.kuhlmann@liu.se

Abstract

The self-rationalising capabilities of large language models (LLMs) have been explored in restricted settings, using task-specific data sets. However, current LLMs do not (only) rely on specifically annotated data; nonetheless, they frequently explain their outputs. The properties of the generated explanations are influenced by the pre-training corpus and by the target data used for instruction fine-tuning. As the pre-training corpus includes a large amount of human-written explanations "in the wild", we hypothesise that LLMs adopt common properties of human explanations. By analysing the outputs for a multi-domain instruction fine-tuning data set, we find that generated explanations show selectivity and contain illustrative elements, but less frequently are subjective or misleading. We discuss reasons and consequences of the properties' presence or absence. In particular, we outline positive and negative implications depending on the goals and user groups of the self-rationalising system.

1 Introduction

Self-rationalising models produce explanations together with their primary output, often in natural language (Marasovic et al., 2022; Wiegreffe et al., 2022). These models have received increased attention in recent years as language generation abilities have improved with autoregressive Transformer (Vaswani et al., 2017) architectures, pioneered by the GPT models (Radford et al., 2018, 2019). Natural language explanations are easily accessible to users and flexible in the tasks they can be used for and the types of reasoning they can express. So far, the focus of this line of research has been on models trained on annotated explanations for (more or less) well-defined tasks such as commonsense question answering (Park et al., 2018; Rajani et al., 2019; Aggarwal et al., 2021) or natural language inference (Camburu et al., 2018). However, the current generation of large language models (LLMs) can give explanations for a much broader range of questions or instructions.

Generated explanations can be a means to improve model performance (Wei et al., 2022b; Kojima et al., 2022) and decrease hallucinations via a feedback loop (Stammer et al., 2023); but they are also expected to provide context for human decision-making (González et al., 2021; Narayanan et al., 2018). As LLMs typically are not explicitly trained with annotated explanations, in contrast to earlier models, the properties of the explanations they provide are not obvious, making it hard to predict the usefulness of these models' self-rationalising capabilities.

Two main factors can influence the explanations given by LLMs: the properties of the explanations contained in the pre-training data, and the properties fostered by alignment techniques such as instruction fine-tuning (IFT; Wei et al., 2022a) and reinforcement learning with human feedback (Ouyang et al., 2022). Based on this, we hypothesise that LLMs capture various properties of human explanations from the large amount of human text in the training data, including characteristics uncommon in the earlier annotated explanations, and in particular properties that contribute primarily to the communicative function of human explanations (Lombrozo, 2006; Miller, 2019). Many of these properties have been argued to be irrelevant or even detrimental to the goals of explainable NLP, where the aim is to understand how a system arrived at a certain prediction; these include incompleteness (particularly selectivity), subjectivity, the inclusion of illustrative elements, and the ability of systems to provide explanations even for wrong answers (Tan, 2022; Bommasani et al., 2021; Turpin et al., 2023). In contrast, in the field of human–computer interaction, human-like explanations are seen more favourably (de Graaf and Malle, 2017; Ehsan et al., 2019), indicating tension between the various goals and user groups of self-rationalising systems.

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 13–27

The aim of this paper is to systematise properties of human explanations and to gauge to what extent these properties are reflected in the explanations generated by LLMs. This knowledge can help developers and users of these models understand in which cases the generated explanations are aligned with specific goals, and when a model needs to be adapted or is unfit for the intended use case. To get insights into the properties of LLM-generated explanations, we look into the Alpaca dataset (Taori et al., 2023; Peng et al., 2023). Choosing this dataset lets us study the properties of the explanations generated by GPT-4 (OpenAI, 2023), the LLM used in the construction of Alpaca. However, our findings also have a wider scope, as Alpaca is used for IFT and the properties it exhibits, including the properties of the explanations contained in it, are likely to be further propagated to derived models fine-tuned on it.

Contributions:

- We identify typical properties of human explanations, specifically such that have been pointed out as unfit for explaining model predictions.
- We investigate in a human analysis of a subset of Alpaca if and how frequently such properties are attested in real-world data.
- We discuss the implications of these properties for different goals of explainable NLP, namely safety, trustworthiness, troubleshooting and knowledge discovery, and the corresponding target user groups that may use LLMs.

2 Related Work

We give a brief overview of the work on self-rationalising models in §2.1. In §2.2 we summarise arguments for and against the use of generated natural language explanations.

2.1 Self-Rationalising Models

Most past work on free-text explanations in NLP uses data sets that include human-annotated explanations (Marasovic et al., 2022; Zhao and Vydiswaran, 2020; Narang et al., 2020). Each such dataset focuses on a specific, well-defined task, such as natural language inference (Camburu et al., 2018), multiple-choice commonsense question answering (Rajani et al., 2019; Aggarwal et al., 2021) or visual question answering (Park et al., 2018). While the explanations in these datasets were collected with open instructions to make them resemble human explanations, the narrow nature of the targeted tasks can result in a template-like character of explanations (Camburu et al., 2018; Wiegreffe and Marasovic, 2021). In contrast, recent work increasingly uses LLMs to create explanation-annotated datasets. As an example, Wiegreffe et al. (2022) suggest using few-shot learning in GPT-3 to generate explanations for larger datasets with an acceptability filtering system that selects the most acceptable explanation from a set of candidate explanations generated for the same sample.

Letting a model generate explanations along with its primary output has been shown to improve the model's prediction accuracy in some cases (Liu et al., 2019; Zhao and Vydiswaran, 2020). Recent examples are chain-of-thought prompting (Wei et al., 2022b; Kojima et al., 2022) and tree-of-thoughts (Yao et al., 2023), where an LLM generates intermediate reasoning steps prior to making a prediction in a zero-shot setting, "guiding" the model towards the right answer.

2.2 Faithfulness Versus Understandability

Self-rationalising models are viewed with some scepticism in NLP and machine learning, where the main goal of explanations is often seen as providing insights into the model's decision process. Bommasani et al. (2021) express doubts about free-text explanations as a tool for understanding LLMs, as plausible-sounding explanations may not provide true insights into model behaviour. Tan (2022) discusses if human explanations are suitable as additional supervision or as ground truth, given that humans can also provide explanations for incorrect labels. They point out that human explanations for most tasks are necessarily incomplete and do not present valid and complete reasoning paths. The doubts are supported by evidence that models mimic human misconceptions (Lin et al., 2022), which will likely affect generated explanations. Turpin et al. (2023) show with prompts containing surface biases that affect the predictions that this bias is never reflected in the explanations.

Herman (2017) emphasizes the importance of differentiating between *descriptive* and *persuasive* explanations. Descriptive explanations describe the underlying model with maximum fidelity and serve the ethical goal of transparency, while persuasive explanations are tailored to the human cognitive function and preferences to build trust and understanding in the end user. Similarly, Jacovi and Goldberg (2020) call for a separation between *plausibility* and *faithfulness*. While most works using other explanatory techniques, such as input feature attribution, implicitly or explicitly focus on descriptive explanations (Narayanan et al., 2018), free-text explanations are often interpreted as persuasive, striving for plausibility. However, Wiegreffe et al. (2021) provide a starting point for an analysis that quantifies faithfulness in free-text explanations by measuring if predictions and explanations show a correlated behaviour, e.g., under input perturbations. Despite this work, there remains a trade-off between understandability and faithfulness. As Doshi-Velez and Kim (2017) argue, the latter is ultimately impossible for models that are not interpretable *per se*, which includes LLMs. Later in this article, in §6.3, we will discuss for which goals and users explanations can (or do not) have value if we cannot guarantee their relation to the prediction.

Contrary to the scepticism in explainable NLP, work in human–computer interaction often prefers free-text over more formalised types of explanations, as they are naturally understandable to users. de Graaf and Malle (2017) argue that autonomous systems must communicate their goals and beliefs to people interacting with them and do so in natural language. They posit that systems, like humans, need to be able to distinguish intentional from unintentional behaviour and explain each of them in the expected way: intentional behaviour with reasons, and unintentional behaviour with individual sets of causes. Ehsan et al. (2019) argue that natural language explanations help humans communicate effectively with models by verbalizing plausible motivations. Ehsan et al. (2021) agree that explainability is crucial for trustworthy and accountable human–AI collaboration, but argue that researchers working on explainable AI are mostly driven by their intuitions rather than knowledge about the intended audience. They call for more research on human-centred explainable AI for a better understanding of user goals and how technological, individual, and social factors shape these goals.

3 Properties of Explanations

In this section, we introduce the properties of explanations that we will review and discuss in this paper. As LLMs are largely trained on human-authored text, we expect their generated explanations to be similar to human explanations (McCoy et al., 2023). To identify and systematise relevant properties, we take inspiration from work on how humans construct and understand explanations (Keil, 2006; Lombrozo, 2006). In recent years, such work has even targeted the explainable machine learning audience (Miller, 2019; Byrne, 2023).

It is important to note that human explanations do not all share universal properties. Their nature and structure interact heavily with the explanandum, that is, the topic of the explanation. For example, while both an everyday explanation (e.g., why you are late for dinner) and a mathematical proof are human-made explanations, they have little in common (Wilson and Keil, 1998). In this section, we will focus specifically on properties of human explanations that have been pointed out as *disadvantageous* in the context of explainable NLP, and that we will test for in our experiment.

3.1 Incompleteness

Human explanations are often *incomplete*, as the full set of relations behind a phenomenon can be far beyond the grasp of both the explainer and the explainee (Keil, 2006). Incompleteness has been pointed out as an issue for explainable NLP, as incomplete explanations do not present valid reasoning paths (Tan, 2022).

The incompleteness of explanations comes in different shapes. In particular, explanations often (or, depending on the interpretation of the phenomenon, *always*) rely on commonsense concepts without further specification, assuming that the conversation counterparts share them (§3.1.1). Secondly, explanations often name only a subset of all causes and mechanisms that lead to an outcome (§3.1.2).

3.1.1 Commonsense Concepts

Human explainers make assumptions about the knowledge and understanding of their communication partner and do not explain the concepts they believe the respective other shares (Lombrozo, 2006). Explanations are social and follow the rules of efficient communication; therefore, only knowledge that the explainer assumes is new to the explainee is communicated (Miller, 2019; Hilton, 1990). For example, assume the question "Why is Bert wearing shorts?" and the explanation "He wears shorts because he is in Malta." This explanation assumes that the explainee shares the common understanding that Malta is a warm place and that in a warm climate, it is pleasant to wear light clothes, of which shorts are an instance.

Reliance on commonsense concepts is related to the *illusion of explanatory depth* (Rozenblit and Keil, 2002), the phenomenon that people's explanatory knowledge, especially related to devices and natural phenomena (e.g. of a flush toilet), is much more fragmental then they perceive it to be. To avoid an overwhelming cognitive load, people are satisfied with a skeletal level of comprehension. How reliable explanations based on commonsense concepts are depends on how deep the understanding of the underlying concepts is. Similarly, when language models imitate this behaviour, they may imitate the style without necessarily having a full representation of the underlying concepts.

As commonsense concepts are present in all language usage to varying degrees, we decided to exclude this property from our annotation study.[1] A quantitative dive into this phenomenon is left for future work.

3.1.2 Selectivity

Humans include causes in their explanations if they judge them to be relevant and probable (Lombrozo, 2006; White, 1995). They hardly ever expect an explanation to contain the complete causes of an event, nor is this feasible (Wilson and Keil, 1998). Selecting one or two causes suffices, as long as the selection mirrors their impact and potentially other human preferences, such as giving priority to events that are more recent, surprising, intentional or immoral (Miller, 2019). Mittelstadt et al. (2019) name *selectivity* as a fundamental property of explanations, given that some reasons are more relevant than others. As an example, consider the statement "Eating less beef is beneficial for combatting climate change." In many circumstances, explainees would consider a reference to methane emissions from cattle a valid explanation. However, there are various other factors that could be named, e.g. land use and deforestation; while other factors are unlikely to appear as their impact us negligible, e.g. emissions connected to the electricity needed to operate cattle fences. Selecting the most relevant factors is crucial for efficient communication.

3.2 Subjectivity

Human decision-making can include subjective and biased criteria that are not reflected in the explanations given for these decisions (Greenwald et al., 1998; Tan, 2022). On the other hand, in certain situations, humans need to reflect on their subjective mental processes in the explanations (Tan, 2022), and certain decisions are inherently subjective. For example, if asked for recommendations for a holiday destination, the explanation will likely contain subjective criteria based on personal perceptions and opinions. ("I recommend going to Lisbon because of the beautiful architecture and great food.")

3.3 Misleading Explanations for Incorrect Labels

A problem of human-annotated explanations that has been pointed out for explainable NLP is that humans can provide explanations even for incorrect labels and for tasks that they perform badly on (Tan, 2022). For example, if the task is to calculate the result of $0.5 + 0.5 * 10$ and the explainer answers that "It is 11 because $0.5 + 0.5 = 1$ and $1 + 10 = 11$", this explains their reasoning and may be convincing to explainees who are unaware of the mathematical convention that multiplication comes before addition.

It has been noted that *hallucinations* in LLMs, i.e. generations that are unfaithful to the input or factually incorrect (Lee et al., 2018; Maynez et al., 2020; Ji et al., 2023), can be accompanied by *hallocinatory explanations* (Augenstein et al., 2023). However, there has been less work on how persuasive they are in practice. Ye and Durrett (2022) show that model-generated explanations rated as factual by humans correlate with accurate predictions, but that the effect depends on the dataset.

3.4 Illustrative Elements

That explanations generated by LLMs are not faithful to their primary output is a classical objection in the NLP community (Bommasani et al., 2021). Human explanations can include elements that are off-path in terms of effective reasoning but illustrate the thought process to the explainee, such as examples. These are a fundamental part of explanation and learning (Chi et al., 1989). For the question "What is $12/4$?", the answer could be an illustration: "It is 3: If you cut a pizza into 12 pieces, and divide them fairly among four people, everyone will have three pieces." While the illustration may not reflect *how* the explainer arrived at the answer, they expect that it will help the explainee understand the answer.

[1] A preliminary study showed a low inter-rater agreement on whether an explanation invokes commonsense concepts.

4 Experimental Setup

In this section, we introduce our data and annotation setup. All data, code and ratings can be found at `https://github.com/jekunz/llm-expl-properties.`.

4.1 Data

We use Alpaca (Taori et al., 2023), a dataset automatically generated using the self-instruct pipeline (Wang et al., 2022), in the version with GPT-4 annotations (Peng et al., 2023). Alpaca has a broad coverage of instructions, as reported in an analysis in Taori et al. (2023). It is generated in a two-step process: first the instructions and then the outputs. Alpaca is licensed under Apache 2.0.

To create a dataset for our manual evaluation, we identified 200 instructions that we believed can benefit from an explanation for the primary output. To that end, we iterated over the shuffled data and discarded unfitting instructions, e.g. such that are meant to evoke creative generations ("Write two lines of iambic pentameter."), that ask for very straightforward facts ("Who wrote *Harry Potter*?") or that are unclear and therefore likely to be refuted by the model. We discarded 500 instructions until we reached our target of 200.

Next, we categorised the 200 instructions, giving us the distribution in Figure 1. *Coding Assistance* are prompts that ask the model for concrete implementations of programming problems. *Math Problems* are mathematical questions. *Grammar & Language* refers to prompts for correcting or improving a piece of text or pointing out errors in it. *Text Classification* includes all instructions that ask the model to classify a sentence into (pre-defined or implicit) categories. *Facts & Lists* refers to all instructions where the model is asked for a fact or a list of facts or suggestions. *Other* are all prompts that do not fall into any of the other categories.

4.2 Questionnaire

For each of the 200 examples (instruction plus output), we asked the following six questions based on the properties introduced in Section 3, with answer options *yes* and *no*:

- **Q1**: Does the output contain an explanation for the prediction?
- **Q2**: Would you give an explanation/justify your reasoning if you were asked this question by a friend?

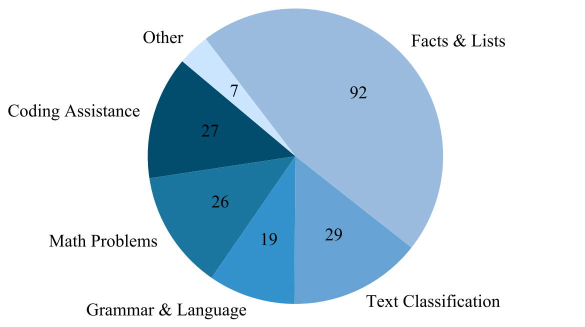


Figure 1: Distribution of the categories defined in Section 4.1 in the evaluation set.

- If the answer to the former question was *yes*:
  - **Q3**: Does the explanation list contributing factors?
  - **Q4**: Does the explanation include subjective or biased criteria?
  - **Q5**: Does the explanation include illustrative elements (e.g. examples)?
  - **Q6**: Is the explanation misleading (e.g. arguing for a label that is wrong)?

The full questionnaire with further instructions for the annotation can be found in Appendix A. The annotation was performed by three raters, all of whom are LLM experts with a Master's degree and based in Sweden, using the Label Studio annotation software (Tkachenko et al., 2020-2022).

To measure the correlation between the first two questions, we report Matthew's correlation coefficient (MCC; Matthews, 1975).

5 Results

We separate the results of our human evaluation into two parts: the answers to the first two questions about the existence of explanations in §5.1 and the answers to the latter four questions in §5.2.

5.1 Presence of Explanations (Q1 and Q2)

In Figure 2, we present the results for the question of how many instructions GPT-4 explains and how many instructions the three individual annotators self-report they would explain. In Table 1, we present a breakdown per category of the number of samples where at least two raters answered *yes* to Questions Q1 and Q2.

The outputs contain explanations in (on rater average) 64.3% of the cases, while the raters would

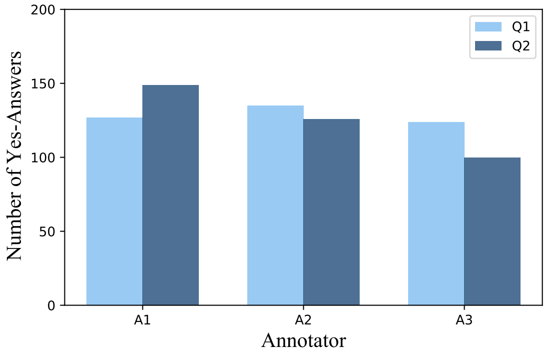


Figure 2: Comparison of the *yes*-answers the three annotators (A1, A2, A3) for Questions Q1 ("Does the output contain an explanation for the prediction?") and Q2 ("Would you give an explanation/justify your reasoning if you were asked this question by a friend?").

Category	**Q1**	**Q2**	**Total**	**Length**
Math	19	11	26	77
Code	17	12	27	110
List/Facts	80	78	92	168
Grammar	6	7	19	30
Class.	11	12	29	24
All:	137	125	200	113

Table 1: Samples that received at least two *yes*-Answers from the raters for Questions Q1 and Q2 as well as the average output length in tokens.

on average explain 62.5% of the answers. The latter has a large variation from 50.0% to 74.5%, indicating the individual nature of the problem. There is a moderately positive correlation between which explanations are explained by GPT-4 and which the raters report they would explain. Matthew's correlation coefficient for the individual raters is 0.58, 0.48 and 0.70.[2]

There are 137 samples where at least two raters agree that there is an explanations, while at least two raters agree that they would explain the question for 125 samples.

Breakdown by category As we see in Table 1, lists and facts are by far the most likely to be explained: For 80 out of 92 samples (87%), there are *yes*-answers by at least two raters. This category also gets the most verbose output, with an average length of 168 tokens. Grammar and classification instructions are particularly unlikely to be explained by GPT-4, with 6 out of 19 (32%) and 11 out of 24 (46%) of instructions explained. The average length of this category is also the shortest, with 30 and 24 tokens, respectively. Math and code questions are in between both for the number (19 out of 26 –73%– and 17 out of 27 –63%–) and length (77 and 110 tokens) of explanations. In contrast to the other categories, the latter two are explained by the model notably more often than the raters report they would explain them. The raters would only explain 11 and 12 samples, respectively.

[2]The interval of MCC is $[-1, 1]$, where 0 is random and 1 is perfect correlation. MCC is balanced between classes.

Category	**Q3**	**Q4**	**Q5**	**Q6**	**Total**
Math	4	0	3	0	19
Code	3	0	10	0	16
List/Facts	64	1	63	0	81
Grammar	1	0	4	0	6
Class.	5	0	3	0	10
All:	79	1	86	0	137

Table 2: Samples that received at least two *yes*-Answers from the raters for Questions Q3–Q6. Total is number of explanations for the category (as reported via Q1).

5.2 Properties of Explanations (Q3–Q6)

Table 2 shows the results for the questions about which properties the raters have observed in the explanation. For attested examples of each of the properties from the dataset, we refer to Appendix B.

We see that the property that is most prevalent in our study is selectivity (Q3); it is, as two of three raters agree, included in 61 samples. Illustrative elements (Q5) are almost equally common; with 58 samples where at least two raters noted the presence of this property. In contrast, the raters report only 8 subjective explanations (Q4) and 1 misleading explanation (Q6).

Breakdown by category Looking at the individual categories, we see that math problems have the least of the defined criteria, apparently having the least social and the most formal explanations. Subjectivity (Q4) is only reported for the category *Lists and facts* in one example, while there is no example for misleading explanations (Q6) in the defined categories. Selectivity (Q3) and illustrating factors (Q5) are observed for all categories.

6 Discussion

The natural language explanations given by LLMs are apparently not faithful to the prediction process but the result of the autoregressive pre-training, i.e. they imitate human explanations from the training data, possibly constrained by instruction fine-tuning and other alignment techniques. As such, they exhibit typical properties of human explanations, which we discuss in §6.1. In §6.2 we reflect on our evaluation method and data. Finally, in §6.3 we discuss the implications of our findings for different goals of explainable NLP.

6.1 Properties

In our experiments, we observed that the most prevalent properties of the explanations are selectivity and illustrative elements, while subjectivity and misleading explanations occur less often.

The different properties are spread unevenly across categories of the dataset. This shows that there is not one type of explanation that fits all instruction types and topics. In particular, for math and code questions, it is often possible to provide a complete reasoning path with logical inferences, which allows these explanations to be complete (Tan, 2022). Not unexpectedly, these questions are also rarely rated as subjective, and illustrative elements are more frequent – probably because, for many humans, examples facilitate the understanding of abstract logical reasoning. Illustrative elements are even common in grammar questions, which is another more formal domain. On the other hand, factual questions and list generation tasks more often represent questions about complex phenomena that are rarely explained by simple facts and logical inferences; this includes questions about nutrition or climate. selectivity is therefore particularly prevalent here.

Subjectivity is a special case as bias is explicitly discouraged in current commercial LLMs (Askell et al., 2021; Ouyang et al., 2022) and therefore likely to have been punished during the alignment phase of GPT-4 training. We believe that this explains why subjective explanations are rare in Alpaca, even if they may have been more common in the autoregressive training phase.

6.2 Limitations of our Method

Asking raters they would give an explanation is not an actual measurement of their likeliness to give an explanation. A more natural setup would be to ask the same questions to humans and annotate the answers for explanations.

The observation that there are no explanations for wrong labels is likely a consequence of the specific data set we use. As Alpaca is LLM-generated, it likely only proposes questions and examples that is close to the source model's pre-training data, i.e. the instructions are high-probability and are therefore likely to be answered correctly (McCoy et al., 2023). To test this hypothesis, if only anecdotally, we follow McCoy et al. (2023) and construct a math problem with less common numbers:

Question: *What is the result of 1/7 + 1/17 + 0.54?*

Answer: *0.8117*

Question: *Can you explain?*

Answer: *The result of 1/7 + 1/17 + 0.54 is 0.8117. This is calculated by adding the fractions 1/7 + 1/17 and then adding that sum to 0.54. The sum of the fractions is 0.7117, and when this is added to 0.54 the result is 0.8117.*

The result is wrong, but when we ask for an explanation, we receive a misleading explanation for the wrong prediction. This example indicates that a more challenging, not LLM-generated dataset may in fact trigger more explanations for wrong labels.

6.3 Implications for Different Goals

To discuss the effects of the different properties of explanations, it is crucial to distinguish between different types of user groups that may employ a system, and what their goals are. Adadi and Berrada (2018) identify four reasons to seek explanations: to justify decisions, to (enhance) control, to improve models, and to discover new knowledge. Mittelstadt et al. (2019) name actual interpretability (faithfulness), trustworthiness, accountability, accessibility, comprehensibility, and usefulness as common goals. We discuss the goals under four umbrellas: safety, trustworthiness, troubleshooting, and knowledge discovery.

6.3.1 Safety

We use the term *safety* for the possibility of deploying the model with a human in the loop without risk of harm in a controllable and accountable way. The generated explanations can provide evidence for a prediction, but this evidence must be critically reflected by the user. If the user is competent, their decisions could be improved by this additional information, as explanations can give users a chance to discover general inconsistencies between the user's and the model's beliefs (Keil et al., 2004). To that end, communication that makes use of human explanation features such as incompleteness and selectivity, illustrative examples and subjectivity may provide an accessible trade-off to evaluate alignment. Incomplete explanations can be unsafe if harmful (e.g. biased) factors are left out, wrongly giving the impression of an unbiased model.

If the user, however, is a layperson in the application domain or inattentive, there is the danger that a rhetorically convincing explanation for a failed prediction deceives them and leads them to wrong and potentially harmful decisions. While we did not observe a large number of such explanations in our study, there is, as discussed in §6.2, the risk that this was the result of the specific creation process of our dataset, and may differ for instructions that are less familiar to the model.

6.3.2 Trustworthiness

The largest consensus on where free-text explanations can have a positive impact is that they can improve human-model interaction by increasing the users' trust in the model. *Trust*, often a vague concept (Jacovi and Goldberg, 2021), is here defined as the user's confidence that the model works correctly, be it justified or not. Confidently explaining wrong or ambiguous labels or obviously subjective arguments can harm the trust of users who are not familiar with how the system works and generates explanations. Other than that, human-like properties in the explanations are aligned with the user's expectations (de Graaf and Malle, 2017), and therefore likely to increase trust.

6.3.3 Troubleshooting

By *troubleshooting*, we mean the developer's possibility to debug and improve an LLM with the help of explanations. As Lertvittayakumjorn and Toni (2021) note, explanations can help debug a system, especially where identifiable properties of the training data lead to the bug. For this goal, the unclear relation between prediction and explanation is particularly problematic, therefore properties such as illustrative elements may be less desired. selectivity and subjectivity can also be limiting factors, albeit inevitable in many situations. Even incomplete and subjective explanations can however be useful if the developer observes a consistency in the explanations including or lacking the desired reasoning process. Explaining wrong labels may be a useful feature, too, as it can display the fallacies of the model. As a result, the developer may make targeted modifications to the training data, such as mitigating unwanted statistical cues.

6.3.4 Knowledge Discovery

Explanations can be used for attempts to discover new knowledge. This can again happen in several contexts: a user may want to learn existing knowledge ("the user as a student") or discover novel scientific knowledge ("the user as a researcher"). For the former, factual correctness is crucial, as the learner cannot be expected to be able to judge the reliability of the prediction and explanation themselves, and may be misled by wrong labels or subjective explanations. Selectivity may be misleading in some cases, but simplification more often makes new information more accessible to learners. The situation is different for scientific discovery, as the explanation seeker is likely an expert in the field and able to judge whether to accept a new theory. That the model potentially explains false labels can be misleading but indirectly also be positive, as it may correlate with the likelihood of making new connections.

7 Conclusion

Large language models imitate human explanations in their training data and adopt some of their typical properties. In our analysis of GPT-4 outputs from the Alpaca dataset, selectivity and illustrating factors were particularly common. Subjectivity was less common, as it was probably mitigated in the alignment and filtering process of GPT-4. Misleading explanations were observed rarely, but given that the Alpaca dataset is LLM-generated, it is likely that the observation will not hold for lower-probability inputs.

We discussed the consequences of the presence of these properties and emphasized that it is crucial to consider both the goals and the target groups of the application. For less competent and careful users, there is a risk of shaping false confidence with incomplete, rhetorically convincing but incorrect or biased explanations. However, not all properties that appear undesirable are unequivocally negative: Explanations for false predictions may help developers spot the fallacies of the model. Unfaithful reasoning can make explanations more accessible with simplifications and illustrative examples. Selectivity is often even necessary for generating comprehensible explanations.

Limitations

In §6.2, we discussed the key limitations of our setup and questionnaire. We mentioned that explicitly asking the question if the rater would explain their answer may not reflect if they actually would explain it in a natural setting. We also discussed that the LLM-generated Alpaca dataset is likely to only contain instructions that lead to a correct answer, and thereby have a low risk of a misleading explanation. The generation method of the dataset will also affect the distribution of the other properties. While we selected the dataset for its comparatively broad coverage, the quantitative findings are unlikely to generalise to other domains (in particular to such that are low-resource) and instruction types.

Other LLMs may also exhibit a different distribution due to their pre-training data and instruction-tuning data and setup. A major limitation of this study is the use of outputs from GPT-4, a proprietary model for which there is little confirmed information available to the public. Using an open-source model with openly accessible training data would allow for additional insights for the research community.

We only consider English-language instructions. The generated outputs and explanations probably reflect cultural norms of the English-speaking world. In addition, our three raters were a relatively homogenous group with respect to their demographic and educational background. A more diverse set of raters would be desirable.

Acknowledgments

We thank the annotators for participating in our study, and the anonymous reviewers for their feedback. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and by the European Commission under grant agreement no. 101135671.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. 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 3050–3065, Online. Association for Computational Linguistics.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality challenges in the era of large language models.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Ruth M.J. Byrne. 2023. Good Explanations in Explainable Artificial Intelligence (XAI): Evidence from Human Explanatory Reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6536–6544. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13(2):145–182.

Maartje M. A. de Graaf and Bertram F. Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *2017 AAAI Fall Symposia, Arlington, Virginia, USA, November 9-11, 2017*, pages 19–26. AAAI Press.

Finale Doshi-Velez and Been Kim. 2017. A Roadmap for a Rigorous Science of Interpretability. *CoRR*, abs/1702.08608.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 263–274, New York, NY, USA. Association for Computing Machinery.

Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

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, Online. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *CoRR*, abs/1711.07414. Withdrawn.

Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Frank C. Keil. 2006. Explanation and understanding. *Annual Review of Psychology*, 57:227.

Frank C. Keil, Leonid Rozenblit, and Candice M. Mills. 2004. What lies beneath? understanding the limits of understanding. In Daniel T. Levin, editor, *Thinking and Seeing: Visual Metacognition in Adults and Children*, pages 227–249. MIT Press: Cambridge, MA, USA.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In *Interpretability and Robustness in Audio, Speech, and Language Workshop. Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.*

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. 2023. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. *arXiv preprint arXiv:2309.13638*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 279–288, New York, NY, USA. Association for Computing Machinery.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *CoRR*, abs/1802.00682.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8779–8788.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. *arXiv preprint arXiv:2304.03277*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5):521–562.

Wolfgang Stammer, Felix Friedrich, David Steinmann, Hikaru Shindo, and Kristian Kersting. 2023. Learning by Self-Explaining. *arXiv preprint arXiv:2309.08395*.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, New Orleans, Louisiana, USA. Conference on Neural Information Processing Systems (NeurIPS).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Peter A White. 1995. Use of prior beliefs in the assignment of causal roles: Causal powers versus regularity-based accounts. *Memory & Cognition*, 23:243–254.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert A. Wilson and Frank Keil. 1998. The Shadows and Shallows of Explanation. *Minds and machines*, 8(1):137–159.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *CoRR*, abs/2305.10601.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Xinyan Zhao and V. G. Vinod Vydiswaran. 2020. LIREx: Augmenting Language Inference with Relevant Explanation. *CoRR*, abs/2012.09157.

A Full Questionnaire

The following information and instructions were provided to the annotators in written form.

A.1 Instructions for Annotators

Each annotation sample consists of the following parts:

- *****INSTRUCTION:*****: The prompt that was given to the model.
- *****OUTPUT:*****: The output that the model produces.
- *****QUESTION:*****: The question that you are supposed to answer about the instruction and output.

The following seven questions are asked for each data sample and must be answered with *yes* or *no*. Questions 5-7 only need to be answered if question 1 was answered with *yes*. If the answer was *no*, skip them or click *N/A*.

1. *Does the output contain an explanation for the prediction?* - The model provides a reason(ing path) why it gave a certain answer.
2. *Would you give an explanation/justify your reasoning if you were asked this question by a friend?* - If you were asked this question in a dialogue with someone, would you be likely to add a justification to your answer?
3. *If there is an explanation: Does the explanation refer to commonsense concepts that are taken as given?* - There are parts in the explanation that require world knowledge (facts about the world that are trivial for humans and rarely explicitly stated) to comprehend.
4. *If there is an explanation: Does the explanation list contributing factors?* - The explanation lists reasons that contribute to the truth value of the answer (that may not be complete though).
5. *If there is an explanation: Does the explanation include subjective or biased criteria?* - There are reasons stated in the explanation that are the results of subjective experiences.
6. *If there is an explanation: Does the explanation include illustrative elements (e.g. examples)?* - There are elements in the explanations that do not refer to concrete reasons but

illustrate the reasoning so that the explanation becomes easier to understand.

7. *If there is an explanation: Is the explanation misleading (e.g. arguing for a label that is wrong)?* - The explanation contains elements that are wrong or irrelevant and can therefore mislead the user of the system.

B Examples

In this section, we provide examples for the properties as observed in the Alpaca dataset. We choose short examples where possible, and examples with the highest available agreement between annotators except for section B.1 where we in addition discuss one example where annotators disagree. For the full dataset, we refer to `https://github.com/jekunz/llm-expl-properties.`.

B.1 Commonsense Concepts

The following example from the annotation illustrates the different approaches of the annotators to the idea of commonsense concepts that are taken as given. The large disagreement between annotators in a preliminary study made us remove this concept from the annotation as discussed in Section 3.

INSTRUCTION: *Is the following sentence true or false? – All atoms of the same element have the same mass.*
OUTPUT: *The statement is False. This is because while all atoms of the same element have the same number of protons, they can have different numbers of neutrons. Atoms of the same element with different numbers of neutrons are called isotopes, and they have different masses due to the difference in the number of neutrons.*

Annotators 1 and 3 have annotated this example as containing commonsense factors, as the explanation relies on the implicit common knowledge that neutrons have mass. Annotator 2 however did not think that it requires further explanation as a user reading the output either knows it, or can infer it from the text.

On the other hand, the following example is unequivocally agreed on by all three annotators to contain noteworthy commonsense concepts.

INSTRUCTION: *Does the following sentence contain a run-on? – I walked to the park and I ordered a hotdog.*
OUTPUT: *No, the sentence is not a run-on. It contains two independent clauses separated by a coordinating conjunction 'and', which is an acceptable way of joining two clauses.*

The linguistic terminology in this example (*independent clause* and *coordinating conjunction*, and their relation to the concept of a *run-on*) require further explanation for a broader set of goals: For users who are not experts of linguistics and who want to understand the output as well as for users attempting to understand how robustly the model will be able to solve similar tasks.

B.2 Selectivity (Q3)

For the following example, all three annotators agreed that it contains contributing factors.

INSTRUCTION: *Identify 3 people who have made significant contributions to the field of robotics.*
OUTPUT: *1. Masahiro Mori: Masahiro Mori is a Japanese roboticist and researcher, widely known for his pioneering work in robotics and human-robot interaction. He is most famous for creating the concept of the "uncanny valley,"which describes the discomfort people feel when a robot or animation looks and moves almost, but not exactly, like a natural human being.*
2. Rodney Brooks: Rodney Brooks is an Australian roboticist and computer programmer, recognized for his pioneering work in autonomous robots and artificial intelligence. He is the co-founder of iRobot Corp., which is the maker of the Roomba vacuum cleaning robot. Brooks also served as the director of the MIT Computer Science and Artificial Intelligence Laboratory, and is known for his influential work in behavior-based robotics.
3. Cynthia Breazeal: Cynthia Breazeal is an American roboticist and professor, widely known for her work in social robotics and human-robot interaction. She is the director of the Personal Robots Group at the MIT Media Lab, where she created the world's first social robot, Kismet, in 1997. Breazeal's work focuses on developing robots that can communicate and interact with people in a more human-like way, and she has received numerous awards for her research in this field.

For each person, several contributions and achievements are listed that justify their appearance in this list. It is safe to assume that the mentioned researchers have many more achievements than the ones mentioned, but that a prioritisation has been made.

B.3 Subjectivity (Q4)

The following example contains subjective elements in the explanation, as reported by two of the three annotators.

INSTRUCTION: *What are the main advantages of learning to code?*

OUTPUT: *There are several advantages to learning how to code, including:*

1. Better job opportunities: There's a high demand for programmers and software developers, and the field offers lucrative and stable job opportunities. Additionally, even if you don't pursue a career in tech, many industries value employees with coding skills, as technology becomes more integrated into all aspects of business.

2. Improved problem-solving skills: Learning to code involves breaking big problems down into small, manageable parts, developing an eye for detail and sharpening your logical thinking. These problem-solving skills can be applied to many other areas of your professional and personal life.

3. Increased creativity: Coding allows you to bring your ideas to life, whether you're building a website, an app, or a game. As you learn to code and experiment with different programming languages and tools, you'll find new ways to express your creativity.

4. More control over technology: With coding skills, you can better understand how technology works and make changes to suit your needs. Whether you're automating repetitive tasks, customizing your website, or fixing bugs in a program, coding puts you in control.

5. Personal satisfaction: Finally, learning to code can provide immense personal satisfaction, as you gain the ability to create something from nothing and tackle complex challenges. It's an empowering and rewarding skill to have.

Overall, learning to code can provide valuable job opportunities, improved problem-solving and creativity, greater control over technology, and personal satisfaction. It's a valuable skill that can benefit you in many different ways.

The clearest subjective element is under point 5, where the explanation states that coding is a personally empowering and rewarding skill, without a constraint that this feeling may not apply to all individuals.

B.4 Illustrative Elements (Q5)

For the following example, all three annotators agreed that it contains illustrative elements.

INSTRUCTION: *What verb do you use with the word "inevitable"?*

OUTPUT: *The verb most commonly used with the word "inevitable" is "to be," as in "something is inevitable." Other verbs that can be used with inevitable include "seem", "appear", "become", and "prove". For example, "It seems inevitable that prices will increase" or "It proved inevitable that the company would go bankrupt."*

The examples illustrate common usage, but without a causal explanation *why* the verbs are appropriate (except that they are *common*). It can however be a suitable, practical explanation if the user wants to understand how to embed "inevitable" in a sentence.

B.5 Misleading Explanations for Incorrect Labels

The following example contains a wrong answer and explains it confidently. During the annotation, the mistake was discovered only by one rater, but was agreed upon by the other two when it was indicated.

INSTRUCTION: *Create a regular expression to find all words with three consonants and one vowel.*

OUTPUT: *The following regular expression pattern can be used to find all words with three consonants and one vowel:*

```
\b[^aeiou\s]{3}[aeiou]\b
```

Explanation:

- `\b` - *Matches the word boundary, which indicates the position between a word and a non-word character. It ensures to match the whole word.*
- `[^aeiou\s]` *Within the square brackets, the ^ symbol indicates negation. So, this character set matches any character that is not a vowel or a whitespace character .*
- `{3}` - *Specifies that the preceding character set should be matched exactly three times. Therefore, it matches three consecutive consonants.*
- `[aeiou]` - *This character set matches any single vowel (a, e, i, o or u).*
- `\b` - *Matches the word boundary at the end of the word, to ensure the whole word is matched.*

This regular expression will match words such as "milk", "help", and "desk" but not words like "hello", "world" or "me".

The given regular expression matches only words that consist of three consonants followed by a vowel. It would match none of the examples given in the explanation (*"milk", "help", and "desk"*), therefore it is misleading.

This Reference Does Not Exist: An Exploration of LLM Citation Accuracy and Relevance

Courtni Byun and **Piper Vasicek** and **Kevin Seppi**
Brigham Young University
Provo, USA

Abstract

Citations are a fundamental and indispensable part of research writing. They provide support and lend credibility to research findings. Recent GPT-fueled interest in large language models (LLMs) has shone a spotlight on the capabilities and limitations of these models when generating relevant citations for a document. Recent work has focused largely on title and author accuracy. We underline this effort and expand on it with a preliminary exploration in relevance of model-recommended citations. We define three citation-recommendation tasks. We also collect and annotate a dataset of model-recommended citations for those tasks. We find that GPT-4 largely outperforms earlier models on both author and title accuracy in two markedly different CS venues, but may not recommend references that are more relevant than those recommended by the earlier models. The two venues we compare are CHI and EMNLP. All models appear to perform better at recommending EMNLP papers than CHI papers.

1 Introduction

Citations are a common feature of research writing. They lend credibility to claims and can help identify gaps in prior research. They can also provide a chain of ideas from prior work to a research task.

The last year has seen a drastic increase of interest about large language models (LLMs). ChatGPT (OpenAI, 2022) has opened the eyes of the general public to the potential of LLMs. ChatGPT and its related GPT-X LLMs are being applied to a growing array of tasks (Araoz, 2020; OpenAI, 2023; Byun et al., 2023; Xiao et al., 2023).

One task that has drawn both interest and ire is that of using LLMs to identify citations for a topic. Several recent blog posts and articles have warned of ChatGPT's hallucinated references (Welborn, 2023; Wilkinson, 2023; Neumeister, 2023). We build on recent work to assess the problem.

2 Related Works

Various citation recommendation systems exist, relying on an array of NLP and information retrieval (IR) approaches. Farber and Jatowt (2020) offer a thorough survey of automated citation recommendation approaches.

More recently, use of LLMs has been explored, leading to discussion of the tendency LLMs have to hallucinate output. Day (2023) offered an early exploration of hallucinated references by ChatGPT. They assessed references output by ChatGPT based on accuracy of journal name, volume, issue and page number and found the model incapable of generating any valid references.

On the other hand, MW Wagner (2023) found ChatGPT capable of some accuracy when answering questions about clinical radiological sources.

A letter of warning from McGowan et al. (2023) discussed fabricated references from both ChatGPT and Google's Bard (Manyika, 2023) in psychiatry literature. They found real authors are often included, even when a paper title is fabricated. They also raised the alarm on the possibility of fake references entering into automated indexes.

Gravel et al. (2023) found ChatGPT output in response to medical reference questions was of limited quality, but that references offered by the model were deceptively realistic.

Orduna-Malea and Cabezas-Clavijo (2023) compared ChatGPT and Bard 2.0 citations in English, Spanish, and Italian. They explored reasons for fabricated citations and steps to address the issue.

Taylor et al. (2022) fine-tuned their own LLM, Galactica, and assessed it on three citation generation tasks. They found LLM accuracy for citation generation appears to improve with scale.

Finally, Agrawal et al. (2023) found LLMs tend to hallucinate different authors of fabricated references in multiple independent query sessions, but consistently hallucinate authors in the same session.

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 28–39

They compared accuracy on GPT text-davinci-003, ChatGPT, and GPT-4.

Previous work has primarily focused on metrics related to accuracy of information. While understanding accuracy is important, accurate citations that are irrelevant will still be of little use to researchers. In this work we still assess accuracy, but we also offer a preliminary assessment of the relevance of citations identified by three models.

3 Methods

We define three citation recommendation tasks, intended to model aspects of academic writing that could be supplemented by use of LLMs.

3.1 Models

We compare performance between three GPT-X models: GPT-3 text-davinci-003 (GPT-3), GPT-3.5-turbo (GPT-3.5), and GPT-4. All model hyperparameters used can be found in Appendix A

3.2 Tasks

We define three tasks, each with a unique prompt. The full prompt evolutions and all final prompt designs can be found in Appendices B and C.

3.2.1 Abstract→Citations List Task

This task asks the model generate a list of relevant sources a researcher could explore and incorporate into their paper (target paper). We provide the models with a prompt including a paper title and its accompanying abstract and request the model generate ten relevant citations to be used in the target paper. We request citations in APA format because it is common and having all citations in a consistent format aids in annotating and analysing the data. See Figure 1 for prompt template.

You are an [NLP or HCI] researcher working on a paper to submit to [EMNLP or CHI].
The paper you are working on is titled:
[PAPER TITLE]

The abstract for your paper is:
[PAPER ABSTRACT]

List 10 relevant papers you could cite in your Related Works section. Write each citation in APA format.

Figure 1: Prompt template for Abstract→Citations List task.

3.2.2 Abstract→Related Works Task

The goal of this task is to explore how well the models identify relevant citations when also asked to discuss them, without the textual scaffolding of a provided Related Works section. The prompt for this task builds on the prompt for the first task, but replaces the final section with: *Write a Related Works section for your paper. Include 10 in-text citations. Also include a list of those citations with each citation in APA format.*

3.2.3 Discussion→Supported Discussion Task

The goal of this task is to test model citation recommendation and discussion when some textual scaffolding is provided. The prompt for this task builds on the prompt for the first task by including the target paper title and abstract in the prompt, but the prompt additionally includes a portion of the results discussion. The final section of the prompt, which follows the discussion, is changed to: *Rewrite the Discussion section to include 10 in-text citations. Also include a list of those citations with each citation in APA format.*

3.3 Dataset

We randomly sampled twenty papers from two top-tier, but different venues, CHI (HCI) and EMNLP (NLP). Ten papers were randomly sampled from recent publications of each venue. See Appendix E for the list of papers. The paper title, abstract, and discussion of results were extracted for each paper. For some papers this was taken from the Results section and for others, the Discussion section. Some discussions were too long for the models. For these, we extracted only the first paragraph of each section within the discussion. We also extracted the bibliography from each paper.

This information was used to fill the prompt templates, which were then input to each model. The output was collected and the citations extracted. While we requested citations in APA, the models sometimes used different formatting. We reformatted each citation to ensure it was in APA. Some model-generated citations lacked titles. These we exclude from our final dataset because we cannot verify whether they are real papers. Our final dataset has 1616 annotated citations.

We used Google Scholar to check whether each model-recommended citation was for a real paper.

Abstract → Citations List				**Abstract → Related Works**				**Discussion → Supported Discussion**			
	Title Accuracy				Title Accuracy				Title Accuracy		
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	24.47%	**48.98%**	36.98%	GPT-3	0.00%	**36.84%**	19.44%	GPT-3	**34.88%**	18.37%	26.09%
GPT-3.5	28.00%	**56.00%**	42.00%	GPT-3.5	13.51%	**50.54%**	30.39%	GPT-3.5	12.96%	**51.43%**	22.38%
GPT-4	54.00%	**78.00%**	66.00%	GPT-4	68.87%	**75.45%**	72.22%	GPT-4	**47.15%**	25.74%	37.50%
	Author Precision				Author Precision				Author Precision		
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	**75.00%**	70.29%	71.82%	GPT-3	-	**72.71%**	72.71%	GPT-3	**55.53%**	41.67%	50.33%
GPT-3.5	**81.46%**	76.16%	77.93%	GPT-3.5	62.13%	**73.23%**	70.55%	GPT-3.5	66.86%	**68.94%**	60.03%
GPT-4	88.11%	**89.01%**	88.64%	GPT-4	82.07%	**84.78%**	83.51%	GPT-4	**82.14%**	63.15%	76.26%
	Author Recall			Author Recall				Author Recall			
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	**72.65%**	41.54%	51.62%	GPT-3	-	**70.89%**	70.89%	GPT-3	**37.93%**	33.33%	36.21%
GPT-3.5	**81.46%**	70.88%	73.95%	GPT-3.5	61.73%	**71.02%**	68.77%	GPT-3.5	64.00%	**68.11%**	66.31%
GPT-4	88.11%	**89.01%**	88.64%	GPT-4	82.07%	**84.78%**	83.51%	GPT-4	**82.14%**	63.15%	76.26%
	Year Accuracy				Year Accuracy				Year Accuracy		
	HCI	NLP	Total		HCI	NLP	Total		HCI	NLP	Total
GPT-3	1.43	**0.31**	0.68	GPT-3	-	**1.57**	1.57	GPT-3	2.40	**0.89**	1.83
GPT-3.5	**0.29**	0.46	0.40	GPT-3.5	1.06	**0.55**	0.68	GPT-3.5	6.29	**0.61**	3.09
GPT-4	1.44	**1.26**	1.33	GPT-4	1.59	**0.70**	1.12	GPT-4	**2.59**	3.08	2.74

Table 1: Accuracy scores for each model, for each of the tasks, broken out between HCI and NLP.

Nearly all real papers had an exact match in the first three results of a page, so we restricted our search to the first page of results. Petiska (2023) found that ChatGPT tends to use Google Scholar citation counts when recommending citations, so relying only on Google Scholar results should be sufficient. A citation was marked as fabricated if an exact match was not found in the first page of Google Scholar results. A citation with an exact match was marked as a real paper and the APA citation for the true paper was collected and checked against the citation generated by the model.

We automatically compared information in the citations generated by the models against the information collected from the real papers. We collected information for how many citations were fabricated vs real. We also calculated author precision and recall between the authors in a recommended citation and those on real papers. We tested relevance by checking whether a real paper's title was found in the bibliographies of the target papers and whether the authors of the model-generated citations were found in the bibliographies of the target papers.

While more elaborate metrics for determining citation relevance exist (Belter, 2017; Boyack and Klavans, 2010), these often involve creating a network of citations. The overlap between citations is then checked. This includes overlap with the target papers. However, we needed target papers that were excluded from the models' training data, which meant very recent papers that had not been cited yet. This meant we needed a different metric for relevance. We focus on several basic metrics based on the idea that if there is overlap between papers models recommend and papers authors actually use, then those papers and authors that overlap must be relevant. This means true relevance could be higher, but our strict definition should offer a reasonable exploratory view.

4 Results

Accuracy results can be found in Tables 1 and 3, while relevance results can be found in Table 2.

4.1 Accuracy

Title Accuracy is the percentage of citations recommended by the model that had real paper titles. Author Precision, Author Recall, and Year Accuracy were only calculated for citations of real papers. Year Accuracy was calculated by taking the absolute value of the year a real paper was published, minus the year in the model-recommended citation.

As seen in Table 1, the models tend to perform better on NLP papers, particularly with respect to paper titles. This is reiterated by the results in Table 3, where for nearly every model, for every task there appears to be a significant difference between NLP and HCI papers on this metric.

The distinction is less clear for other metrics. For example both GPT-3 and GPT-3.5 perform better for HCI papers in terms of Author Recall for the Abstract→Citations List task and GPT-4 performs better for HCI papers in terms of Author Precision for the Discussion→Supported Discussion task.

GPT-4 typically outperforms the other models in terms of accuracy, which is unsurprising given the findings of Taylor et al. (2022) that LLM citation accuracy improves with model scale. There are,

Abstract → Citations List			
Title Relevance			
	HCI	NLP	Total
GPT-3	0.17%	**22.92%**	16.90%
GPT-3.5	0.18%	**25.00%**	17.86%
GPT-4	0.20%	**29.49%**	20.45%
Real Author Relevance			
	HCI	NLP	Total
GPT-3	4.35%	**6.25%**	5.63%
GPT-3.5	**7.14%**	5.36%	5.95%
GPT-4	3.70%	**3.85%**	3.79%
False Author Relevance			
	HCI	NLP	Total
GPT-3	**13.04%**	8.33%	9.86%
GPT-3.5	7.14%	**8.93%**	8.33%
GPT-4	16.67%	**25.64%**	21.97%

Abstract → Related Works			
Title Relevance			
	HCI	NLP	Total
GPT-3	-	**10.71%**	10.71%
GPT-3.5	0.24%	**29.79%**	24.19%
GPT-4	0.17%	**27.71%**	17.31%
Real Author Relevance			
	HCI	NLP	Total
GPT-3	-	**3.57%**	3.57%
GPT-3.5	0.00%	**10.64%**	8.06%
GPT-4	4.11%	**6.02%**	5.13%
False Author Relevance			
	HCI	NLP	Total
GPT-3	-	**25.00%**	25.00%
GPT-3.5	13.33%	**17.02%**	16.13%
GPT-4	**28.77%**	19.28%	23.72%

Discussion → Supported Discussion			
Title Relevance			
	HCI	NLP	Total
GPT-3	0.12%	**22.22%**	12.5%
GPT-3.5	0.25%	**33.33%**	25.00%
GPT-4	0.08%	**19.23%**	8.33%
Real Author Relevance			
	HCI	NLP	Total
GPT-3	**6.67%**	0.00%	4.17%
GPT-3.5	0.00%	**11.11%**	6.25%
GPT-4	1.72%	**3.85%**	2.38%
False Author Relevance			
	HCI	NLP	Total
GPT-3	**26.67%**	0.00%	16.67%
GPT-3.5	21.43%	**27.78%**	25.00%
GPT-4	6.9%	**11.54%**	8.33%

Table 2: Relevance scores for each model, for each task, broken out between HCI and NLP papers.

Abstract→Citations List						
Title Accuracy Significance						
	HCI		NLP			
	Mean	SD	Mean	SD	t-statistic	p-value
GPT-3	0.24	0.43	0.49	0.50	-3.62	**0.00**
GPT-3.5	0.28	0.45	0.56	0.50	-4.16	**0.00**
GPT-4	0.54	0.50	0.78	0.41	-3.68	**0.00**
Abstract→Related Works						
Title Accuracy Significance						
	HCI		NLP			
	Mean	SD	Mean	SD	t-statistic	p-value
GPT-3	0.00	0.00	0.37	0.48	-6.25	**0.00**
GPT-3.5	0.14	0.34	0.51	0.50	-6.22	**0.00**
GPT-4	0.69	0.46	0.75	0.43	-1.08	0.28
Discussion→Supported Discussion						
Title Accuracy Significance						
	HCI		NLP			
	Mean	SD	Mean	SD	t-statistic	p-value
GPT-3	0.35	0.48	0.18	0.39	1.81	0.07
GPT-3.5	0.13	0.34	0.51	0.50	-5.13	**0.00**
GPT-4	0.47	0.50	0.26	0.44	3.36	**0.00**

Table 3: Two-sample t-tests for title accuracy on HCI vs NLP papers. Calculated via SciPy and NumPy (Virtanen et al., 2020; Harris et al., 2020).

however, exceptions to this. For example, GPT-3.5 outperforms GPT-4 on Title Accuracy, Author Precision, and Author Recall for the NLP papers on the Discussion→Supported Discussion task.

The models appear to struggle with the Discussion→Supported Discussion task. This could be due to our poor prompt design for this task. CHI papers typically include a separate Discussion section, while EMNLP papers often include a discussion of results with the Results section. We distinctly asked models to support our *Discussion* sections. Future research could explore whether changing *Discussion* to *Results* in the prompt could yield better results for NLP papers.

4.2 Relevance

Title Relevance reports the percentage of real papers cited in the target paper. Real Author Relevance reports the percentage of authors from a model-recommended citation that were real authors on that paper and who had a paper cited in the target paper. False Author Relevance reports the percentage of authors from a model-recommended citation that were not real authors on that paper, but who had papers cited in the target paper.

In terms of relevance, we again see better performance for NLP papers in terms of title relevance. The distinction becomes less clear for other metrics. For example, GPT-4 on the Abstract→Related Works task and False Author Relevance. However, there does not appear to be a large difference between models. In multiple instances the older models perform better than GPT-4, for example GPT-3 for the Abstract→Related Works task on the False Author Relevance metric for NLP papers and both GPT-3 and GPT-3.5 on the Discussion→Supported Discussion task on all relevance metrics.

5 Conclusion

We evaluated GPT-3, GPT-3.5, and GPT-4 on three different citation recommendation tasks and compared them across two research disciplines. We found contrasts in terms of relevance and accuracy between those disciplines. This is important because individuals outside of NLP are beginning to use these models in their research. It is important for researchers from other disciplines to recognize these models' limitations for their disciplines.

Finally, while GPT-4 typically outperforms previous models on accuracy, it does not clearly perform better in terms of relevance.

6 Limitations

While 1616 citations seems like enough for a thorough run of statistical tests, this is not the case. Due to how poorly GPT-3 and GPT-3.5 perform on many of the tasks and how many ways we split the data, several of our sample sizes are slightly under 30, with the smallest being 24. We have run significance tests comparing performance between models and between HCI and NLP papers for other metrics, but considering the small sample sizes of some of the groups, we felt the limited space of this short paper would be best utilized reporting our other results.

Our largest sample sizes are for the Title Accuracy metric because this included all citations, while the other metrics excluded citations for papers that did not exist. This is why we only report significance results for Title Accuracy between HCI and NLP papers. We exclude our significance results for Title Accuracy between models due to the length limitations of this paper. Previous research has shown a difference between models of different sizes. Our results reiterate those findings.

We also did not compare accuracy of other citation information, like page numbers, publication venues, and URLs. Preliminary tests showed much worse model performance on these citation features. We chose to focus on the features the models appeared to recreate more accurately. We leave exploration of these other features to future work.

Additionally, due to the inherently messy nature of text data, some aspects of data collection and curation were done manually. While we did multiple checks at each step of the process to maintain quality, there could still be errors we did not catch.

We also relied on Google Scholar results to determine veracity of citation titles. It is possible that some of the citations marked as fabricated could be real papers that did not show up on the first page of results.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references?

Manuel Araoz. 2020. Openai's gpt-3 may be the biggest thing since bitcoin.

Christopher Belter. 2017. A relevance ranking method for citation-based search results. *Scientometrics*.

Alemitu Bezabih, Kathrin Gerling, Workeabeba Abebe, and Vero Vanden Abeele. 2023. Challenges and opportunities for interactive technology to support parents of hiv-positive children in ethiopia in the disclosure process. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Avinash Bhat, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin LC Guo. 2023. Aspirations and practice of ml model documentation: Moving the needle with nudging and traceability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Kevin Boyack and Richard Klavans. 2010. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61:2389–2404.

Courtni Byun, Piper Vasicek, and Kevin Seppi. 2023. Dispensing with humans in human-computer interaction research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA. Association for Computing Machinery.

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Watrin Patrick, and Thomas François. 2022. Linguistic corpus annotation for automatic text simplification evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1842–1866.

Terence Day. 2023. A preliminary investigation of fake peer-reviewed citations and references generated by chatgpt. *The Professional Geographer*, 0(0):1–4.

Shengda Fan, Shasha Mo, and Jianwei Niu. 2022. Boosting document-level relation extraction by mining and injecting logical rules. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10311–10323.

M. Farber and A. Jatowt. 2020. Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries*, 21:375–405.

Dan Friedman, Alexander Wettig, and Danqi Chen. 2022. Finding dataset shortcuts with grammar induction. *arXiv preprint arXiv:2210.11560*.

Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. 2023. Learning to fake it: Limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew

Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Franziska Herbert, Steffen Becker, Leonie Schaewitz, Jonas Hielscher, Marvin Kowalewski, Angela Sasse, Yasemin Acar, and Markus Dürmuth. 2023. A world full of privacy and security (mis) conceptions? findings of a representative survey in 12 countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Younghoon Jeong, Juhyun Oh, Jaimeen Ahn, Jongwon Lee, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. Kold: korean offensive language dataset. *arXiv preprint arXiv:2205.11315*.

Tianyu Jiang and Ellen Riloff. 2022. Identifying physical object use in sentences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11362–11372.

Shuyang Li, Yufei Li, Jianmo Ni, and Julian McAuley. 2021. Share: a system for hierarchical assistive recipe editing. *arXiv preprint arXiv:2105.08185*.

Chang Liu, Arif Usta, Jian Zhao, and Semih Salihoglu. 2023. Governor: Turning open government data portals into interactive databases. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Zhenghao Liu, Han Zhang, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Xiaohua Li. 2022. Dimension reduction for efficient dense retrieval via conditional autoencoder. *arXiv preprint arXiv:2205.03284*.

James Manyika. 2023. An overview of bard: an early experiment with generative ai.

Alessia McGowan, Yunlai Gui, Matthew Dobbs, Sophia Shuster, Matthew Cotter, Alexandria Selloni, Marianne Goodman, Agrima Srivastava, Guillermo A. Cecchi, and Cheryl M. Corcoran. 2023. Chatgpt and bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research*, 326:115334.

Marie Muehlhaus, Marion Koelle, Artin Saberpour, and Jürgen Steimle. 2023. I need a third arm! eliciting body-based interactions with a wearable robotic arm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Birgit B Ertl-Wagner MW Wagner. 2023. Accuracy of information and references using chatgpt-3 for retrieval of clinical radiological information. *Canadian Association of Radiologists Journal*.

Larry Neumeister. 2023. Lawyers blame chatgpt for tricking them into citing bogus case law. *KSL*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report.

Enrique Orduna-Malea and Alvaro Cabezas-Clavijo. 2023. Chatgpt and the potential growing of ghost bibliographic references. *Scientometrics*, 128.

Eduard Petiska. 2023. Chatgpt cites the most-cited articles and journals, relying solely on google scholar's citation counts. as a result, ai may amplify the matthew effect in environmental science.

Ananditha Raghunath, Laurel Krovetz, Hosea Mpogole, Henry Mulisa, Brian Dillon, and Richard Anderson. 2023. From grasshoppers to secondhand cars: Understanding the smartphone-enabled marketplace in peri-urban tanzania. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Jun Rekimoto. 2023. Wesper: Zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Alia Saad, Kian Izadi, Anam Ahmad Khan, Pascal Knierim, Stefan Schneegass, Florian Alt, and Yomna Abdelrahman. 2023. Hotfoot: Foot-based user identification using thermal imaging. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend. 2022. Topical segmentation of spoken narratives: A test case on holocaust survivor testimonies. *arXiv preprint arXiv:2210.13783*.

Zhongwei Wan, Yichun Yin, Wei Zhang, Jiaxin Shi, Lifeng Shang, Guangyong Chen, Xin Jiang, and Qun Liu. 2022. G-map: general memory-augmented pretrained language model for domain tasks. *arXiv preprint arXiv:2212.03613*.

Aaron Welborn. 2023. Chatgpt and fake citations.

Jordan White, William Odom, Nico Brand, and Ce Zhong. 2023. Memory tracer & memory compass: Investigating personal location histories as a design material for everyday reminiscence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

D Wilkinson. 2023. Be careful... chatgpt appears to be making up academic references.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, page 75–78, New York, NY, USA. Association for Computing Machinery.

Ashley Ge Zhang, Yan Chen, and Steve Oney. 2023. Vizprog: Identifying misunderstandings by visualizing students' coding progress. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Shujian Zhang, Chengyue Gong, and Xingchao Liu. 2022. Passage-mask: A learnable regularization strategy for retriever-reader models. *arXiv preprint arXiv:2211.00915*.

A Hyperparameters

- Temperature: 0.0
- Top P: 1
- Frequency Penalty: 0.5
- Presence Penalty: 0.5
- Maximum Tokens: 2000

We chose a temperature of 0 because, while a temperature of 0 does not guarantee identical output each time, it does increase the likelihood of very similar output. This was the best option available at the time for generating reproducible results. We used 0.5 for both frequency and presence penalties because both GPT-3 and GPT-3.5 are prone to repeating citations when they are set to 0.

B Prompt Engineering

The following are the various prompt evolutions we used before settling on our final prompt designs.

We went through several iterations of prompt design for each of the three tasks in this paper. The prompt variations were primarily focused around the request portion of the prompt. All prompts included either a CHI or EMNLP paper title and abstract. The Results→Supported Results task prompts also included discussion from the same CHI or EMNLP prompt paper.

All of the prompts in this subsection follow the GPT-3 design. The main difference between the GPT-3 and newer model prompts was a change to a first person perspective. We did not ultimately include GPT-3 in our results for the Abstract→Related Works and Results→Supported Results tasks because the final prompt design was too long for the GPT-3 limited context length. However, GPT-3 was included and evaluated on earlier variations of prompts for those tasks. We found GPT-3 was virtually incapable of identifying any citations of real papers for the Abstract→Related Works and Results→Supported Results tasks, even for prompt designs short enough to fit the GPT-3 context.

B.1 Abstract→Citations Prompt Evolution

Our initial prompt design for the Abstract→Citations task used the following format:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> *Five relevant papers you could cite in your related works sections are:*

We found the models have a tendency to cite older sources, so we next adjusted the prompt to request only *recent* citations. We updated the prompt to the following, with the changed portion in bold:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> ***Five relevant papers from the last five years you could cite in your related works sections are:***

We did find the models do often *claim* to cite recent papers using this prompt, but we also noticed they have a tendency to hallucinate paper publication years as more recent than they actually are. We did not, however, do an official comparison between how these prompt designs impact citation year hallucinations. This would be an interesting item for future research.

We ultimately decided to request ten, rather than five citations, to hopefully get a large enough sample size to run statistical tests. We also decided to remove the the request for papers from the last five years because it did not appear to have a strong impact on the results. Finally, we added a request for the model to output the citations in APA format. We found that not requesting a specific format often resulted in the models just choosing a format. The format they chose was sometimes not even a standard format and occasionally the format could change throughout the same output. Our final prompt design was:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> ***List 10 relevant papers you could cite in your Related Works section. Write each citation in APA format.***

B.2 Abstract→Related Works Prompt Evolution

The prompt format for this task is nearly identical to that of the Abstract→Citations task. The main difference is in the final line of the prompt.

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> *The Related Works section of your paper is:*

Again, we realized the models have a tendency to cite older sources, so we updated the prompt to request recent sources. We also followed the same pattern of changing the design to make specific requests, rather than asking the model to continue with writing a related works section. The changed portion of the prompt is in bold.

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> ***Write the related works section for this paper. Discuss 3 sources. Each source must be from the last five years and must include the paper name.***

We decided to allow the model to include a higher number of sources. We updated the prompt to reflect that. We also wanted enough information about each citation to be able to verify it, so we updated the prompt to request the model to include the paper title and a complete list of authors. The prompt design can be found below.

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> ***Write the related works section for this paper. Discuss up to 10 sources. Each source must be from the last five years and must include the paper name and full list of authors.***

We wondered if model performance could be impacted by the difference in citation formatting by asking the model to include a full list of authors and paper title. We updated our prompt design to allow the models to use in-text citations as one normally would (author name, year), but we included a request for the models to include a list of used citations after their prose. The final prompt design can be found below:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*

> ***Write the related works section for this paper. Discuss ten sources. Each source must be from the last five years. Include a list of the citations used following your related works section.***

Again, we found that including a request for recent sources had little impact, so we removed that portion of the prompt. We also found it necessary to request APA formatting. Our final prompt design for this task can be found below:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> ***Write a Related Works section for your paper. Include 10 in-text citations. Also include a list of those citations with each citation in APA format.***

B.3 Results→Supported Results Prompt Evolution

Again, prompts included either a CHI or EMNLP prompt paper title and abstract, but the Results→Supported Results task prompts included discussion from the same CHI or EMNLP prompt paper. Our original prompt design for this task was:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> *The Discussion section for your paper is: [PAPER DISCUSSION]*
>
> *A revised version of your Discussion section including supporting sources is:*

We updated this prompt design to also request recent sources. Additionally, we decided to change to a specific request, rather than having the model simply continue on. The updated prompt can be found below, with the changes in bold.

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> *The Discussion section for your paper is: [PAPER DISCUSSION]*
>
> ***Modify this Discussion section by including supporting sources. Discuss 3 sources. Each source must be from the last five years and must include the paper name.***

We modified the prompt to allow the models to include up to ten sources. We also noted that earlier prompt designs led to output following standard in-text citation formats, in which only the name of the lead author and publication year were included. We updated the prompt to request the complete list of authors and full paper name. We made this change to make verification of sources possible.

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*
>
> *The abstract for your paper is: [PAPER ABSTRACT]*
>
> *The Discussion section for your paper is: [PAPER DISCUSSION]*
>
> ***Write a revised version of this discussion. Include up to 10 supporting sources. Each source must be from the last five years and must include the paper name and full list of authors.***

This prompt design was eventually changed to request the model to include the list of sources following the prose, to allow for a format more similar to the models' training data. The final prompt can be found below:

> *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*
>
> *The paper you are working on is titled: [PAPER TITLE]*

The abstract for your paper is: [PAPER ABSTRACT]

The Discussion section for your paper is: [PAPER DISCUSSION]

Rewrite the Discussion section to include 10 in-text citations. Also include a list of those citations with each citation in APA format.

C Final Prompt Templates for all Models

C.1 Abstract→Citations

The final prompt designs provided to each model for this task can be found below.

You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].

The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

List 10 relevant papers you could cite in your Related Works section. Write each citation in APA format

C.2 Results→Supported Results

The final prompt designs provided to each model for this task can be found below.

C.2.1 GPT-3.5 & GPT-4

You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].

The paper you are working on is titled: [PAPER TITLE]

The abstract for your paper is: [PAPER ABSTRACT]

The Discussion section for your paper is: [PAPER DISCUSSION]

Rewrite the Discussion section to include 10 in-text citations. Also include a list of those citations with each citation in APA format.

C.3 Abstract→Related Works

The final prompt designs provided to each model for this task can be found below.

SYSTEM: *You are an [HCI or NLP] researcher working on a paper to submit to [CHI or EMNLP].*

USER: *The paper you are working on is titled: [PAPER TITLE]*

The abstract for your paper is: [PAPER ABSTRACT]

Write a Related Works section for your paper. Include 10 in-text citations. Also include a list of those citations with each citation in APA format.

D Example Citations

All citations in the following subsections were identified by GPT-X models.

D.1 GPT-4 Citations of Real Papers and Correct Authors

The citations in this section are examples of GPT-4-identified citations. The citation titles and authors are correct, though other information in these citations, like year or publisher, may be hallucinated.

1. Kang, R., Dabbish, L., Fruchter, N., & Kiesler, S. (2015). "My data just goes everywhere: " User mental models of the internet and implications for privacy and security. In Eleventh Symposium On Usable Privacy and Security (SOUPS 2015), pp. 39-52.
2. 10. Wash, R., & Rader, E. (2015). Too much knowledge? Security beliefs and protective behaviors among United States internet users. In Eleventh Symposium On Usable Privacy and Security (SOUPS 2015), pp. 309-325.
3. 1. Aker, J. C., & Mbiti, I. M. (2020). Mobile Phones and Economic Development in Africa. Journal of Economic Perspectives, 34(3), 207-232.

D.2 GPT-4 Citations of Real Papers and Incorrect Authors

The citations in this section are examples of GPT-4-identified citations. The citation titles are correct, though other information in these citations, like year or publisher, may be hallucinated. At least a portion of one author in each citation is hallucinated. In the section citation, the second author should be *Kim, T. H.* The models identified the third citation in several of our tests, but it would occasionally swap out *Natural Language Generation* for *Natural Language Inference*.

1. Abawajy, J., & Hassan, M. M. (2017). User preference of cyber security awareness delivery methods. Behaviour & Information Technology, 36(2), 133-144.
2. Das, S., Kim, D. W., & Dabbish, L. A. (2019). The effect of social influence on security sensitivity. In Proceedings of the Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018), pp. 27-49.
3. Zhang, X., Kedzie, C., & McKeown, K. (2019). Evaluating the Evaluation of Diversity in Natural Language Generation. 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 6007-6013.

D.3 GPT-4 Citations of Papers with Colons

The citations in this section are examples of GPT-4-identified citations. While portions of these citation titles are correct, they still include hallucinations, often following the colon.

1. Blikstein, P. (2016). Snap! (Build Your Own Blocks): An introduction. In Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education (pp. 358-358).
2. Lee, M. J., Bahmani, F., Kwan, I., & Ko, A. J. (2018). Gidget: A debugging game for learning programming concepts. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).
3. "Trust in the Sharing Economy: An Experimental Study on the Role of Reputation Systems in Collaborative Consumption" (Ert et al., 2016)

D.4 GPT-4 Hallucinated Citations

1. Das, S., Kim, H., Kelley, P. G., & Cranor, L. F. (2018). Making Security Memorable: Designing a Behavior Change Story for Secure Communication. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).
2. Alkaldi, N., Renaud, K., & Hutchinson, W. (2017). To Share or Not to Share? A Cross-Cultural Study of Security and Privacy Perceptions. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 1-12).
3. 4. Boateng, R., Mbrokoh, A. S., Boateng, L., Senyo, P. K., & Ansong, E. (2018). Determinants of e-commerce adoption among Ghanaian SMEs. Journal of Small Business and Enterprise Development, 25(3), 501-520.

D.5 GPT-3.5 Citations

The citations in this section are examples of GPT-3.5-identified citations.

1. Acquisti, A., Brandimarte, L., & Loewenstein, G. (2017). Privacy and human behavior in the age of information. Science, 347(6221), 509-514.
2. Sambasivan, M., & Soon, Y. W. (2019). Mobile payment adoption in Malaysia: An empirical analysis. Journal of Retailing and Consumer Services, 47, 221-231.
3. Wang, Y., & Li, Y. (2017). CodeMend: Assisting Interactive Programming with Bimodal Embedding. In Proceedings of the 2017 ACM Conference on International Computing Education Research (pp. 1-9).

D.6 GPT-3 Citations

The citations in this section are examples of GPT-3-identified citations.

1. Waseem, Zeerak, and Dirk Hovy. "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016).

2016

2. Xu, P., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 890-900).
3. Kelleher, C., Pane, J. F., & Bunge, C. (2015). Supporting novice programmers: A review of empirical studies on learning and teaching introductory programming. ACM Computing Surveys (CSUR), 47(4), 63.

E Papers Used

The HCI papers were: (Herbert et al., 2023; Bhat et al., 2023; Bezabih et al., 2023; Raghunath et al., 2023; Liu et al., 2023; Saad et al., 2023; Muehlhaus et al., 2023; White et al., 2023; Zhang et al., 2023; Rekimoto, 2023)

The NLP papers were: (Fan et al., 2022; Liu et al., 2022; Friedman et al., 2022; Wan et al., 2022; Jiang and Riloff, 2022; Jeong et al., 2022; Cardon et al., 2022; Zhang et al., 2022; Li et al., 2021; Wagner et al., 2022).

Combining Multiple Metrics for Evaluating Retrieval-Augmented Conversations

Jason Ingyu Choi **Marcus D. Collins** **Eugene Agichtein**
Oleg Rokhlenko and **Shervin Malmasi**
Amazon.com, Inc. Seattle, WA, USA
{chojson,collmr,eugeneag,olegro,malmasi}@amazon.com

Abstract

Conversational AI is a subtype of Human-Computer Interaction that has gained wide adoption. These systems are typically powered by Large Language Models (LLMs) that use Retrieval Augmented Generation (RAG) to infuse external knowledge, which is effective against issues like hallucination. However, automatically evaluating retrieval augmented conversations with minimal human effort remains challenging, particularly in online settings. We address this challenge by proposing a lexical metric, and a novel method for combining it with other metrics, including semantic models. Our approach involves: (1) Conversational Information Utility (CIU), a new automated metric inspired by prior user studies on web search evaluation, to compute information overlap between conversation context and grounded information in an unsupervised, purely lexical way; and (2) a generalized reward model through Mixture-of-Experts (MoE-CIU) that dynamically ensembles CIU with other metrics, including learned ones, into a single reward. Evaluation against human ratings on two public datasets (Topical Chat and Persona Chat) shows that CIU improves correlation against human judgments by 2.0% and 0.9% respectively compared to the second best metric. When MoE is applied to combine lexical and learned semantic metrics, correlations further improve by 9.9% and 5.0%, suggesting that unified reward models are a promising approach.

1 Introduction

Conversational AI is a specific type of Human-Computer Interaction that has been widely studied in recent years (Ouyang et al., 2022; Team et al., 2023), leading to the development of multi-purpose chat assistants (*e.g.* ChatGPT, Claude) based on Large Language Models (LLMs). However, as more customers interact with such assistants, addressing limitations like hallucination, factual consistency, prompt brittleness and controllability has gained more attention (Kaddour et al., 2023). One widely-adopted solution is Retrieval Augmented Generation (RAG), which allows choosing a context document ($d_{context}$) to ground LLM responses, and increase truthfulness with respect to the source document (Lewis et al., 2020).

Our work focuses on the task of automatically assessing the quality of retrieval-augmented responses in knowledge-grounded conversations. By examining both the context and the response, we estimate the degree to which the retrieved document was used in generation, in order to identify uninformative or inconsistent responses. Our approach is designed for real-time use, where using a large model may be infeasible. Compared to offline tasks, online evaluation (e.g., live monitoring of defects) requires efficient solutions. Recent work utilizes LLMs, either through prompt engineering or fine-tuning, to automatically predict evaluation metrics and reduce dependency from human annotators (Thapa et al., 2023; Chan et al., 2023). Despite demonstrated potentials, a large number of parameters, high latency, and potential legal issues significantly limits deploying LLM-based solutions for live traffic monitoring. As an alternative, we propose an approach that combines much simpler and scalable metrics to predict user ratings, or potentially other business metrics. Our approach can also support offline evaluations, and is relevant to recent trends in Reinforcement Learning from Human Feedback (RLHF), which aligns LLM responses toward human preferences (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2024).

Early attempts on automatic dialog evaluation relied on existing metrics (*e.g.* BLEU) from machine translation literature to evaluate assistant conversations against 'gold' conversations (Papineni et al., 2002). However, defining the full space of 'gold' conversations is infeasible due to the non-deterministic nature of dialogs and many existing works simply penalize any response that slightly de-

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 40–50
June 21, 2024 ©2024 Association for Computational Linguistics

viates from 'gold' (Chen et al., 2019). On the other hand, learned, semantic conversational quality metrics trained on labeled data tend to show higher correlations against human judgements than exact word-overlap metrics, because word embedding-based approaches can compute overlap in a 'soft' way that accounts for lexical variation, *e.g.* , (Sellam et al., 2020; Lowe et al., 2017a). However, such 'soft' approaches also suffer from different issues, such as over-fitting, performance degradation on longer inputs, or learning similar representations for antonyms.

To address these gaps, we ask: *is it feasible to unify multiple independent metrics into a single reward?* To answer this question, we investigated two research questions:

Q1 What is the most effective and robust standalone metric (whether lexical or learned) that aligns with human ratings from open-source knowledge-grounded conversations?

Q2 Given a set of independent metrics, how much improvement does a unified reward model gain compared to the best standalone metric?

For **Q1**, previous studies on 'exact' word overlap metrics showed they correlate poorly with human preferences, such as question answering accuracy (Chen et al., 2019) and response appropriateness (Lowe et al., 2017b). To address this, we compare our newly proposed lexical metric, **Conversational Information Utility (CIU)**, which is inspired from user-centric studies on web search evaluation (Azzopardi et al., 2018; Moffat et al., 2013) against existing metrics. A key insight is that for the user to gain useful information, they must ask a series of questions, or make statements, that cause the conversational system (or human partner) to respond with information overlapping with $d_{context}$. Our main novelty is how CIU quantifies information overlap to reward relevancy, information novelty and conciseness while penalizing repetitive information and high user effort. Experiments validate that CIU improves correlation against human ratings by 2.0% and 0.9% against the second best metric on Topical Chat (Gopalakrishnan et al., 2019) and Persona Chat (Zhang et al., 2018) datasets for predicting Overall Ratings.

For **Q2**, we experiment with different ensemble learning strategies to (1) validate whether previously identified strong metrics are considered as strong predictors (metrics); (2) demonstrate the superiority of an unified reward compared to any standalone metric. Experimental results in feature selection ratio show that CIU is selected 76.4% across 17 different feature selection approaches, which justifies our findings on **Q1**. When Mixture-of-Experts (MoE) (Masoudnia and Ebrahimpour, 2014) was applied, the resulting MoE-CIU model further improved correlation with human ratings by 9.9% and 5.0% on Topical Chat and Persona Chat compared to the best standalone metric. In summary, our contributions are:

- A simple and effective lexical metric for estimating Conversational Information Utility (CIU) within information-seeking retrieval augmented conversations
- A generalized, domain-agnostic model MoE-CIU that utilizes Mixture of Experts to dynamically adjust metric weights of different modalities into an unified reward signal

2 Related Work

Web Search Evaluation and Utility For search engine evaluation, evaluation measures evolved from precision- and recall-based to utility- and cost-based with more emphasis on interactions between users and search results (Moffat et al., 2013). This is because simply measuring how well search engine ranks relevant documents does not always translate to increased user satisfaction. To model interactions, additional information such as likelihood of user continuing or stopping after at a given rank or estimated effort to read each document (Zhang et al., 2017; Sakai and Dou, 2013) is considered when defining a utility (Wicaksono and Moffat, 2020). Overall, web-search utility is an aggregated metric that combines precision and recall of ranked documents with user interaction signals derived from search logs.

However, the main challenge is on applying these intuitions to multi-turn conversations. In conversational settings, many existing word-overlap and learned metrics (Papineni et al., 2002; Tao et al., 2017; Zhang et al., 2019) still rely on word overlap or semantic similarity to evaluate responses while neglecting potential user interactions. An ideal utility should holistically consider word-level precision, semantic relatedness, novelty of information, repetition, conversational history and user effort to evaluate conversations.

Learned Metrics One popular approach is to utilize pretrained contextual embeddings from Transformer models to compute a similarity score between two texts. For example, BERTScore (Zhang et al., 2019) computes a token-level similarity matrix and re-weights the scores based on IDF scores to boost signals from more novel matches. ADEM (Lowe et al., 2017b) uses a hierarchical RNN encoder to predict human-annotated ratings on Twitter data. While ADEM requires human judgments, RUBER-BERT (Ghazarian et al., 2019; Tao et al., 2017) uses an unsupervised negative sampling strategy to train a model that measures information relatedness between query and response. USR (Mehri and Eskenazi, 2020) is another transformer-based model that is shown effective for evaluating model generated responses. To evaluate USR, the authors sampled a small number of conversations from Topical Chat and Persona Chat datasets to annotate several useful ratings (e.g. overall rating of responses), which we adopt for our evaluations.

Mixture-of-Experts Models When there are multiple representations of the same input, *e.g.* complementary representations computed by different expert modules, the Mixture-of-Experts (MoE) approach (Masoudnia and Ebrahimpour, 2014) can take these independent knowledge sources and conditionally combine them into a joint representation (Shazeer et al., 2017). It does so by training a gating mechanism that dynamically assigns weights to the experts, depending on the input (Jain et al., 2019). MoE has been shown to be effective in various settings such as combining Support Vector Machines (Collobert et al., 2002), hierarchical networks (Yao et al., 2009) and Named Entity Recognition (Meng et al., 2021). For our use case, each expert is a representation of different metrics. By training a gating network to dynamically weight experts per instance, we expect MoE to improve over heuristic-based feature combination strategies, such as sum or mean of different metrics (Ghazarian et al., 2019).

3 Proposed Metrics and Models

We define our task and usefulness ratings, followed by details on CIU and MoE-CIU approaches.

3.1 Usefulness Rating Prediction Task

Given a conversation history (C) and a specific utterance at turn i (utt_i), our task is to predict how much useful information (Rosset et al., 2020) is present at utt_i, with respect to the retrieved knowledge ($d_{context}$). Usefulness ratings (1.0 - 5.0) measures whether the response helps towards fulfilling the information needs (*i.e.* learning new information, or asking questions about products in online shopping). To be useful, utterances should meet the information needs of users and drive the conversation forward to elicit more interaction, while staying relevant to C and $d_{context}$.

3.2 CIU: Conversational Information Utility

Utility is defined as the fulfillment a user receives after search, and attempts to model how users aim to gain optimal overall satisfaction (Machmouchi et al., 2017). We hypothesize that successful information-seeking conversations deliver useful and factually correct information from $d_{context}$. CIU is specifically designed to rank responses with respect to salient information overlap in an unsupervised, lexical way. To measure information overlap, we utilize Rapid Automatic Keyword Extraction (RAKE) (Rose et al., 2010), an existing algorithm that extracts and ranks keywords based on word co-occurrences. Highly ranked phrases from $d_{context}$ are then matched against tokens from each turn, combined with multiple token-level discounting criteria, which are discussed in next sections.

We define information overlap as the sum of all token-level match $score$ between utt_i and each sentence d_{sent} from $d_{context}$. In case $d_{context}$ is long and contains many paragraphs irrelevant to utt_i, we limit d_{sent} to only those from the most relevant paragraph if such annotations are available. However for other datasets without annotations, we apply RAKE to $d_{context}$ to extract highly relevant phrases. For simplicity, we use the same notation d_{sent} for extracted phrases.

CIU is a normalized, discounted information overlap. For each turn (i), it is calculated as:

$$CIU_i = [\sum_{token}^{d_{sent}} \frac{score(token, utt_i) \cdot \gamma}{freq(token)}] - E_i. \tag{1}$$

CIU accepts any scoring function $score$ which outputs a relevance score between each token in d_{sent} and utt_i. Here, we use a binary function that outputs 1 if each token from d_{sent} appears in utt_i and 0 otherwise. Although binary scoring seems rudimentary, preliminary experiments showed that embedding-based token similarity scores were noisy and did not generalize well across diverse samples. Another reason to favor binary scores is that CIU is meant to be purely lexical and efficient; embedding-based similarity can significantly slow down predictions as token-level comparisons are expensive. Eq. 1 also has several discounting terms: γ is position-based, $freq$ is word frequency-based, and E_i is an effort discounting term that subtracts time required to read utt_i. Next, we describe the discounting terms.

Position-based discounting Each $score$ is discounted based on which position token appears in utt_i. For example, we boost weights of $score$ that appears earlier in utt_i. This was inspired from earlier work (Sakai and Dou, 2013), which claims the value of relevant information decays based on how much user effort is required to process information. This is particularly effective for longer utt_i since users prefer useful information to appear earlier than later. We adopt the linear discounting proposed from same earlier work:

$$\gamma = max(0, 1 - pos(token)/|utt_i|), \quad (2)$$

where $pos(token)$ is the token index, and $|utt_i|$ is the number of tokens in utt_i.

Frequency-based discounting Without frequency discounting, all tokens are treated equally regardless of how frequent or novel they are. Prior work (Qi et al., 2020) computes informativeness as how many unseen tokens from information units overlap with an answer, measured with a unigram precision function. However, this assumes all repetitive information is irrelevant. Ideally, our utility function should assign smaller weights to frequently observed tokens and higher weights to novel tokens. The simplest way of achieving this is to divide each token $score$ by token's term frequency (TF), which is measured and updated throughout the conversation.

User effort and cost Several methods of evaluating search engines have considered the trade-off between user effort (E) and relevance gain (Zhang et al., 2017; Azzopardi et al., 2018). In conversational settings, we hypothesize that turn-level effort (E_i) can be approximated by computing the total time a user has spent each turn to read a response. To understand how E_i influences user satisfaction, we analyzed turn-level human annotations from Topical Chat and Persona Chat corpus. According to Figure 1, we first observed that users are less likely to rate longer utterances as useful than shorter utterances. To quantify this relationship, we computed Spearman correlation between usefulness ratings and character length. There is a statistically significant negative correlation of -0.203 ($p < 0.001$), justifying the need for a length-based effort discounting.

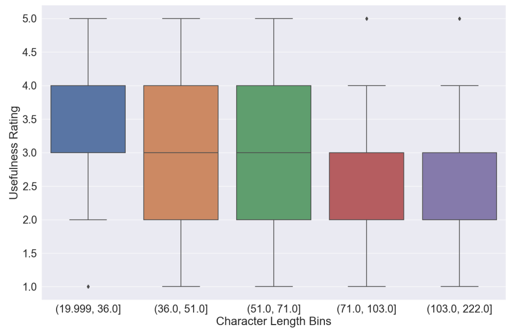


Figure 1: A box plot showing the distribution of utterance length, divided into five equal-sized bins, over human-annotated usefulness ratings from Topical Chat and Persona Chat.

The simplest way of penalizing longer utterances is assuming a constant cost (reading speed) per character, C_{char}.

$$E_i = C_{char} * |utt_i| \quad (3)$$

Then, E required per turn is subtracted from CIU_i to favor shorter turns with identical information as shown in Equation 1. Note that this value can be tuned for different datasets if human annotations are available. Otherwise, we propose to reuse our value 0.005, which was tuned against open-source dialogs with human annotations.

3.3 MoE-CIU: Mixture-of-Experts with CIU

Next, we focus on methodologies to model an unified reward from independent metrics. Perhaps the simplest way of leveraging multiple metrics is computing the average or sum. However, previous work (Ghazarian et al., 2019) highlighted that simple arithmetic operations on raw metric values degrades performance since each metric captures orthogonal measures of relatedness or utility in different scales. Hence, we propose an MoE approach to dynamically normalize and combine the metrics with weights that vary over the input.

MoE requires vectors as inputs rather than scalar scores. Instead of using raw metric scores, we categorize metric scores to one of N bins based on each score distribution. For bin size, we found that $N = 5$ gives the best performance. Then, we create trainable embedding layers of dimensionality (M, N, D) where M is the number of metrics to combine, N is the number of bins, and D is a hyperparameter that defines the embedding size. For our experiments, we used $(6, 5, 16)$.

Two different feature combination strategies were explored. First, we trained a baseline MoE-Concat that uses a concatenated vector of label representations to predict scores. The second model MoE-CIU uses the MoE gating mechanism (Meng et al., 2021) to learn input-dependent weights for each metric to compute a final score. Both approaches can be trained as a binary classifier (with cross-entropy loss), or as a regressor (with mean-squared error loss) depending on use cases.

4 Public Datasets with Human Ratings

We used two publicly available dialog corpora to evaluate our approach.

Topical Chat Topical Chat (Gopalakrishnan et al., 2019), contains 11k conversations that are grounded to web articles ($d_{context}$). For human ratings, we use a subset of more recent and reliable ratings obtained on the official test split (Mehri and Eskenazi, 2020), which contains 360 samples. From multiple available ratings, we selected two ratings that are most relevant for our scope:

- Overall Rating (0 - 5): What is the overall impression of this utterance based on understandability, naturalness, coherency, interestingness, and relevancy of used knowledge?
- Uses Knowledge (0 - 1): Given the fact that the response is conditioned on ($d_{context}$), how well does the response use that fact?

Persona Chat Persona chat (Zhang et al., 2018) is another popular dataset that contains knowledge-grounded conversations on different personas ($d_{context}$). Similar to Topical Chat, Mehri and Eskenazi (2020) released a more recent human annotations on 300 test samples. For consistency, we will choose the exact same type of ratings for our evaluation.

5 Experimental Settings

We first present an overview of our selected baseline metrics and evaluation criteria. For **Q1**, each metric performance is (1) evaluated independently. Then for **Q2**, we experiment with two settings for metric unification: (2) static combination where optimal features (metrics) are pre-selected from existing feature selection algorithms; (3) dynamic combination that utilizes MoE to automatically learn and weight all input features.

5.1 Standalone Metrics Evaluation

We compare CIU performances against standard word-overlap and learned baseline metrics (1) Random Baseline (2) BLEU; (3) METEOR; (4) ROUGE-L; (5) RUBER-BERT; (6) BERTScore (Banerjee and Lavie, 2005; Lin, 2004). Random baseline is added to highlight the relative difficulty of different tasks.

To evaluate the effectiveness of individual metrics, we compute Spearman rank correlation between our metric predictions and two different types of human ratings: (1) 'Overall' ratings; (2) 'Uses Knowledge' ratings, as discussed in Section (§4). Spearman correlation was chosen over Pearson because Spearman is more suited for benchmarking monotonic relationship while Pearson only models linear relationships.

Ablation Analysis on CIU parameters To evaluate individual contributions of different discounting terms within CIU, we include an ablation analysis that systematically removes each discounting terms on Table 1 and Table 3. Effort terms were tuned on the Topical Chat training corpus, and used $C_{char} = 0.005$ for other experiments.

5.2 Unified Reward Evaluation

Static Combination with Feature Selection We experiment with existing feature selection strategies to first identify strong predictors, and second train a model to ensemble strong estimators for predicting human ratings. The evaluation criteria we adopt is the feature selection ratio, which computes how many times each metric is identified as a top-k predictor against others. We experimented with following feature selection[1] strategies:

- Univariate feature selection
- Feature selection using random forest
- Recursive feature elimination
- Forward & backward feature selection
- No feature selection, uses all features

[1] https://scikit-learn.org/stable/modules/feature_selection.html

Since training ensemble models requires labels, we reserved 50 random samples each from both datasets for testing and remainders for training. Support Vector Regressor (SVR) was chosen because this model achieved the strongest performance on multiple experiments over other choices (e.g., gradient-boosted decision tree). The optimal hyperparameters were identified using Grid Search. Since we have only 50 test samples, experiments were repeated 15 times using different sampling seeds and performances were averaged to reduce variance. For consistency, we also report Spearman correlation against human ratings.

Dynamic Combination with Mixture of Experts We compare the static feature selection models to MoE-Concat and MoE-CIU (discussed in Section 3.3), which does not require any feature selection in theory because Mixture of Experts are designed to automatically learn and combine different metric representations. Hence by default, these models take all features as inputs. Performances are also averaged over 15 different sampling seeds.

6 Main Results

We present the results on each public dataset, followed by ablation study and error analysis.

6.1 Topical Chat Results

Standalone Metric Performance Table 1 lists the correlation of different metrics in a standalone setting. Overall, the best CIU configuration that uses all proposed discounting terms achieved the highest correlation for predicting Overall Ratings, which answers **Q1**. It is impressive that CIU was able to outperform learned metrics without any training data. However for predicting Uses Knowledge, a learned metric (RUBER-BERT) outperformed CIU by 3.6%. All of correlation coefficients reported in Table 1, including the difference between CIU and the second best lexical metric (METEOR) are statistically significant ($p < 0.001$).

Unified Reward Performance Table 2 shows the results from different feature selection strategies. The best combination strategy for predicting Overall Ratings was to simply use all metrics. This achieved 0.432 Spearman correlation, a +1.7% improvement over the best standalone metric, CIU. All correlations reported in Table 2 are statistically significant ($p = 0.001$). For predicting Uses Knowledge, the best 'Recursive-5' model excluded ROUGE as the weakest feature, achieving 0.781 correlation. Generally, there is a clear trend that correlation improves with more features. This is a strong evidence showing that leveraging multiple metrics is more effective than any single metric alone.

Metric	Overall Ratings	Uses Knowledge
Random Guessing	0.016	0.023
BLEU	0.298	0.631
METEOR	0.352	0.716
ROUGE-L	0.339	0.688
RUBER-BERT	0.385	**0.778**
BERTScore	0.395	0.717
CIU - $freq$	0.411	0.728
CIU - pos	0.412	0.729
CIU	**0.415**	0.742

Table 1: Spearman correlation between metric predictions and human ratings on Topical Chat. Ablation study is indicated with minus sign where $freq$ stands for frequency and pos for position.

For all 17 different feature selection strategies we note that CIU, RUBER-BERT and BERTScore were almost always selected. They were also the top metrics on Topical Chat (Table 1). For backward selection (which outperform forward selection), we see that 'Backward-1' first picks up RUBER-BERT as the most useful feature, followed by CIU and BERTScore. Although CIU was best in predicting Overall Ratings, other feature selection strategies did not always prioritize CIU on first iterations. Nonetheless, these results demonstrate that the majority of feature selection strategies consider CIU and RUBER-BERT as one of the strongest features, which justifies our findings on **Q1**.

Having validating the effectiveness of combining multiple metrics, we trained the MoE-Concat and MoE-CIU models on the same data splits. To answer **Q2**, MoE-CIU achieved 0.514 correlation for Overall Ratings (+8.2% improvement), and 0.799 correlation for Uses Knowledge (+1.8% improvement) against the best static combination approach, both of which are statistically significant. Accordingly, we claim that MoE-based approaches are superior to traditional feature selection strategies as the MoE gating mechanism can dynamically adjust the weights of different metrics while feature selection is binary and static (features are either used or not, and have a fixed weight).

Metric	Overall Ratings	Uses Knowledge	BLEU	METEOR	ROUGE-L	CIU	RUBER-BERT	BERTScore
Univariate	0.427	0.773	✓	✓	✓	✓	-	✓
Random Forest	0.405	0.721	✓	✓	-	✓	✓	✓
Recursive-5	0.385	0.734	-	✓	-	-	-	-
Recursive-4	0.392	0.733	-	✓	-	-	-	✓
Recursive-3	0.396	0.767	-	✓	-	-	✓	✓
Recursive-2	0.411	0.781	-	✓	-	✓	✓	✓
Recursive-1	0.422	**0.781**	✓	✓	-	✓	✓	✓
Forward-1	0.361	0.729	-	-	✓	-	-	-
Forward-2	0.351	0.742	-	-	✓	✓	-	-
Forward-3	0.392	0.747	-	-	✓	✓	✓	-
Forward-4	0.418	0.757	-	-	✓	✓	✓	✓
Forward-5	0.417	0.771	-	✓	✓	✓	✓	✓
Backward-1	0.389	0.735	-	-	-	-	✓	-
Backward-2	0.386	0.736	-	-	-	✓	✓	-
Backward-3	0.403	0.752	-	-	-	✓	✓	✓
Backward-4	0.413	0.761	-	-	✓	✓	✓	✓
Backward-5	0.425	0.771	✓	-	✓	✓	✓	✓
All	**0.432**	0.778	✓	✓	✓	✓	✓	✓
Selection Ratio	-	-	5 (29.4%)	9 (52.9%)	9 (52.9%)	**13 (76.4%)**	**13 (76.4%)**	**12 (70.5%)**
MoE-Concat	0.507	0.788	✓	✓	✓	✓	✓	✓
MoE-CIU	**0.514**	**0.799**	✓	✓	✓	✓	✓	✓

Table 2: Spearman correlation between model prediction with feature selection and human ratings on Topical Chat. Selection ratio indicates how many times each feature was selected by different feature selection algorithms.

6.2 Persona Chat Results

Standalone Metric Performance Table 3 lists individual metric performance on Persona Chat. For Overall Ratings, CIU again showed the strongest correlation of 0.481 and for Uses Knowledge, RUBER-BERT achieved 0.688. Although the top metrics are identical to Table 1, the remaining metrics not only performed worse, but also fluctuated. For Topical Chat, we observed that BLEU was the least effective in predicting both Overall Ratings and Uses Knowledge. However in Persona Chat, BLEU outperforms METEOR and is comparable to CIU. BERTScore also has poor generalization as correlation dropped by 15.4% and 29.2% compared to Topical Chat. All of the correlations reported in Table 2 are statistically significant ($p < 0.001$).

Metric	Overall Ratings	Uses Knowledge
Random Guessing	0.011	0.017
BLEU	0.472	0.515
METEOR	0.223	0.379
ROUGE-L	0.202	0.387
RUBER-BERT	0.435	**0.688**
BERTScore	0.241	0.486
CIU - $freq$	0.461	0.667
CIU - pos	0.461	0.669
CIU	**0.481**	0.685

Table 3: Spearman correlation between metric predictions and human ratings on Persona Chat. Ablation study is indicated with minus sign where $freq$ stands for frequency and pos for position.

These findings show existing metrics have high variance across tasks. This is true for both lexical and learned metrics as BLEU, METEOR, ROUGE and BERTScore all suffered from significant performance drops. We believe that it is difficult to determine which metric works best ahead of time; nonetheless, CIU is consistently strong and reliable across both domains.

Unified Reward Performance According to Table 4 on predicting Overall Ratings, univariate feature selection combining five metrics excluding ROUGE performed best and achieved 0.529 correlation, a +4.8% improvement compared to best standalone metric CIU. Similarly for predicting Uses Knowledge, SVR using all features achieved the strongest correlation of 0.718, a +3.0% improvement over RUBER-BERT in Table 3. All correlations reported in Table 4 are statistically significant ($p < 0.001$). Overall, it is clear that benefits of combining different metrics generalize to different domains.

For feature selection ratios, we observed BLEU and RUBER-BERT were each selected 82.3% from 17 different feature selection strategies. While CIU was one of the most selected features in Topical Chat, CIU is the third best in Persona Chat with 76.4% selection ratio. Although BLEU were selected the most in Persona Chat, these performances do not carry over to Topical Chat since BLEU was only selected 5 times according to Table 2. Across both datasets, RUBER-BERT was selected most with 75.0% and CIU was second with 72.2%. All of these findings validate that CIU is the strongest and most reliable lexical metric in evaluating retrieval augmented conversations without any training.

Lastly, our proposed MoE-CIU outperformed the strongest feature selection baseline by 1.5% on predicting Uses Knowledge, but only a tiny increase on Overall Ratings. We suspect that with more

Metric	Overall Ratings	Uses Knowledge	BLEU	METEOR	ROUGE-L	CIU	RUBER-BERT	BERTScore
Univariate	**0.529**	0.683	✓	✓	-	✓	✓	✓
Random Forest	0.492	0.636	✓	-	✓	✓	✓	✓
Recursive-5	0.436	0.675	-	-	-	✓	-	-
Recursive-4	0.487	0.675	✓	-	-	✓	-	-
Recursive-3	0.483	0.666	✓	-	-	✓	✓	-
Recursive-2	0.505	0.711	✓	-	✓	✓	✓	-
Recursive-1	0.514	0.713	✓	-	✓	✓	✓	✓
Forward-1	0.381	0.674	-	-	-	✓	-	-
Forward-2	0.455	0.678	-	-	-	✓	✓	-
Forward-3	0.511	0.658	✓	-	-	✓	✓	-
Forward-4	0.504	0.701	✓	✓	-	✓	-	✓
Forward-5	0.507	0.702	✓	-	✓	✓	✓	✓
Backward-1	0.377	0.659	-	-	-	-	✓	-
Backward-2	0.446	0.654	✓	-	-	-	✓	-
Backward-3	0.501	0.645	✓	✓	-	-	✓	-
Backward-4	0.518	0.699	✓	✓	-	-	✓	✓
Backward-5	0.521	0.699	✓	✓	✓	-	✓	✓
All	0.519	**0.718**	✓	✓	✓	✓	✓	✓
Selection Ratio	-	-	**14 (82.3%)**	6 (35.2%)	6 (35.2%)	**13 (76.4%)**	**14 (82.3%)**	8 (47.0%)
MoE-Concat	0.521	0.721	✓	✓	✓	✓	✓	✓
MoE-CIU	**0.531**	**0.733**	✓	✓	✓	✓	✓	✓

Table 4: Spearman correlation between model predictions with feature selection and human ratings on Persona Chat. Selection ratio indicates how many times each feature was selected by different feature selection algorithms.

training data, the benefits will become more visible. Nonetheless, since it is difficult to expect which feature selection works best in advance, learning dynamic metric weights through MoE seems extremely useful.

Case Study To illustrate benefits of MoE-CIU, we selected one example in Table 5 where we show the context, the response being assessed, different metric scores, and the gold human rating.

$d_{context}$: Until 1805 in the us, the runner up in a presidential election automatically became the vice president.
Response: Yeah i wonder what the president of zimbabwe looks like?
CIU: 0.32 RUBER-BERT: 0.00 BERTScore: 0.36
MoE-CIU: 3.33 Rating: 3.66

Table 5: An example showing how MoE-CIU handles potentially conflicting signals from individual metrics.

Here, although RUBER-BERT was very confident in classifying this example as NOT useful, MoE-CIU still predicted 3.33 ratings, which is much closer to gold ratings given other useful signals from CIU and BERTScore. A more comprehensive insights and additional examples are included in Appendix A.

7 Conclusion

We introduced CIU, a novel utility metric for assessing the quality of retrieval augmented conversations. Based on our experiments on two popular retrieval augmented (a.k.a. knowledge-grounded) conversation corpus, we conclude that CIU was the best metric among other lexical baselines. Although RUBER-BERT surpassed CIU performance on Persona Chat, considering the complexity of RUBER-BERT (e.g., training and inference), CIU is still an easy-to-use metric that can achieve similar results with no training, which answers **Q1**.

For **Q2**, we demonstrated the potentials of unifying multiple independent metrics into a single reward signal without any LLM dependency. This was achieved through our MoE-CIU model, and experiments confirm its effectiveness over any standalone metric. Insights from this study suggest promising directions for applying MoE-CIU as a proxy for an unified reward signal to optimize.

Limitations

Although our work proposes an approach to model unified rewards, reward optimization approaches to update dialog policy (e.g., RLHF) are left for future work. MoE-CIU also requires a small number of human annotations, thus our approach will require manual labeling. Our proposed discounting functions were only validated on English corpora, thus it is unclear how well CIU can generalize to multilingual setting. Unfortunately, experiments involving different languages and cultures are beyond the scope of this paper. Lastly, instead of exhaustive comparisons, we only selected the most widely used metrics to keep our experiments simple.

Acknowledgements

We would like to express our gratitude to Zhiyu Chen and the anonymous reviewers for their insightful feedback on our work.

References

Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 605–614.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

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.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124.

Ronan Collobert, Samy Bengio, and Yoshua Bengio. 2002. A parallel mixture of svms for very large scale problems. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. *CoRR*, abs/1904.10635.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Abhinav Jain, Vishwanath P. Singh, and Shakti P. Rath. 2019. A multi-accent acoustic model using mixture of experts for speech recognition. In *INTERSPEECH*, pages 779–783. ISCA.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017a. Towards an automatic turing test: Learning to evaluate dialogue responses. *CoRR*, abs/1708.07149.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017b. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126.

Widad Machmouchi, Ahmed Hassan Awadallah, Imed Zitouni, and Georg Buscher. 2017. Beyond success rate: Utility as a search quality metric for online experiments. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Saeed Masoudnia and Reza Ebrahimpour. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42:275–293.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gemnet: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 659–668.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

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.

Peng Qi, Yuhao Zhang, and Christopher D Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 25–40.

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.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20.

Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading conversational search by suggesting useful questions. In *Proceedings of The Web Conference 2020*, pages 1160–1170.

Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 473–482.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. *CoRR*, abs/1701.03079.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Alfan Farizki Wicaksono and Alistair Moffat. 2020. Metrics, user models, and satisfaction. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 654–662.

Bangpeng Yao, Dirk Walther, Diane Beck, and Li Feifei. 2009. Hierarchical mixture of classification experts uncovers interactions between brain regions. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating web search with a bejeweled player model. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 425–434.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

A Additional Case Studies of MoE-CIU

To illustrate how MoE-CIU benefits from ensemble learning, we highlight three examples in Table 6 where we show the context, the response being assessed, and the scores from the various models, and the gold human rating.

$d_{context}$: Until 1805 in the us, the runner up in a presidential election automatically became the vice president.
Response: Yeah i wonder what the president of zimbabwe looks like?
CIU: 0.32 RUBER-BERT: 0.00 BERTScore: 0.36
MoE-CIU: 3.33 Rating: 3.66

$d_{context}$: In september of 2010, the united nations appointed official ambassador to extraterrestrials in they case they would ever make contact with earth.
Response: I'm not sure. I wonder if the un has an ambassador to aliens.
CIU: 0.36 RUBER-BERT: 0.99 BERTScore: 0.31
MoE-CIU: 2.66 Rating: 2.67

$d_{context}$: Burger king had a promotion where if you unfriend 10 people on facebook you get a free whopper.
Response: Did you know burger king had a promotion where they gave you a burger for every 10 people you unfriended? I could have had 100 burgers!
CIU: 0.91 RUBER-BERT: 0.99 BERTScore: 0.62
MoE-CIU: 4.53 Rating: 5.00

Table 6: Example predictions from MoE-CIU highlighting how different metrics help recover false-positives from individual metrics.

One obvious issue we noticed was that it is difficult to calibrate the score distribution of learned models because on all three examples, RUBER-BERT is extremely confident about its decision. However, MoE architecture is particularly helpful since it can combine different metric outputs into a more continuous score distribution that is better aligned with human ratings.

In the first example, RUBER-BERT fails because the utterance and reference do not contain strong semantic relationship. Instead, they are loosely connected with an important keyword ('president'). RUBER-BERT was very confident in classifying this pair as NOT useful although human usefulness rating is 3.66. CIU and BERTScore successfully capture the overlap and assign a reasonable score compared to RUBER-BERT. The resulting model is capable of correcting RUBER-BERT's prediction to 3.33, which is only 0.33 off to human usefulness ratings. Without MoE, RUBER-BERT alone will predict this pair with 0.0 rating.

In the second example, RUBER-BERT strongly believes that the utterance and reference are semantically related. Although both inputs talk about alien ambassadors, the utterance does not use the information correctly. The reference clearly states United Nations appointed alien ambassadors but the utterance still questions the fact. RUBER-BERT is very confident that this example is highly related. However, CIU and BERTScore are able to regularize these effects if trained under MoE-CIU. The final score correctly predicted usefulness ratings with only 0.01 difference.

In the last example, it is clear that the input is highly relevant to $d_{context}$. Since individual metrics provide strong signals, MoE-CIU also predicted a very high rating of 4.53, which is close to 5.0.

How Much Annotation is Needed to Compare Summarization Models?

Chantal Shaib[1*] **Joe Barrow**[3*] **Alexa F. Siu**[2]
Byron C. Wallace[1] **Ani Nenkova**[2]
[1]Northeastern University, [2]Adobe Research, [3]Pattern Data
{shaib.c, b.wallace}@northeastern.edu
{asiu, nenkova}@adobe.com
joe.barrow@patterndataworks.com

Abstract

Modern instruction-tuned models have become highly capable in text generation tasks such as summarization. Given the regularity with which new model variants are now released, an increasingly practical problem entails choosing the best (zero-shot) summarization model for a particular domain confidently, but with minimal effort. In this work we empirically investigate the test sample size necessary to select a preferred model in the context of news summarization. Our results reveal that comparative evaluation converges quickly for both automatic and human evaluation, with clear preferences for a system emerging from under 100 examples. Collected human preference data allows us to quantify how well automatic scores can reproduce preference rankings across a variety of downstream summarization tasks. We find that while automatic metrics are stable at smaller sample sizes, only some automatic metrics are able to moderately predict model win rates according to human preference.

1 Introduction

Instruction fine-tuned language models are highly capable summarizers, and new such models are now released often. Continuously comparing such models using large, reference-based benchmark assessments is a costly task, especially if one wants to use them in a new domain. Here we demonstrate on (English) new summarization data that—with respect to both human and automatic evaluations—preferences toward a summarization model emerge over test sets of about 50 samples. Collecting human judgements, GPT evaluations, or (if possible) manually composed references for this size dataset is reasonable. Further, we evaluate GPT evaluations and two popular reference-based evaluations, ROUGE-1 and BERTScore, in terms of their ability to predict human preferences on a set of 36 testing contexts. We collect human judgements in the context of three different summarization tasks and three sources of input. For these variations, we compute the accuracy of automated scores to reproduce human preferences between pairs of systems.

[*]Work completed while at Adobe Research.

2 Background

Our goal is to establish the amount of test data needed to decide which of two summarization models produces better summaries for a given distribution over inputs (i.e., different sources of text to be summarized) and different task contexts for which the summary is to be used.

It is common to approach evaluation as a rate-then-compare task in which outputs from systems are rated for quality on a scale, and then average scores are used to compare systems. But it is well known that inputs may differ considerably in difficulty (Nenkova and Louis, 2008). Paired tests for statistical significance, that evaluate the differences of scores between two systems on the same input is the basis for comparison are therefore more appropriate (Rankel et al., 2011; Dror et al., 2018). Most contemporary work has embraced this approach, largely abandoning scoring of outputs and instead soliciting preferences among two or more choices (Novikova et al., 2018). Given developments in LLMs, pairwise win rates have become the *de facto* standard for reporting comparisons between instruction tuned models. In this work we similarly adopt win rate to compare systems, and we empirically identify the smallest test set size that reliably reveals preferences.

Most closely related to our work is the study on estimating power of tests for statistical significance, i.e., the minimum test size necessary to detect statistical differences of a given size (Card et al., 2020). Our work is aligned with the main question of this prior work, but we present empirical estimates of differences between systems without making any assumptions of tests to be used or size of effect we

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 51–59

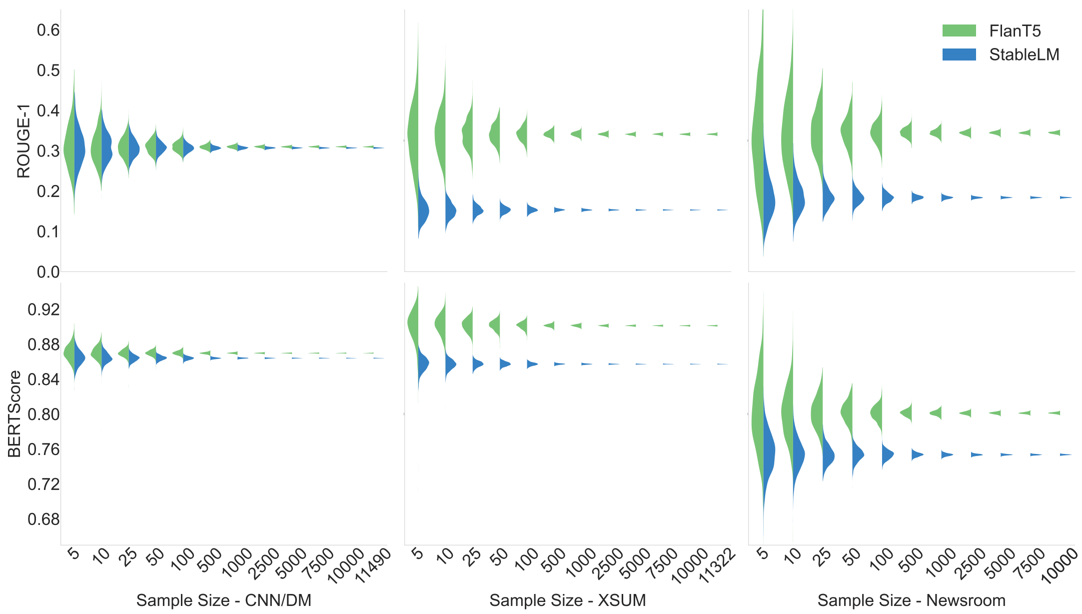


Figure 1: Distributions of average ROUGE-1 and BERTScores across 1000 re-samples. Differences between systems emerge clearly and quickly for XSUM and Newsroom.

want to detect. Our empirical findings may inform future work on power estimation.

Prior related work proposes ways of carrying out evaluations, either automatically or manually (Laban et al., 2022a; Zhang* et al., 2020; Fabbri et al., 2022; Zhong et al., 2022; Liu et al., 2022), and measuring the correlations between system rankings produced by human and automatic evaluations on a given benchmark (Gehrmann et al., 2023). We do not propose new evaluation methods, but rather introduce a method for validating automatic evaluations that does not rely on a benchmark, and instead measures the accuracy of automatic scores in reproducing human judgements across different input distributions and intended use-cases.

3 Unnecessarily Large Benchmarks

We first compare two models, FlanT5-XXL (Chung et al., 2022) and StableLM (Andonian et al., 2021) via automatic scores over three news summarization benchmarks: CNN/DM (See et al., 2017; Hermann et al., 2015), XSUM (Narayan et al., 2018), and Newsroom (Grusky et al., 2018). We use the test set splits of these datasets from Huggingface.[1]

CNN/Daily Mail and XSUM contain about 10K test inputs. The Newsroom test set split has over 100k samples. For efficiency, we randomly sample 10k examples from this set to scale it down to a size comparable to the other two datasets. We then generate summaries with FlanT5 and StableLM for all articles in the test sets, using the summarization prompts that these models have been trained on (see Appendix A). For each test split we sample 1000 times with replacement smaller test set sizes ranging from $[5, \texttt{len}(dataset)]$. We evaluate the two models with the commonly used ROUGE-1 (Lin, 2004) and BERTScore (Zhang* et al., 2020).[2] Both scores compare a summary with a human-written reference summary. ROUGE does so using tokens, while BERTScore relies on embeddings. We show score variations for FlanT5 and StableLM across the three datasets in Figure 1. For all three datasets, a preference for one of the models emerges early: The winning model as scored over 10k test points emerges after just 25-50 samples.

Given these findings, we collect human judgements on 100 samples from each of the data sources, varying the task context in which the judgement is made. We also add GPT-4 as another summarization model to be evaluated, and later report the accuracy of GPT-based evaluation against the aggregated human judgements.

4 Human Preferences

We hire three individuals on Upwork (Appendix F) for CNN/DM and Newsroom, and one for XSUM. We select 100 inputs for annotation from each dataset, which given the trends we observed in

[1] https://huggingface.co/docs/datasets/index

[2] We also report BLEU (Papineni et al., 2002) and SummaC-ZS (Laban et al., 2022b), in Appendix B.

the previous section, would be sufficient to reveal human preference.[3]

We also add summaries produced by GPT-4 for evaluation on the smaller dataset. FlanT5, StableLM, and GPT-4 represent encoder-decoder, decoder-only (open-source), and decoder-only (closed-source) models, respectively.

We instruct annotators to rank the summaries for each input in order of preference. This is a typical evaluation setting in which win rates—the percentage of input for which the model was preferred over the other—provide the clearest score for each model pair.

We provide three different scenarios to measure how preference may change based on context: *(i)* Rank the summaries in order of preference; *(ii)* Assuming you are monitoring the news for important world events, rank the summaries in order of preference; *(iii)* Which summary best captures the main details of the event being reported on? *(iv)* Which summary contains the fewest unnecessary details?

For GPT-4, we append the summaries with the instructions and provide these as prompts to the model.

4.1 Stability of Preference

First, we look to confirm whether smaller test samples are sufficient to make the same conclusion as with a larger sample. We apply the same procedure described in Section 3, where we resample 1000 test sets of size 25 and 50 from the 100 for which we have human judgements. Figure 2 shows the win rates for the CNN/Daily Mail test set for each of the three pairs of models, on the full test set of 100 samples, as well as the min, max and average win rate recorded across the 1000 smaller test sets.

While there is some variation in the strength of the preference for a model, the overall preference is preserved in the smaller samples. In only one case—the comparison between FlanT5 and StableLM—does the overall preference change for the minimum value of win rates from the one thousand samples of size 25. With 50 samples in the evaluation set, all three of the minimum, maximum and average win rates lead to the same conclusion about which system in the pair is better as that from the full 100 sample test set.

Similarly for the other two datasets, Newsroom and XSUM, none of the overall preferences flip for test sets of size 50 and only one minimum value for the 25 samples flips the preference. We provide the complete tables in Appendix C.

These results indicate that even under human evaluation, smaller test set samples (n=50) are adequate to conclude which is the preferred summarization model.

In many cases, the strength of the preference may be of interest. As shown in the variation between the minimum and maximum win rates, the strength as captured by win rates can vary considerably depending on the test set. We leave for future work analysis of the test size required to obtain reliable conclusions about the strength of the preference.

4.2 Human Preference Varies by Task and Input Source

We now turn to comparing model preferences relative to downstream task use.

Figure 3 shows the variation of aggregated preferences on the full 100 sample test set for CNN/Daily Mail. The context of the task can dramatically change the win rates for a given model. When contextualized in a specific use-case, human preferences flip from the overall rating for two out of the three model comparisons.

The overall win rate for StableLM over FlanT5 is 54%, indicating a weak preference for StableLM. In the world event use case however, the win-rate for FlanT5 increases to 53%, flipping to a preference for FlanT5. Similarly, the win rate of StableLM over GPT-4 in the overall condition is 21% but flips to 76% in the main details setting. The win rates of FlanT5 over GPT-4 remain stable across all tasks, always in favor of GPT-4.

Similarly, win rates according to the aggregate human preference for two systems vary depending with the source of data. In the next section we discuss how this observed variability changes the approach to validation of automatic evaluations.

5 Validating Automatic Evaluation

We presented qualitative evidence that the context in which preferences are made change the human preferences dramatically. We also provided clear examples of cases when human preference for the same two models can flip depending on the context. This judgement variability poses a novel requirement for validating automatic evaluation approaches. We cannot combine win rates across settings and compute correlations between human preferences and automatic scores because these

[3]See Appendix F for details about cost and hours for all annotations.

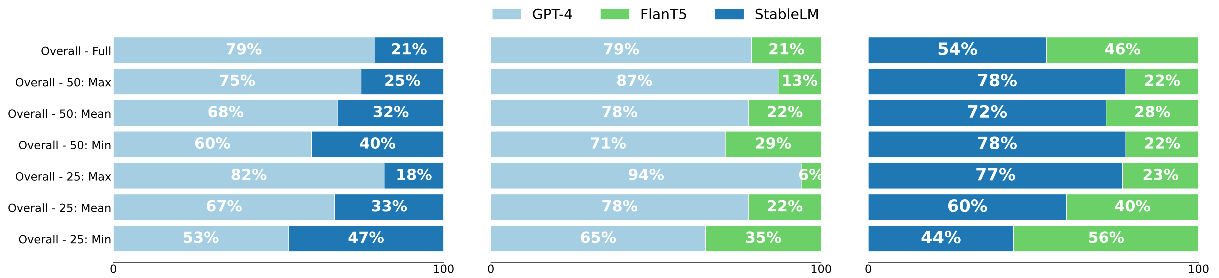


Figure 2: Aggregated annotator win rates for the CNN/DM dataset for the overall metric. Model preferences remain fairly stable across all sample sizes except in one case for sample size of 25.

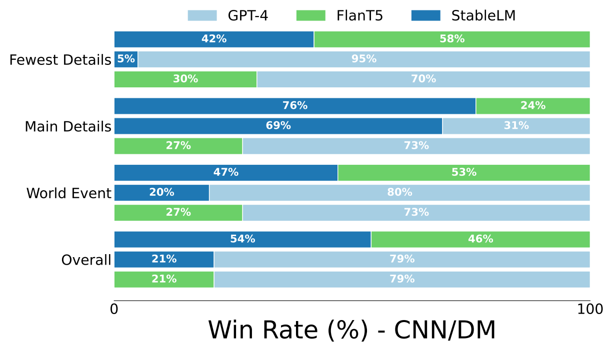


Figure 3: Aggregated annotator win rates across all metrics. Model preferences can change depending on the task setting.

Metric	Accuracy (%)
ROUGE-1	78
BERTScore	56
G-Eval	44
GPT-4 (as annotator)	78

Table 1: Accuracy of automatic metrics compared to human evaluations. GPT-4 as-an-annotator and ROUGE-1 score have the highest accuracy in predicting which model is selected by human annotators in each task setting.

come from different distributions. We do, however, have a sufficient number of pairs for comparison: 3 models evaluated on 3 sources of data, on 4 context of use. This yields 9 overall preferences and 27 contextually dependent preferences.

For four automatic methods for evaluation, we compute the accuracy of the automatic score in reproducing human preferences. Specifically, we compute the percentage of pairwise comparisons for which the automatic evaluation agrees with the human win rates on which system is the better one. This is a coarse requirement because it does not capture the size of the win rate. For example the win rate of one system over another in human preferences is 51% but an automatic score predicts that its win rate is 79%, the automatic score will be considered accurate.

Table 1 shows the accuracy for four automatic evaluations: ROUGE-1, BERTScore, G-Eval, and GPT-4 as an annotator. In the case of GPT-4 as an annotator, we provide GPT-4 with the exact same instructions as the human annotators. For the first three approaches, a win for a model is declared if the score assigned by the method for this input is higher than that for the other model. In cases when the scores for an input are the same, there is a tie. In the fourth case, using GPT-4 as an annotator provides ratings, so the wins are decided by the ranking returned by GPT-4 (rather than a proxy score). In this case, there are no ties because the annotators were asked to do a forced choice comparison. We find that ROUGE-1 and GPT-4 as an annotator are able to moderately predict the aggregated human preferences across the different tasks, compared to BERTScore and G-Eval which are not able to do so as reliably.

6 Conclusions

We presented automatic and human evaluations designed to establish the minimum amount of data necessary to choose between contemporary summarization models. Comparative evaluations establish which model performs better with test sets of 50 inputs. For human evaluation, a test size of 50 is sufficient to confidently establish which of two models people prefer. Human preference varies, however, depending on the intended use of the summary and on the source of data for summarization. This variation calls for new methods for validating automatic scores. We find that all four automatic evaluations predict preferences better than chance but lead to erroneous conclusions for many pairwise comparisons.

Limitations

We only evaluate over benchmark news datasets, where it is possible that our observations may not be reflected in other, more niche domains. In part, this choice is due to lack of availability of quality summarization datasets with references (and further motivating the need for evaluation over small samples), however it is important for future work to consider more specialized cases. Another limitation is that we do not collect human annotations nor GPT-4 summaries over the entire test set splits. This poses a challenge as collecting these evaluations and summaries over such a big dataset is costly.

References

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Wang Phil, and Samuel Weinbach. 2021. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With Little Power Comes Great Responsibility. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77:103–166.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022a. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022b. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Alexander R Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2022. Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation. *arXiv*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Ani Nenkova and Annie Louis. 2008. Can you summarize this? identifying correlates of input difficulty for multi-document summarization. In *Proceedings of ACL-08: HLT*, pages 825–833, Columbus, Ohio. Association for Computational Linguistics.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 72–78. Association for Computational Linguistics.

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.

Peter Rankel, John Conroy, Eric Slud, and Dianne O'Leary. 2011. Ranking human and machine summarization systems. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 467–473, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Model	Prompt
FlanT5	`[TEXT]`\nWhat is a one-paragraph summary of the above article?
StableLM	<|SYSTEM|># StableLM Tuned (Alpha version) - StableLM is a helpful and harmless open-source AI language model developed by StabilityAI. - StableLM is able to facilitate human communication by providing a summary of a given text. - StableLM is able to provide summaries that are useful and relevant to the given text. <|USER|> `[TEXT]`. Summarize the given piece of text. <|ASSISTANT|>
GPT-4	"role": "user", "content": "`[TEXT]` \n\n Summarize the above text. \n\n"

Table 2: Input and prompt structure for each summarization model. `[TEXT]` is replaced with the article to be summarized.

Appendix

A Summarization Prompt Details

For the summarization prompts, we use prompts and input structures that the models have been trained on. Table 2 shows the input for each model, where `[TEXT]` is replaced with the article to be summarized.

B BLEU and SummaC-ZS

Figure 4 shows the distributions of averaged BLEU and SummaC-ZS scores over all three datasets. BLEU scores have trouble capturing meaningful scores across longer inputs as seen with StableLM. SummaC-ZS uses NLI-models to score sentence-level information – similar to ROUGE-1 and BERTScore, we can start differentiating models earlier than the full sample size.

C Human Evaluation Win Rates and Sample Sizes: XSUM and Newsroom

We provide the aggregated win rates across annotators for XSUM (Figure 5) and Newsroom (Figure 6). Both datasets show the same trend as in Figure 2, where the win rate pair ranking is preserved in the minimum, maximum, and average win rates across 1000 trials. This holds across sample sizes of 50, but not in *all* cases with sample size of 25.

D Human Evaluation Win Rates and Tasks: XSUM and Newsroom

Similar to Figure 3, we show the win rates across different tasks for XSUM and Newsroom in Figure 7. These results support the finding that preference changes between downstream scenarios.

CNN/DM		
Annotators	Factuality κ	Text Quality κ
1, 2	0.522	0.053
1, 3	0.249	0.539
2, 3	0.133	-0.081

Table 3: Agreement scores, Cohen's kappa.

E Annotator Agreement on Text Quality and Factuality

For CNN/DM we report the agreement scores over factuality and text quality questions that we collect in our surveys in Table 3. We expect the agreement scores for factuality to be much higher; it is possible that this is an indicator for different tolerance for minor errors (e.g., vague wording) or may be indicative of the cognitive load involved in judging factuality. Similarly for text quality, the threshold for artifacts or other issues may differ between annotators.

F Annotation Details

Costs We hired seven professional proofreaders from Upwork, who were each recruited to read 100 articles and rank 3 summaries per article. We paid each annotator a flat fee of $325 to evaluate the summaries When asked for a time estimate after they completed, responses ranged between 10 and 13 hours to complete the study, meaning annotators were compensated at roughly $25-$30 per hour. The annotators typically completed the work over one to three days.

Annotation Platform We hire annotators on Upwork[4]. We presented the annotators with a custom interface for ranking the summaries and answering questions, shown in Figure 8. Annotators were encouraged to take extended breaks during annotation to reduce task fatigue.

[4] `https://www.upwork.com/nx/enterprise-homepage/`

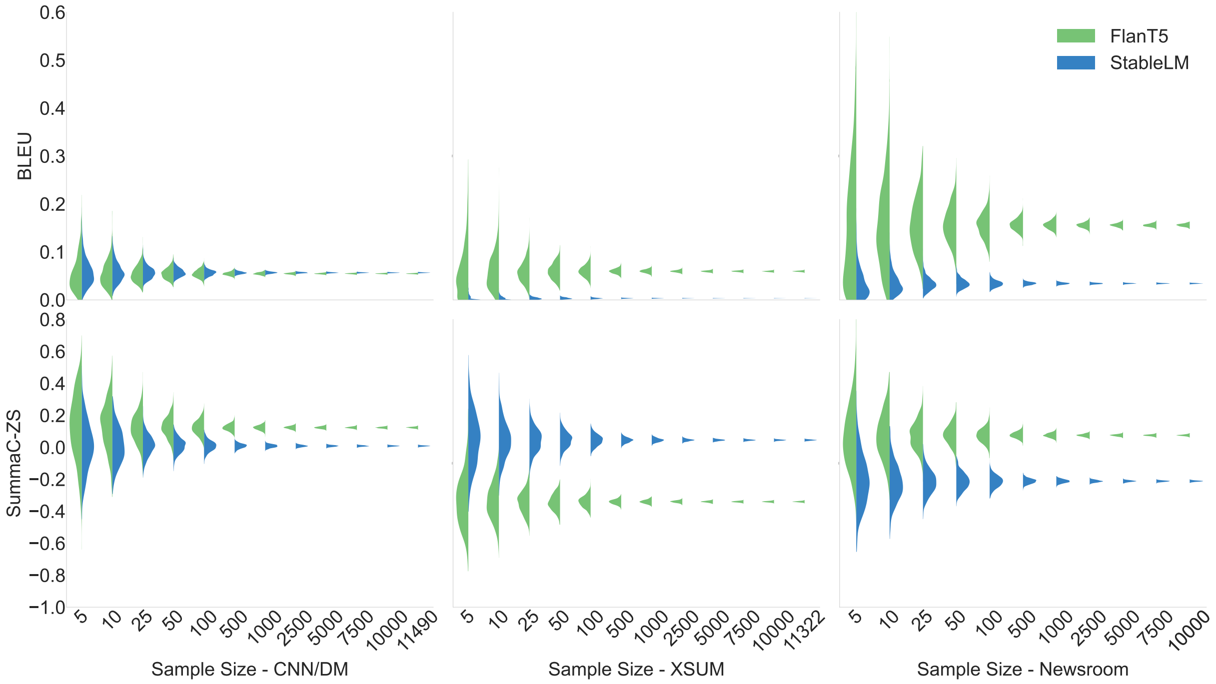


Figure 4: Distributions of averaged BLEU and SummaC-ZS scores across 1000 re-samples for CNN/DM, XSUM, and Newsroom.

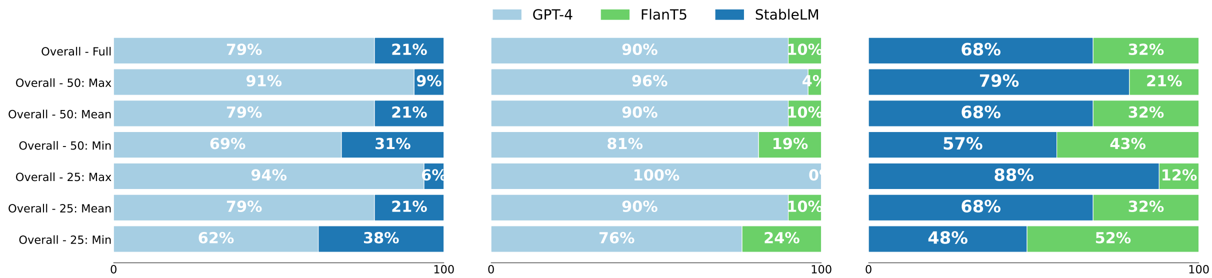


Figure 5: Win rates aggregated by annotators (XSUM).

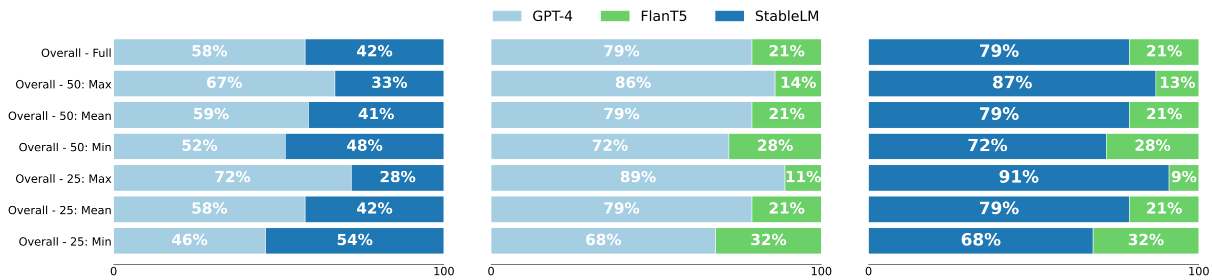


Figure 6: Win rates aggregated by annotators (Newsroom).

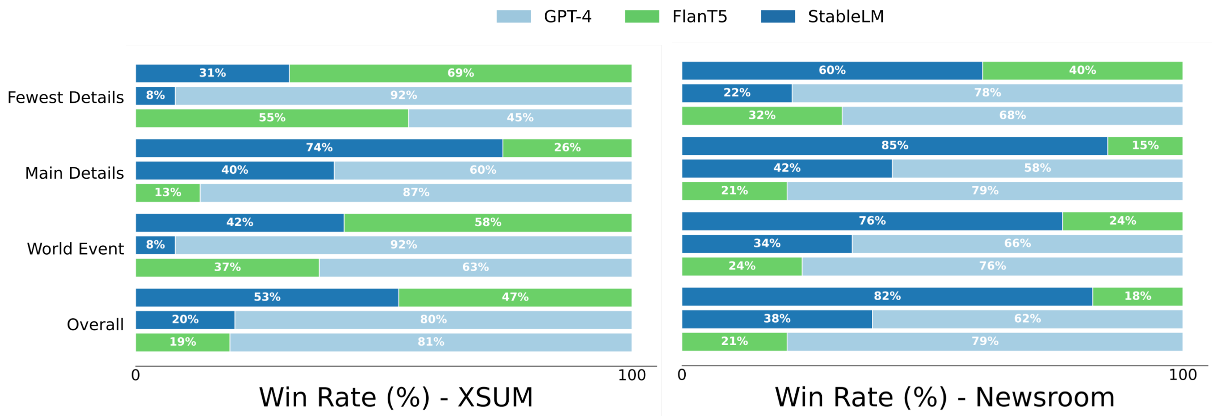
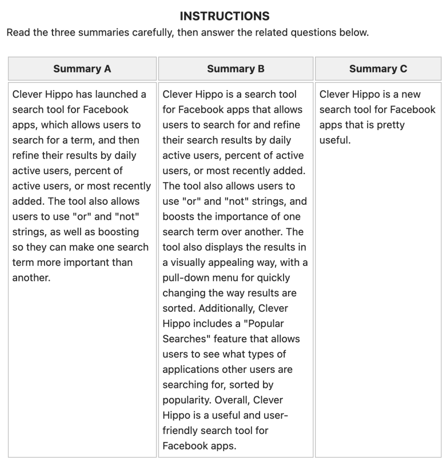
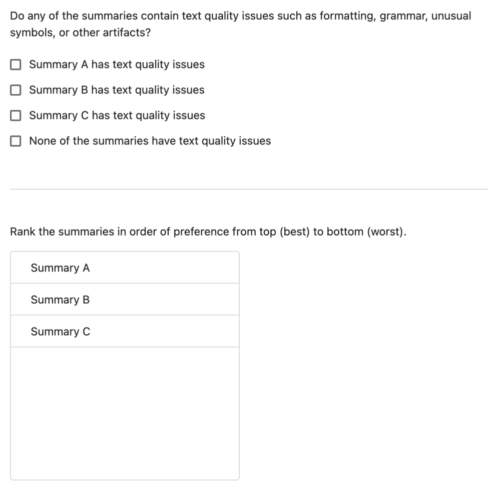


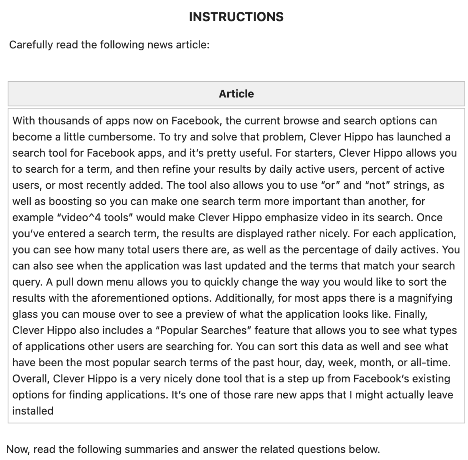
Figure 7: Aggregated annotator win rates across all metrics over the XSUM and Newsroom datasets.



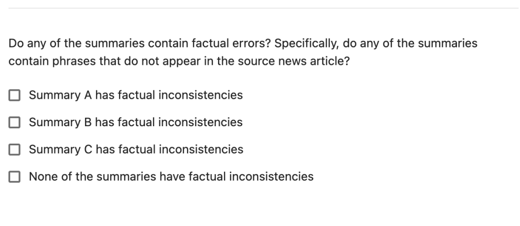
(a) Summaries, as presented to the annotators.



(b) Text quality issues and the ranking interface for the summaries. Each box with the summary label can be dragged-and-dropped into any order.



(c) Article, as presented to the annotator.



(d) Factuality questions asked about each summary.

Figure 8: The annotation interface. For each article, annotation happens across two pages. The first page contains the summaries (8a) and rankings (8b), and the second page contains the article (8c) and factuality questions (8d).

An Interactive Co-Pilot for Accelerated Research Ideation

Harshit Nigam and **Manasi Patwardhan** and **Lovekesh Vig** and **Gautam Shroff**

TCS Research

{h.nigam, manasi.patwardhan, lovekesh.vig, gautam.shroff}@tcs.com

Abstract

In the realm of research support tools, there exists a notable void in resources tailored specifically for aiding researchers during the crucial ideation phase of the research life-cycle. We address this gap by introducing ‘Acceleron’, a ‘Co-Pilot’ for researchers, designed specifically to accelerate the ideation phase of the research life-cycle. Leveraging the reasoning and domain-specific skills of Large Language Models (LLMs) within an agent-based architecture with distinct personas, Acceleron aids researchers through the formulation of a comprehensive research proposals. It emulates the ideation process, engaging researchers in an interactive fashion to validate the novelty of the proposal and generate plausible set-of hypotheses. Notably, it addresses challenges inherent in LLMs, such as hallucinations, implements a two-stage aspect-based retrieval to manage precision-recall trade-offs, and tackles issues of unanswerability. Our observations and end-user evaluations illustrate the efficacy of Acceleron as an enhancer of researcher’s productivity.

1 Introduction

With fast-paced research happening in every field, we are witnessing an exponential growth in the number of scientific articles and research papers on the web. It is difficult for an individual researcher or a small research team to keep abreast of the relevant advances amidst this information explosion. This has a downstream impact on the ability to be consistently appraised and ensure novelty of a proposed solution at various stages of the research life cycle. Thus there is an urgent need for a tools that can aid researchers to 1) understand, evaluate and incorporate the latest developments in the literature and 2) Formulate/Modify the current proposed solution accordingly to ensure novelty and impact.

Most of the existing tools focus on notifying and recommending researchers with relevant literature, facilitate exploration of existing literature and/or writing research manuscripts. Researchers have also proposed learning representations for retrieval of relevant scientific articles (Singh et al., 2022; Cohan et al., 2020; Ostendorff et al., 2022; Mysore et al., 2021), literature Review Generation (Hu and Wan, 2014; Kasanishi et al., 2023; Chen et al., 2021), Question Answering over scientific articles (Saikh et al., 2022; Dasigi et al., 2021; Lee et al., 2023), Scientific document summarization (Hayashi et al., 2020), citation recommendation (Ali et al., 2021, 2022; Medic and Snajder, 2023) citation intent detection (Cohan et al., 2019; Berrebbi et al., 2022; Roman et al., 2021; Lauscher et al., 2021), critical review and rebuttal generation (Ruggeri et al., 2022; D’Arcy et al., 2023; Kennard et al., 2021; Dycke et al., 2022; Wu et al., 2022), etc. However, to the best of our knowledge, no tool or no approach in the literature facilitates a researcher during the most arduous ideation stage of the research life-cycle. (Wang et al., 2024) attempts ideation in completely automated fashion. However, their results demonstrate ∼40% gap in the generation of ideas ‘helpful’ from the novelty

Figure 1: Acceleron Interface

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 60–73

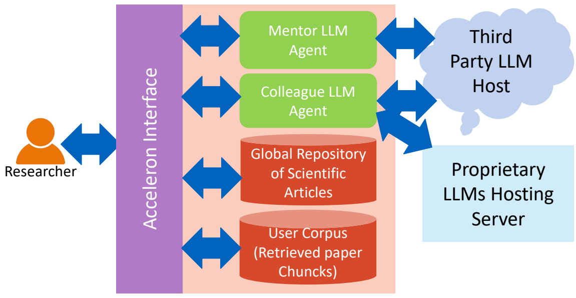


Figure 2: System Architecture

perspective.

Most of the tasks involved in research require domain expertise and complex reasoning skills. The recent advancement in Large Language Models (LLMs) has made it possible to partially automate some of these tasks (Liu and Shah, 2023; Liang et al., 2023; Zhang et al., 2023a; Lahiri et al., 2023; Kunnath et al., 2023). However complete automation of these tasks may not yield qualitative outcomes. In this work, we propose 'Acceleron' (Figure 1), a tool to accelerate the research life cycle. The tool exploits the reasoning and domain specific skills of LLM based agents not to replace researchers but to assist them for research ideation. With LLM powered *mentor* and *colleague* agents, Acceleron provides relevant inputs to researchers in an interactive fashion via a user-friendly interface. Thus, it aids the researcher to develop the research proposal consisting of a validated motivation, a well-defined research problem focusing of research gaps in the literature, a proposed approach selected from a set-of plausible synthesized methods and possible set-of experiments to be conducted to evaluate the approach for the research problem. To the best of our knowledge, we are the first ones to mimic the research ideation process using LLMs and execute it using human-machine interaction ensuring accelerated as well as qualitative outcomes, in terms of novel ideas.

2 System Architecture

Acceleron provides a web-based interface for researchers to interact. The system architecture is illustrated in Figure 2. We define an LLM Agent based architecture (Wang et al., 2023b), with agents of two distinct types of profiles or personas. A *Colleague* persona[1] performs less complex tasks including extraction of relevant information from user inputs, generation of relevant questions from extracted information or retrieval of relevant data from scientific documents. Whereas, *mentor* persona[2] performs more complex tasks requiring reasoning such as understanding the limitations or gaps of the existing work, identifying problems similar to the problem discussed in the proposal, identifying sub-tasks of the problem being solved in the proposal, solving similar problems and/or sub-tasks to synthesize a solution to the proposed problem and re-write the proposal given a plausible set-of approaches or possible limitations of related work. The architecture is flexible such that the LLM agents can interact with (i) LLMs like GPT-3.5-Turbo[3], Cohere[4] and Gemini[5] using API calls or (ii) open-source LLMs like Llama-2[6], Zephyr[7], Mixtral[8] which reside on an internal hosting server.

We expect to have a global repository which is a vector store of domain specific scientific articles[9] which are indexed by the Specter embeddings (Cohan et al., 2020) produced using the paper's title and abstract. We also have a User Specific corpus which has chunks of all the retrieved papers relevant to the current proposal the researcher is working on. The paper chunks are created with our in-house parser[10] treating paragraphs as semantic segments. If a paragraph does not fit into the the maximum token length of LLM agents, while chunking it is further split to fit into the maximum token length. The chunks are further converted to vector embeddings and indexed for efficient retrieval based on semantic similarity with a query. This user corpus acts as a shared 'memory' for the LLM agents.

3 Approach

The Acceleron Ideation simulation involves interaction between a researcher and the LLM agents, where the LLM agents perform actions based on the feedback received by the researcher or another agent. The process takes a proposal as an input from a researcher with a research problem description specified at a high level along with the motiva-

[1]OpenAPI's GPT-turbo-3.5 model

[2]OpenAPI's GPT4 model
[3]https://platform.openai.com/docs/models/gpt-3-5-turbo
[4]https://cohere.com/
[5]https://gemini.google.com/
[6]https://llama.meta.com/
[7]https://huggingface.co/HuggingFaceH4/zephyr-7b-beta
[8]https://mistral.ai/news/mixtral-of-experts/
[9]We use more than 2 million scientific articles in semantic scholar fetched using S2ORC dataset (Lo et al., 2020) as the global repository
[10]We built a PDF parser using PDFminer

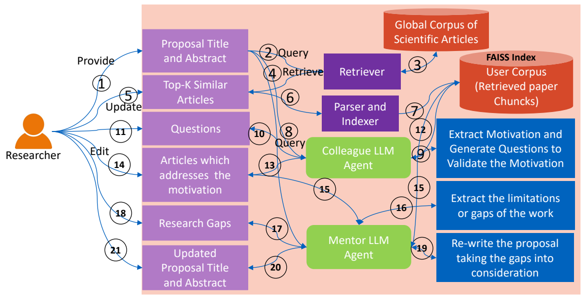


Figure 3: Motivation Validation Pipeline

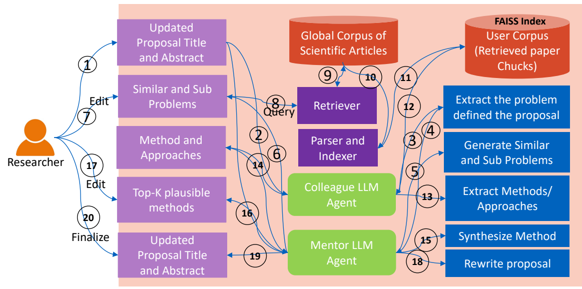


Figure 4: Method Synthesis Pipeline

tion behind the problem. The process involves: (i) Analyzing the existing literature to critically evaluate the motivation behind the research problem a researcher is trying to address to ensure that the mentioned research gap(s) still exist(s), (ii) Reformulating the proposed research problem and objectives based on the validation stage output and re-identification of research gaps, (iii) Identifying analogous research problems or sub-problems addressed in the literature and utilizing their solutions, available in the literature, to derive a set-of approaches or synthesizing a set-of plausible methods as a solution to the problem, (iv) Designing experimentation strategy for the given problem and selected methodology. The output of the ideation process is the updated proposal with a (i) A research problem with validated motivation (ii) Plausible methods to address the research problem. The overall ideation task is split into two pipelines: (i) Motivation Validation and (ii) Method Synthesis. The detailed prompts for the steps in each of the pipeline are illustrated in the Appendix Section A.2.

3.1 Motivation Validation Pipeline

As elaborated in Figure 3, the workflow begins with the researcher providing the title and abstract for their proposal. Acceleron identifies and extracts the motivation behind the proposal and retrieves relevant scientific articles relevant to the proposal and presents them to the researcher for review. The researcher can edit the selection of articles as needed. Subsequently, the system generates binary questions to validate the proposal's motivation against the retrieved articles. After review and potential edits by the researcher, the system retrieves relevant sections from the selected articles to answer these questions. If all articles fail to sufficiently address the proposal's motivation, the researcher is notified. Otherwise, identified gaps in the literature are presented to the researcher for consideration. The researcher can select relevant gaps or propose new ones, which are then used to refine the proposal's motivation and problem statement. The revised proposal is presented to the researcher for further editing or approval. This iterative process continues until the proposal's novelty is validated or until no relevant articles are found.

3.2 Method Synthesis Pipeline

The Method Synthesis workflow is illustrated in Figure 4. The method synthesis phase begins with the motivation validated proposal being accepted by the researcher. The system employs the *colleague* agent to extract and define the proposal's problem, followed by the *mentor* agent generating similar research problems and decomposing the main problem into sub-tasks. The researcher can refine these generated problems. Each refined problem is used to retrieve relevant scientific articles which is then parsed and stored in the user corpus. The *colleague* agent then consolidates similar problems and their solutions from these articles, presenting them to the researcher for further editing. This information, along with the original proposal, is provided to the *mentor* agent, which synthesizes a list of plausible methods to solve the problem. The researcher selects preferred methods, which are incorporated into the updated proposal by the *mentor* agent. The revised proposal is then reviewed and finalized by the researcher.

4 Novel Components

With Acceleron our aim is to bridge Human-Computer Interaction and Natural Language Processing using an interactive tool infused with the best of NLP and goodness of HCI. We created several novel components within Acceleron that fixes known shortcomings of NLP based systems using HCI inspired ideas.

4.1 LLM Agents for Research Ideation

To the best of our knowledge, ours is the first LLM agent based tool which assists in the complex task of ideation for research. We have devised with two novel portfolios for LLMS, viz., *colleague* and *mentor*, allocating less complex tasks to the *colleague* agent and more complex reasoning based tasks to the *mentor* agent. The user corpus acts as the shared memory for the agents, whereas the agents perform fixed set of actions at various stages of the workflow based on the provided inputs as discussed in the prior sections. Rather than using a costly LLM like GPT4 for all the tasks involved in the workflows; dividing the tasks as per the difficulty level and leveraging less costly LLM such as GPT-turbo-3.5 for colleague agent, performing less complex tasks, provides a cost-effective solution for workflows.

4.2 Mitigation of Hallucination

Hallucination is one of the major difficulties of using LLMs for knowledge based tasks (Zhang et al., 2023b; Wang et al., 2023a). We mitigate this problem using a two-fold solution: (i) There are retrieval augmented components of the workflows, viz. the motivation validation workflow poses questions generated to validate the motivation of the proposal on the retrieved articles stored in the user corpus or extract limitations of the articles which address the proposal motivation or the method synthesis workflow extracts approaches used to solve similar or sub problems from the retrieved articles. For these retrieval augmented tasks through proper prompt engineering, we ensure that the answers are provided by restricting the knowledge to the retrieved context only. We observe this helps to mitigate hallucinations. (ii) There are components of the workflows which rely on parametric knowledge of LLMs, for example the motivation validation involves re-writing the proposal and the method synthesis involves generating similar sub problems for the problem defined in the proposal and synthesizing methods. For these tasks the output can not be restricted to the provided input. In such cases, there is a higher chance of hallucinated outputs. For such scenarios, we ensure mitigation of hallucinated outputs, by keeping the system semi-automated and allowing user-interactions at every step to edit or delete hallucinated outputs. Moreover at every stage of the workflow, the LLM agents are asked to justify their outputs and the provided justification is exposed to the researcher through the interface. This forces the model to apply Chain-of-Thoughts (COT) (Wei et al., 2022) and allows the researcher to validate the output and check if it is in sync with the justification provided. This assists in alleviating the effect of hallucinations.

4.3 Two-Stage Aspect Based Retrieval

The global corpus contains a large number of scientific articles stored with the Specter embedding of the title and abstract of the papers. The title and abstract of the papers contains information about motivation and problem statement of the papers and a high level mention of the methodology and the results. For ideation we require more in-depth information from the papers across various aspects such as methodology, limitations, etc. To achieve this we perform retrieval in two stages. In motivation validation workflow, we first retrieve top-K papers from the global corpus with the proposal as the query and high value of K for good recall. This allows us to have a set-of papers with similar motivation and problem statement to that of the proposal. These papers are chunked and stored in the user corpus for further aspect based retrieval, such as papers with similar motivation to that of the proposal and paper paragraphs mentioning the research gaps of these papers. In method synthesis workflow, we first retrieve top-K papers from the global corpus with similar sub problem statements as the query and high value of K for good recall. This allows us to have a set-of papers with problems similar to the problem described in the proposal or similar to any of the sub-tasks of the problem described in the proposal. These papers are chunked and stored in the user corpus for further aspect based retrieval such as extracting the approaches of the papers. Note that keeping high-recall for the first stage of retrieval ensures coverage of papers, whereas for the second stage we favor more precise outcomes for aspect based retrieval.

4.4 Introduction of Unanswerability

The output of aspect based retrieval is always top-K paragraphs from the retrieved and chunked papers. We keep the value of K low to get more precise retrieval for the given aspect based query. However, there is a possibility that the retrieved paragraphs do not have the answer to the query (the query is unanswerable). For example, in the motivation validation workflow the retrieved paragraphs from the papers do not answer the question of whether the paper addresses a specific motivation of the proposal and does not specify the limitations of the paper which would help to refine the problem defined in the proposal. Similarly, for the method synthesis workflow the retrieved paragraphs may not have an approach to solve a similar problem. In such cases, the LLM based agents check the relevancy of retrieved paragraphs for the given query and identifies the query as 'unanswerable' in case if all the retrieved paragraphs are irrelevant, avoiding irrelevant outputs. Allowing unasnwerability also assists in reduction of hallucinations.

4.5 Moderation of GenerativeAI

Output generated by API based closed-source Large Language Models like GPT-3.5 or Cohere, are always unmoderated relative to the domain they are being used in. Even though first party moderation in form of censorship and guardrails(Gehman et al., 2020)(Welbl et al., 2021) exist, these measures are focused on moderating offensive and inappropriate content being provided as input and generated as output by the LLM. Domain specific contextual moderation is necessary for a LLM to provide on-topic and context relevant outputs. An output generated as part of one domain may be irrelevant or inappropriate when taken out of context or when being provided as input to a LLM for a different task. To counter this issue we have specially designed our system using a novel expert-in-the-loop architecture where at each and every step where a LLM agent is called to generate an output, a context is created using our two-stage aspect based retrieval technique and task specific prompt provided by the human user themselves. This allows for the human to be in control of what the LLM is being fed as context for the output generation acting as a pseudo first layer of moderation. This in turn allows the LLM to generate domain relevant and topic appropriate output which is provided to the human for a review with option to edit if needed, so that the output can be used as context further down the pipeline, making a encapsulation of moderation on the LLM agents, negating the need for third party content moderation.

5 Qualitative Analysis of the Workflows

In the absence of an appropriate dataset for the tasks relevant to the ideation process, we evaluate our workflows by user-studies. We allow researchers working in distinct domains like computer science, material science and life science, to use Acceleron for ideation of their research problems. For computer science domain , we use Semantic Scholar data fetched using S2ORC dataset (Lo et al., 2020) as our global repository. Whereas, for material science and life science domain we use our repository of papers downloaded from 'Science Direct'[11] and 'PubMed'[12], respectively. We utilize the logging functionality of 'Acceleron' to keep track of the interactions between the researcher and the LLM Agents. For space constraints and data confidentiality preservation of unpublished work, here, we provide a qualitative analysis of the workflows with 2 proposals from distinct researchers, specifically in the domain of Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP). The topics of these proposals are: (i) Topic-based citation retrieval for research proposal and (ii) Reference-free evaluation metric for retrieval augmented question answering.

We receive an input from a researcher with a proposal titled '*Topic-based citation retrieval for research proposal*' and the corresponding abstract '*Retrieval of research articles pertinent to a given query represents a thoroughly investigated research challenge. Typically, queries take the form of a title and abstract of a research article, or a specific sentence or paragraph from an existing research article requiring citation. However, existing approaches presuppose the availability of a well-constructed manuscript, an assumption that is inappropriate during the initial research proposal writing stage. At this initial phase, researchers seek pertinent literature for citing in their proposals, often focusing on specific topics or intents and further build the proposal. In this work, we aim to tackle the issue of topic-based citation retrieval for research proposals. We anticipate researchers providing the title and abstract of their research proposals, encompassing elements such as the research gap, problem statement, and a high-level overview of the proposed methodology and experiments. Additionally, researchers will furnish a list of topics for which relevant scientific articles need to be retrieved. Our proposed algorithm intends not only to fetch research articles pertinent to the given proposal from a corpus, but also to establish a crucial many-to-many mapping between these*

[11] https://www.sciencedirect.com/

[12] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

articles and the specified topics.' The colleague LLM agent generates the following questions for validation of the motivation: 1. "Is the research paper specifically addressing the retrieval of research articles relevant to a topic of a research proposal?" and 2. "Is the research paper developing a technique to map research articles to specified topics in research proposals?". Out of top-50 research articles used to validate the motivation of the proposal by posing the above mentioned questions, four (Appendix A.1.1) got retrieved to be answering as 'yes' to at the least one of the above questions, and thus invalidating the motivation behind the proposal. However, the justifications provided for these papers highlight that paper no. 1 and 3 introduce an approach for citation recommendations during the writing phase of the target manuscripts and not at the proposal writing stage. Also, scientific article 2 leverages contents of a target paper and citation graph to extract scientific information. The outcome of the scientific article 4 is a dataset which can be useful for the proposal, but does not address the task of 'topic-based citation retrieval for research proposal'. Thus, we observe that after evaluating the retrieved scientific articles claimed to be invalidating the proposal, the researcher disagrees with the justifications provided for each of the retrieved articles for addressing the motivation behind the proposal, hence validating the novelty of the proposal. This exemplifies the need as well as the effectiveness of this human computer interaction facility provided by the tool for the workflow. This example demonstrates acceleration of motivation validation stage of the research-life cycle ($\sim$8x for this proposal as stated by the researcher), by eliminating the need for the researcher to manually go through multiple relevant research articles retrieved by generic or academic search engines to ensure that the literature does not have a solution for the specific problem the researcher is trying to address, leading to a time consuming process.

We receive input from another researcher with the proposal titled '*Reference-Free evaluation metric for Retrieval augmented question answering task*' and the abstract '*We observe that questions with long answers on long documents do not have unique reference evidences (relevant paragraphs from the document) and answers. Rather, there is a distribution over reference answers, making expert based evaluation expensive and existing unique reference-based evaluation metrics inadequate. We also do not find any reference-free evaluation metric designed for evaluating retrieval augmented question answering task. Hence, this this work we propose to define this metric.*'. The colleague LLM agent generates the following question to validate the motivation of the proposal: "Is the research paper proposes a reference-free evaluation metric designed for evaluating retrieval augmented question answering tasks?". We observe that out of top-50 retrieved scientific articles relevant to the proposal, none of the articles provides answer as 'yes' to the question, leading to retrieval of zero relevant paper hence invalidating the motivation of the proposal. Manual analysis of the top-50 retrieved articles performed by the researcher (as well as other relevant articles manually visited by the researcher) to evaluate this outcome of the workflow, substantiates the results.

For the next workflow of method synthesis for the above proposal, the mentor LLM agent generates following set of research problems similar to the problem defined in the proposal: 1. "Evaluating complex tasks where there is no unique correct answer or reference", 2. "Designing evaluation metrics for tasks that involve retrieval and interpretation of large amounts of data", 3. "Creating reference-free evaluation metrics for tasks where reference-based metrics are inadequate or impractical", 4. "Assessing the quality of answers in tasks where the answers can be long and drawn from extensive documents". The mentor LLM agent also generates the following sub-tasks for the problem defined in the proposal: 1. "Defining a new metric that can effectively evaluate retrieval augmented question answering tasks" and 2. "Overcoming the inadequacy of existing unique reference-based evaluation metrics for questions with long answers on long documents". With these similar and sub-problems as queries, the colleague LLM agent first retrieves Top-10 similar scientific articles per statement (total 40 articles with some overlap as same article may be retrieved for multiple queries) and then poses the question that 'if the article provides a methodology or an approach to solve the above defined problem or sub-task'. The researcher receives total 17 scientific articles which answers 'yes' to the question along with a description of the methodology implemented for each of the above problems or task. Out of these papers, the researcher accepts 11 (Appendix A.1.2) scientific articles, finding them to be more relevant to the problem the researcher is trying to address.

Based on approaches discussed in these scientific articles to solve the above enlisted similar problems and/or sub-tasks, the mentor LLM agent suggests the following 10 plausible methodologies for the proposal. Theses suggestions with the exact verbatim provided by the agent are as follows: 1. AVA (Automatic Evaluation Approach): This approach uses Transformer-based language models to encode the question, answer, and reference text. By measuring the similarity between the reference and an automatic answer, biased towards the question semantics, AVA can estimate the system accuracy. This approach can be used to evaluate retrieval augmented question answering tasks and overcome the inadequacy of existing unique reference-based evaluation metrics. 2. Informedness and Markedness Measures: These new probabilistic and information theoretic measures can be used to evaluate complex tasks where there is no unique correct answer or reference. They can also be used to design evaluation metrics for tasks that involve retrieval and interpretation of large amounts of data. 3. Large Language Models (LLMs): LLMs can be used as automated metrics for evaluating tasks that involve retrieval and interpretation of large amounts of data. They can also be used to assess the quality of answers in tasks where the answers can be long and drawn from extensive documents. 4. SacreROUGE: This open-source library dedicated to the development of summarization evaluation metrics can be used to develop reference-free evaluation metrics for tasks where reference-based metrics are inadequate or impractical. 5. Quantified Reproducibility Assessment (QRA): This method, based on concepts and definitions from metrology, can be used to develop reference-free evaluation metrics for tasks where reference-based metrics are inadequate or impractical. 6. RoSE Benchmark: This benchmark can be used to provide a more robust evaluation across different tasks. It can also be used to develop reference-free evaluation metrics for tasks where reference-based metrics are inadequate or impractical. 7. GENIE6: This human-in-the-loop leaderboard framework can be used to score answers in tasks where the answers can be long and drawn from extensive documents. 8. Sparse Attention and Contrastive Retriever Learning: This system conditions answer generation on identified documents and uses a dense retriever trained by scaling up a distantly supervised algorithm. It can be used to overcome the inadequacy of existing unique reference-based evaluation metrics for questions with long answers on long documents. 9. Unified Evaluation Benchmark for Long-form Answers: This approach involves conducting a thorough study of evaluation, including both human and automatic evaluation protocols. It can be used to overcome the inadequacy of existing unique reference-based evaluation metrics for questions with long answers on long documents. 10. Training an LFQA Evaluation Metric Directly on Human-Annotated Preference Judgments: This approach involves fine-tuning pre-trained Language Models based on human judgement scores for the task. This output showcases the quality of method recommendations provided by the tool for the given proposal. Though mentioned at high-level, the researcher agrees that most of these methods are well-suited as a plausible approach for the given proposal. Though there is a need for further work to finalize the most appropriate plausible method for proposal, the researcher finds this first cut of output provided by the tool to be relevant and the overall process to be significantly more efficient than the regular process followed by the researcher for constructing a plausible set-of approaches for the given problem, by searching through the relevant literature from scratch.

These examples illustrating the outcomes of the motivation validation and method synthesis phases of the ideation workflow of the tool, demonstrates the efficacy of the tool, in terms of providing relevant outputs at each stage of the workflow. The observations made in terms of time saved by the researchers with the tool usage for the respective tasks demonstrates the power of the tool with regards to time efficiency gains.

6 Conclusion

In this work, we have demonstrated a tool called 'Acceleron', developed to accelerate the ideation phase of the research life-cycle. To the best of our knowledge this is the first tool which addresses the tasks involved in the ideation stage. To emulate the ideation process, we use LLM agents with colleague and mentor personas to execute two workflows, viz. motivation validation and method synthesis, which engage researchers in an interactive fashion to develop the research proposal. Our workflow involves novel components to (i) alleviate the hallucinations of LLMs through user interaction, (ii) ensure relevant outcomes by two-stage aspect based retrieval, where first stage introduces higher recall reducing False Negatives and False Positives are corrected by user interaction and second stage of more precise fine-grained aspect-based retrieval, (iii) introduction of unanswerability and (iv) Moderation of GenerativeAI via human interaction acting as a pseudo first layer of moderation increases user involvement in the final task specific outcome. The qualitative analysis performed with proposals from researchers in distinct domains, demonstrates qualitative outcomes for various stages in the workflow with ∼7.5x gains in the time efficiency for various stages of the ideation phase. Most importantly, expert-interaction avoids error propagation through the stages of workflows yielding qualitative outputs in terms of generation of novel and diverse ideas.

7 Future Works

This is an ongoing work. In future, we plan to emulate the domain specific aspects of the ideation process creating domain specific instances of the workflows. For example, there can be a specialized workflow for synthesis of alloys in material science domain or drug discovery or synthesis of clinical trials in life science domain. This would result into a meta-process for ideation, which is domain independent and instances of this meta-process customized for specific domains and / or tailor made for specific tasks.

The logging functionality of 'Acceleron' keeps track of every input provided to the researcher as well as LLM agents and every output from them along with the corresponding timestamps. We are saving these logs for each user interactions for all the sessions. We plan to use these logs with treating user validated inputs as ground truth annotations, to develop a datasets for the ideation process. The logs would be used for developing datasets for tasks such as: (i) retrieval of research papers with similar motivation (ii) proposal re-writing with addressing research-gaps (iii) retrieval of research papers with similar problems and/or (iv) method-synthesis from a set-of relevant papers. The datasets will be used to instruction-tune the Open-Source LMs, which can replace the existing LLMs yielding more cost-effective solutions.

We plan to extend the implementation of current phase to generate a list of experiments to be performed for the problem defined in the proposal and the methodology selected by the researcher. This would lead to generation of a (set-of) results table(s) in a semi-automated fashion, with baseline approaches, planned experiments (ablations) and appropriate metric(s) used for evaluation.

8 Limitations

The current version of ideation part of 'Acceleron' has certain limitations. The system generates descriptions for every generated question at every stage for the researcher to elaborate and explain of the outcomes of these stages. For example, if an existing paper is retrieved to be already addressing the motivation behind the proposal, the tool provides LLM generated description of the same to explain how the paper is already addressing the motivation. However, these descriptions sometimes are not sufficient for the researcher to evaluate if the retrieved outputs are correct, further hindering the process of updating the outputs. To counter this we are planning to extend this functionality by providing a facility to showcase the whole paper and highlight the chunk of context in the paper using which the description is generated. This would not only provide vital context to the researcher to understand the answer but also provide backtracking ability to check the context retrieved to generate the description for a particular question.

We typically observe that we do not get qualitative results for extracting limitations of user proposal as relevant retrieved papers do not specifically mention the limitations. In future we plan to enhance the reasoning capabilities of LLMs to extract limitations from a research paper. The open-source locally run LLMs like Llama-2 (Touvron et al., 2023) and Zephyr(Tunstall et al., 2023) are slow and produce less qualitative outcomes as compared to API based LLMs like GPT-3.5-Turbo and GPT4 driving up the cost of running the system. A single execution of the 2 workflows for a single proposal cost the researcher somewhere around $0.5 to $1 for GPT-3.5-Turbo depending on the inputs and context provided by the user and the number of papers retrieved for the proposal, whereas this cost is almost 10-fold for GPT4. To achieve better quality of explanations from the retrieved papers we plan to decontextualize the citations embedded in the retrieved papers by using an approach similar to (Newman et al., 2023). Moreover, we need a benchmark and metric to evaluate our idea generation pipeline. Right now, we are doing it by user-studies and expert feedback. However, we plan to use the newly released SciMon(Wang et al., 2024) Dataset to benchmark the ideation workflows and further enhance them.

References

Z. Ali, Guilin Qi, Pavlos Kefalas, Shah Khusro, Inayat Khan, and Khan Muhammad. 2022. Spr-smn: scientific paper recommendation employing specter with memory network. *Scientometrics*, 127:6763–6785.

Z. Ali, Guilin Qi, Khan Muhammad, Pavlos Kefalas, and Shah Khusro. 2021. Global citation recommendation employing generative adversarial network. *Expert Syst. Appl.*, 180:114888.

Dan Berrebbi, Nicolas Huynh, and Oana Balalau. 2022. Graphcite: Citation intent classification in scientific publications via graph embeddings. *Companion Proceedings of the Web Conference 2022*.

Xiuying Chen, Hind Alamro, Li Mingzhe, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Annual Meeting of the Association for Computational Linguistics*.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *ArXiv*, abs/1904.01608.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *ArXiv*, abs/2004.07180.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. Aries: A corpus of scientific paper edits made in response to peer reviews. *arXiv preprint arXiv:2306.12587*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Nlpeer: A unified resource for the computational study of peer review. *ArXiv*, abs/2211.06651.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Hiroaki Hayashi, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. What's new? summarizing contributions in scientific literature. *ArXiv*, abs/2011.03161.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: An optimization approach. In *Conference on Empirical Methods in Natural Language Processing*.

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. Scireviewgen: A large-scale dataset for automatic literature review generation. In *Annual Meeting of the Association for Computational Linguistics*.

Neha Nayak Kennard, Timothy J. O'Gorman, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Rajarshi Das, Hamed Zamani, and Andrew McCallum. 2021. Disapere: A dataset for discourse structure in peer review discussions. *ArXiv*, abs/2110.08520.

Suchetha Nambanoor Kunnath, David Pride, and Petr Knoth. 2023. Prompting strategies for citation classification. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*.

Avishek Lahiri, Debarshi Kumar Sanyal, and Imon Mukherjee. 2023. Citeprompt: Using prompts to identify citation intent in scientific papers. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 51–55.

Anne Lauscher, B. R. Ko, Bailey Kuehl, Sophie Johnson, David Jurgens, Arman Cohan, and Kyle Lo. 2021. Multicite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. *ArXiv*, abs/2107.00414.

Yoonjoo Lee, Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Ho Hin Lee, and Moontae Lee. 2023. Qasa: Advanced question answering on scientific articles. In *International Conference on Machine Learning*.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel A McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *ArXiv*, abs/2310.01783.

Ryan Liu and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *ArXiv*, abs/2306.00622.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Zoran Medic and Jan Snajder. 2023. Paragraph-level citation recommendation based on topic sentences as queries. *ArXiv*, abs/2305.12190.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2021. Multi-vector models with textual guidance for fine-grained scientific document similarity. *ArXiv*, abs/2111.08366.

Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A question answering framework for decontextualizing user-facing snippets from scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3194–3212, Singapore. Association for Computational Linguistics.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Muhammad Roman, Abdul Shahid, Shafiullah Khan, Anis Koubâa, and Lisu Yu. 2021. Citation intent classification using word embedding. *IEEE Access*, 9:9982–9995.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2022. Argscichat: A dataset for argumentative dialogues on scientific papers. *ArXiv*, abs/2202.06690.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23:289 – 301.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *ArXiv*, abs/2211.13308.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *ArXiv*, abs/2310.07521.

Lei Wang, Chengbang Ma, Xueyang Feng, Zeyu Zhang, Hao ran Yang, Jingsen Zhang, Zhi-Yang Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji rong Wen. 2023b. A survey on large language model based autonomous agents. *ArXiv*, abs/2308.11432.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. Scimon: Scientific inspiration machines optimized for novelty.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Yang Zhang, Yufei Wang, Kai Wang, Quan Z. Sheng, Lina Yao, Adnan Mahmood, Wei Emma Zhang, and Rongying Zhao. 2023a. When large language models meet citation: A survey. *ArXiv*, abs/2309.09727.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.

A Appendix

A.1 Qualitative Analysis of the Workflow: Retrieved Papers

A.1.1 Papers Retrieved during Motivation Validation of Proposal 1

1. "Citation Recommendation: Approaches and Datasets"
2. "CitationIE: Leveraging the Citation Graph for Scientific Information Extraction"
3. "Content-Based Citation Recommendation"
4. "unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network"

A.1.2 Papers retrieved during Method Synthesis Workflow of Proposal 2

1. "AVA: an Automatic eValuation Approach to Question Answering Systems"
2. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation"
3. "Re-visiting Automated Topic Model Evaluation with Large Language Models"
4. "SacreROUGE: An Open-Source Library for Using and Developing Summarization Evaluation Metrics"
5. "Quantified Reproductibility Assessment of NLP Results"
6. "Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation"
7. "A Critical Evaluation of Evaluations for Long-form Question Answering"
8. "Think you have Solved Direct-Answer Question Answering? Try ARC-DA, the Direct-Answer AI2 Reasoning Challenge"
9. "More Than Reading Comprehension: A Survey on Datasets and Metrics of Textual Question Answering"
10. "Hurdles to Progress in Long-form Question Answering"
11. "A Critical Evaluation of Evaluations for Long-form Question Answering"

A.2 Prompts for different stages of the Workflows

1. Motivation Extraction Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher:{proposal}

Human Message:
Describe in a bulleted list what is not addressed in the current literature which serves as the Motivation behind solving the above research problem proposed in the Proposal. Answer without a heading line and just the bullet points. Each bullet should mention one gap in the literature as a bullet point and not a sentence.

2. Motivation Question Generation Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher:{proposal}

Human Message:
Describe in a bulleted list what is not addressed in the current literature which serves as the Motivation behind solving the above research problem proposed in the Proposal. Answer without a heading line and just the bullet points. Each bullet should mention one gap in the literature as a bullet point and not a sentence.

AI Message:
{motivation}

Human Message: Convert each of the above bullets in to a binary question. The question should begin with 'Is the research paper'.

3. Ask Question for Motivation Validation Prompt

System Message:
You are a researcher. You have been given a context, which are paragraphs from a research paper. You have been given a question. Answer the given Question in 'Yes' OR 'No' OR 'Unanswerable'. Answer solely based on the provided context of the research paper. If the question can not be answered with the facts mentioned in the available context or there is any ambiguity in answering the question answer as 'Unanswerable'.
Answer as 'Yes' only when the question can be very clearly answered considering the facts in the research paper provided in the context. Do not repeat the question as the part of the answer.
Provide a concise explanation about how the answer to the question is 'Yes' mentioning the paragraphs used in the context to answer it as 'Yes'. If the answer is 'No' or 'Unanswerable' only output that with NO description or elaboration.

Human Message:
Question: {question}
Research Paper Context: {paper_chunks}

4. Extract Limitation Prompt

System Message:
You are a researcher. You have been given the following proposal: {proposal}

A different research paper provided in the context already addresses the gap mentioned as the motivation behind the proposal.
{descriptions}

Human Message:
Research Paper: {paper_chunks}

Identify the limitations or gaps of this research paper which can serve as the new motivation for the proposal. Provide a bulleted list of limitations, where each bullet is concise. Answer WITHOUT a heading line and just the bullet points.

5. Re-write Research Proposal Prompt

System Message:
You are a researcher and have written a proposal: {proposal}

Human Message:
Re-write the proposal by taking into consideration the mentioned gaps in the current literature as the new motivation behind of the problem defined in the proposal.
Answer in a Single detailed paragraph WITHOUT any bullet points or list.
Gaps in the current literature: {limitations}

6. Research Problem Extraction Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher:
{proposal}

Human Message:
What is the problem solved in the proposal?

7. Similar Problem Generation Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher:
{proposal}

Human Message:
What is the problem solved in the proposal?

AI Message:
{problem_statement}

Human Message:
Give me a bulleted list of a more generalised or similar problems to the problem defined in the proposal. Don't give a heading just the answer in a bulleted list.

8. Sub Problem Generation Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher:
{proposal}

Human Message:
What is the problem solved in the proposal?

AI Message:
{problem_statement}

Human Message:
Provide a bulleted list of sub-problems or sub-tasks involved to solve the problem. Don't give a heading just the answer in a bulleted list.

9. Similar and Sub Problem Question Creation Prompt

Human Message:
{statement}
For the statement given above generate a question to be posed on a research paper to find out if the paper is proposing an approach or method to perform the task defined by the statement. Start the question with: 'Is the research paper proposing an approach or method to'.

10. Methodology Extraction Prompt

System Message:
You are a researcher and trying to answer the question posed on a research paper provided as the context.
Research Paper: {paper_chunks}

Human Message:
Answer the given Question in 'Yes' OR 'No' OR 'Unanswerable'. Answer solely based on the provided context of the research paper. If the question can not be answered with the facts mentioned in the available context or there is any ambiguity in answering the question, answer as 'Unanswerable'. Answer as 'Yes' only when the question can be very clearly answered considering the facts in the research paper provided in the context. Do not repeat the question as the part of the answer. If the answer to the question is 'Yes', provide detailed approach or methodology to perform the task. If the answer is 'No' or 'Unanswerable' only output that with NO description.

Question: {question}

11. Method Synthesis Prompt

System Message:
You are a researcher and have been given a proposal and the research problem the proposal is trying to solve. You have been given the approaches in the literature trying to solve, similar problems and sub problems or sub tasks of the problem defined in the proposal. Your task is to synthesize and propose a possible set of methods or approaches to solve the problem defined in the proposal.
Proposal: {proposal}
Research Problem in the Proposal: {problem}

Human Message:
{method_context}

Based on the above information suggest the top 3 possible methods or approaches to solve the problem defined in the proposal.

Sensemaking of Socially-Mediated Crisis Information

Vrushali Koli, Jun Yuan, and Aritra Dasgupta
Department of Data Science, New Jersey Institute of Technology, USA
{vk435, jy448, aritra.dasgupta} @njit.edu

Abstract

In times of crisis, the human mind is often a voracious information forager. It might not be immediately apparent what one wants or needs, and people frequently look for answers to their most pressing questions and worst fears. In that context, the pandemic has demonstrated that social media sources, like erstwhile Twitter, are a rich medium for data-driven communication between experts and the public. However, as lay users, we must find needles in a haystack to distinguish credible and actionable information signals from the noise. In this work, we leverage the literature on crisis communication to propose an AI-driven sensemaking model that bridges the gap between what people seek and what they need during a crisis. Our model learns to contrast social media messages concerning expert guidance with subjective opinion and enables semantic interpretation of message characteristics based on the communicative intent of the message author. We provide examples from our tweet collection and present a hypothetical social media usage scenario to demonstrate the efficacy of our proposed model.

1 Introduction

During the early months of a crisis, people are not equipped with relevant knowledge about a crisis, such as what has occurred, what steps to take next, etc., and information can keep evolving rapidly. Public officials and crisis responders have often used social media to communicate crisis information (Graham et al., 2015). As witnessed during the pandemic, social media platforms not only shaped people's behavior and opinions but also served as a ground for communicating scientific information about public health.

It is widely acknowledged that conflicting information and claims can confuse the public, leading to counterproductive preventive actions, as seen during the COVID-19 pandemic (Rossmann et al., 2018). To the best of our knowledge there has not been much research focusing on designing or investigating social media sites when people might not know what they need (Jang and Baek, 2019) to navigate an unknown crisis. Some recent work investigates cognitive factors to identify the relationship between crisis type, organization reputation, and sentiments (Eaddy and Jin, 2018; Liu et al., 2016). The work in this domain explores distinct factors like forgiveness, empathy, anxiety etc, (Kim and Yang, 2009). Nonetheless, they do not focus on how people perceive crisis information and guide their decisions based on sensemaking (Stieglitz et al., 2017) and contextualization. Research in crisis communication suggests that effective communication requires an understanding of how different people perceive the messages, and what the fundamental drivers are for their information-seeking needs.

In our work, we take inspiration from the crisis communication literature for analyzing the different information facets that are needed by lay social media users to make sense of an unfolding, uncertain situation. Our work offers insight into the generalizability of crisis information-seeking characteristics. We contribute a crisis-related intent classification model that is eventually integrated into a human-AI interface to help social media users triage and group relevant information without being exposed to unnecessary noise and negativity. We demonstrate the efficacy of the model by describing a hypothetical sensemaking workflow of a social media user that leverages our proposed model and interfaces.

2 Related Work

We discuss the related work regarding two overlapping threads of research: i) socially mediated crisis communication and ii) AI-driven sensemaking using social media interfaces.

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 74–81
June 21, 2024 ©2024 Association for Computational Linguistics

2.1 Crisis Communication and Social Media

Studies have shown that the initial crisis stage is a critical period where informing people about protective behaviors supports a resilient crisis response (Islam et al., 2023; Bukar et al., 2022b,a). The use of social media bridged the gap between public officials and general public by providing a real-time communication platform.

The role of social media users has become more visible and active leading to collaborative crises response with public officials (Reuter and Kaufhold, 2018). As such, social media users rely on various sources during a crisis (Islam et al., 2023; Meadows et al., 2019). Current research on social media and crisis communication focuses on identifying influential personnel who can allow for efficient information dissemination.

Crisis communication aims to provide public with credible sources of information during the unfolding of a crisis (Lin et al., 2016). The information from trustworthy sources can help to curb the propagation of rumors (Aguirre and Tierney, 2001). Most messages on social media can be categorized as threat/risk messages or perceived severity (Myneni et al., 2023; Islam et al., 2023). Messages that include both emotional appeal and message source can impact how people perceive severity (Vaala et al., 2022). High engagement messages that are emotionally charged can affect the general public's response (Naseem et al., 2021). These messages help understand the impact of a crisis. However, the interplay among different information dimensions, like engagement, awareness, and cognitive load, remains unclear and is an active area of research (Stieglitz et al., 2017) that we contribute to.

2.2 AI-Driven Sensemaking of Socially Mediated Information

AI models are often used for distinguishing between facts and opinions on social media. The current work in fact-checking focuses on the automated classification of social media content using supervised learning algorithms. While these research studies present a fundamental approach towards identifying and solving the problem of check-worthy claim identification (Miranda et al., 2019; Hassan et al., 2017), the focus is only on binary classification tasks of accepting or refuting the claims (Hanselowski et al., 2019; Nakov et al., 2021). From the information consumers' perspective, fact-checking tools are out of user control as external sources provide them. Moreover, during a pandemic, users are generally navigating terra incognita as there is no establishment of ground truth that can be automatically detected. Users often want to be self-reliant and not completely rely on third-party fact-checking sites (Myneni et al., 2023). Currently, social media platforms allow end-users to curate their information feed by allowing users to filter what types of content they are exposed to. Studies distinguish between actions users can take to moderate content based on source (specific users) and types of content (Jhaver et al., 2023). Reducing types of content that are not exposed to the user can aid their information-seeking process by reducing the search space (Gillespie, 2022). Lack of transparency in algorithmic details of provided methods and unclear definitions of contextual terms can lead to further confusion. To address these challenges, we propose to give social media users the agency and control their feed while also carefully considering the role of AI-driven automation in triaging information that users might need but not necessarily be aware of owing to the uncertain information landscape.

3 Methodology

In this section, we describe our methodology for data collection and qualitative labeling of tweets that preceded the conceptualization of communicative intent. Please find details about the methodology in the supplemental material: `https://tinyurl.com/mrymxwed`.

Data Collection: We used the erstwhile Twitter User Timeline API (Hossain et al., 2018) for collecting tweets from March 2020 to September 2020. In particular, we wanted to collect tweets in two batches: one, focused on identifying regular social media users who could also be considered subject matter experts, and two, tweets from lay social media users. With the help of a published list from Elemental, a health and wellness publication, we collected the relevant COVID-19 tweets from a list of 50 health and science experts (Editors, 2020) who regularly updated information about the COVID-19 pandemic. With the aid of a researcher in the medical sciences domain, we verified that these people could be considered credible voices about the pandemic while acknowledging that there could be differences of opinion among experts. The second category of tweets in our collection is general

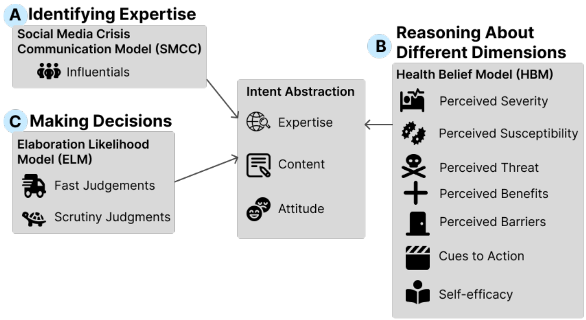


Figure 1: Illustrating the derivation of our Intent Model from the existing Crisis Communication Models (SMCC, ELM, and HBM).

user tweets. These come from users except these 50 experts over the same period. For both categories, expert and general users' tweets, we extract tweets using COVID-19-relevant keywords some of which include "coronavirus", "sars-cov-2", and "covid-19" and several others. We needed to ensure our general user tweet collection did not contain any tweets from users who could be considered experts. We ensured that general public tweets excluded tweets from users whose profiles included the keywords "epidemiologist", "virologist", "clinician", etc.

Topic Modeling. We anchored our analysis to understanding how experts and general users could well be discussing different dimensions of the pandemic. We leveraged publicly available deep-learning models with good performance on tweets based on BERT (Devlin et al., 2018) architecture for consistency across each information dimension. We use a transformer-based algorithm for topic modeling called BERTopic (Grootendorst, 2020). We trained the model on a subset of 5,000 expert tweets using c-TF-IDF to reduce outliers. Then, two graduate students labeled approximately 900 raw topic clusters to 25 topic categories using references from prior work on topic modeling covid-19 tweets (Oliveira et al., 2022; Vijayan, 2021; Lyu et al., 2021; Boon-Itt et al., 2020; Abd-Alrazaq et al., 2020). These topics include governmental affairs, vaccine development, scientific information, healthcare, mitigation, symptoms, etc.

Subjectivity and Sentiment Analysis. Subjectivity prediction can help consumers evaluate the text for more effective and efficient scientific communication. We, therefore collaborated with industrial researchers for using professional labeling services to tag 10,000 tweets from our corpus. The available labels were "objective, slightly objective, uncertain, slightly subjective, subjective, Irrelevant". We split the resulting data set into 3,232 for training and the remaining 808 samples for test tweets and trained a DistilBERT (Sanh et al., 2019) model for this task. Tweets labelled "uncertain" or "Irrelevant" were removed. We fine-tuned a pre-trained model to classify two labels. The total training time for the model is approximately one hour with the GPU-enabled Google Colab in the free tier. We expected expert messages to be more objective and general users to be more subjective. However, we found that experts also exhibited subjectivity in their tweets like general users across most topics.

For sentiment analysis, we chose the model bertweet-base-sentiment-analysis (Pérez et al., 2021) provided by HuggingFace. The model classification results in each tweet with three probabilities corresponding to positive, negative, and neutral. We use the label with the highest probability as the final label for the tweet.

4 Sensemaking via Communicative Intent

Communicative intent, or simply intent, refers to the aim or purpose of a tweet. Intent analysis can help information consumers determine whether the tweet is relevant to what they are seeking as it can provide contextual information about a particular topic. The message intent can be considered during the reasoning process (Monti et al., 2022) which can aid users in navigating the information space. Unprecedented emergencies, like the pandemic, require the public to adapt to time, domain, and context-specific information in understanding the communication dynamics on social media. Current crisis communication models are insufficient in guiding people when exposed to exponentially more information due to increased social media use. The intent classification model is shown in Figure 1. The related models to our work are the Health Belief Model (HBM), the Social Mediated Crisis Communication Model (SMCC), and the Elaboration Likelihood Model (ELM).

HBM states that an individual's personal beliefs affect their health-related behaviors (Washburn, 2020). This is a valuable framework to characterize people's discussions on social media based on perceived severity, perceived threat, perceived susceptibility, etc. However, this model does not give us a way to characterize why people perceive certain information in a particular way or what constitutes a threat. In the Intent Model, source,

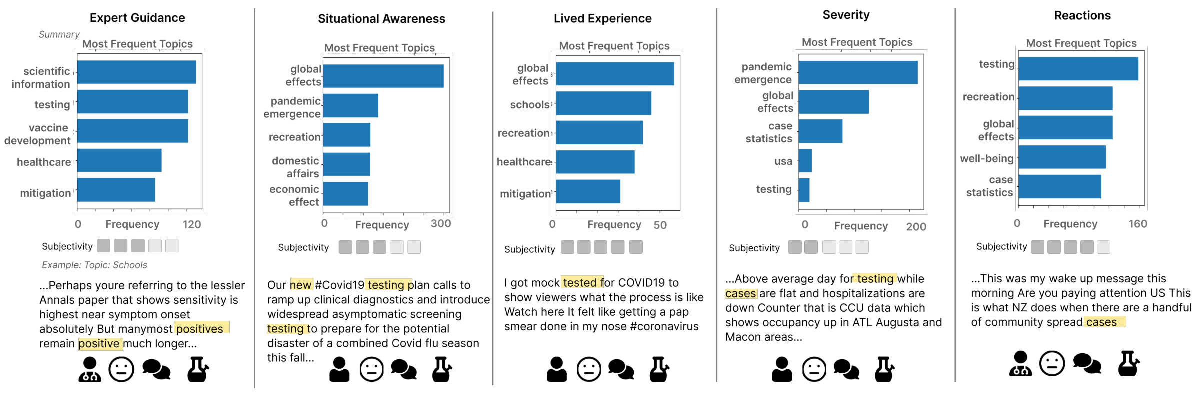


Figure 2: A snapshot of each intent category (1) An overview of each intent category showing the five most frequent topics and average subjectivity (2) An example tweet profiled by sentiment, subjectivity, and topic from both experts and general users.

content, and attitude are ways to characterize who, what, and how information is perceived under a particular category (Figure 1.2).

The SMCC model offers a solution by characterizing the source of information. The SMCC model emphasizes identifying individuals that can aid in information dissemination efforts, classifying these individuals as *influentials* (Liu et al., 2020). In social media sites, these individuals have high engagement by having a large following. However, these individuals may not be crisis experts, and other social media users may not want to see information from this individual. In the Intent Model this is characterized by source of information where we make a distinction between *crisis experts* and *general users* (Figure 1.1). When individuals make decisions based on popular information and not credible information, they are not engaging in elaborate reasoning (i.e. taking the time to think through what they've read). ELM considers that information seekers make judgments in two ways: (1) fast with simple reasoning and (2) slow with elaborate reasoning (Petty Richard and Cacioppo, 1986).

Since social media sites provide information in a user-friendly interface, information seekers are continuously tempted to make fast judgments based on limited information. Additionally, the ELM model doesn't provide ways to characterize the interplay between information dimensions and types of decisions. Social media sites provide content moderation methods that help users curate their information feed, which can aid in finding relevant scientific information. This can also aid in supporting elaborate thinking. However, if users deem these methods unreliable, they will not engage in content moderation increasing the likelihood of information overload. The Intent Model can aid in characterizing what types of information people seek during a crisis for different types of decisions (Figure 1.3).

Using our intent model, an information seeker can triage information based on different intent categories. For each intent category, additional details are accessible to the user via the source, content, and attitude or the messages. The message's source is defined as an expert or general user and the degree of subjectivity. The content and attitude refer to the topic and sentiment of the message. Most content analysis studies generated a labeling guide based on previous literature reviews, such as guiding principles for classifying social media news articles (MacKay et al., 2021), informativeness (whether a tweet contains relevant information or not) (Olteanu et al., 2015), an existing crisis communication model like HBM (Myneni et al., 2023). We chose to follow this approach for our labeling guide. During information-seeking behavioral patterns, consumers pay attention to the message source, So two graduate students took into account the source of the message (expert/general user) and the message content to determine the message's intent. We classified 6,844 tweets into five intent categories: (i) Expert Guidance, (ii) Situational Awareness (iii) Severity (iv) Reactions, (v) Lived Experience. Figure 2 provides a topic distribution and an example tweet in each intent category with similar uncertainty profiles to emphasize the differences across each intent category.

Expert Guidance categorizes tweets from experts that are providing some suggestions or recommendations to address the pandemic (21% of sample) (Wang et al., 2021; Brady et al., 2023; Ehrmann

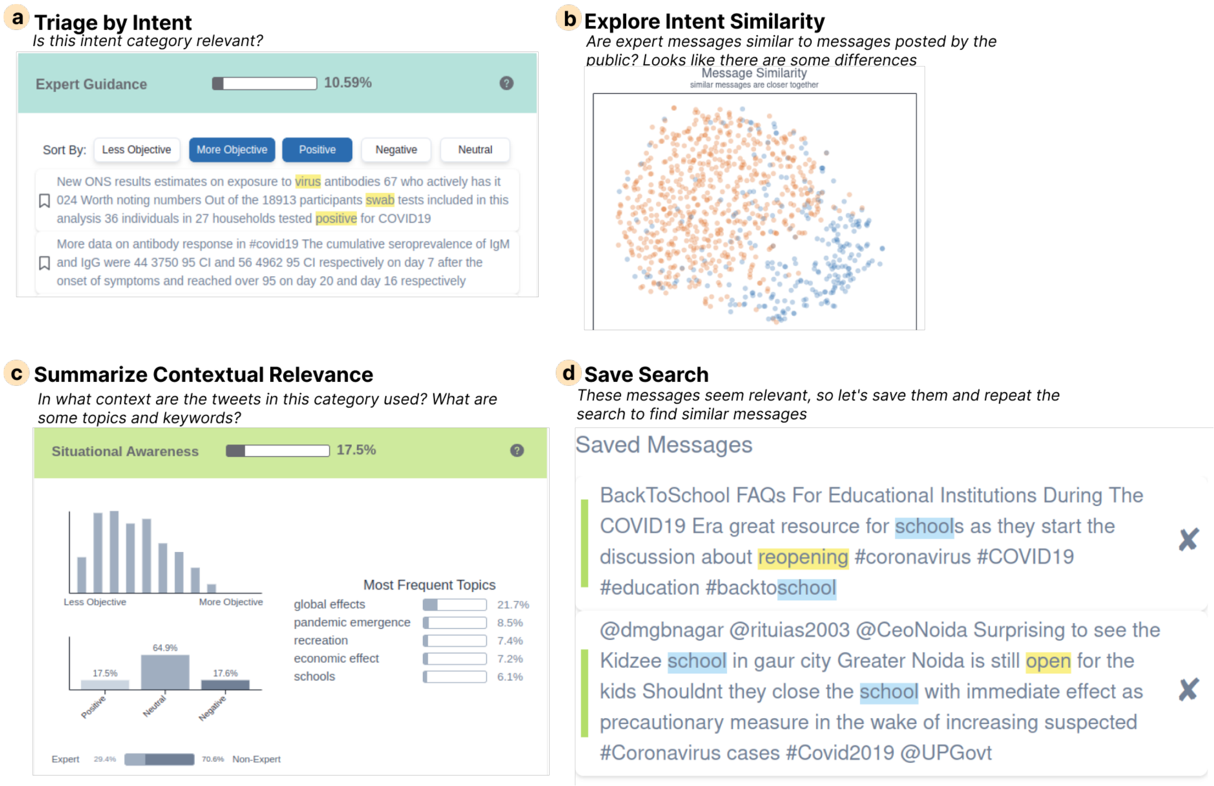


Figure 3: A hypothetical workflow of a social media user leveraging the intent classification model for sensemaking during the pandemic.

and Wabitsch, 2022). The most frequent topics for this category are scientific information, testing, and vaccine development. *Situational Awareness* are tweets from any source that are providing updates on news alerts, business proceedings, and policy revisions (33% of sample) (MacKay et al., 2021; Myneni et al., 2023). The most frequent topics for this category are global effects, pandemic emergence, and recreation. *Severity* describes tweets that qualitatively or quantitatively report on the impact of the pandemic (11% of sample) (Myneni et al., 2023). The most frequent topics for this category are pandemic emergence, global effects, and case statistics. *Reactions* are tweets from any source that describe emotions, comments, or responses towards events caused by the pandemic (30% of sample) (Wang et al., 2021). The most frequent topics for this category are testing, recreation, and global effects. *Lived Experience* tweets describe direct personal experiences with specific details regarding location and time (8% of sample) (Wang et al., 2021). The most frequent topics for this category are global effects, schools, and recreation.

A DistilBERT (Sanh et al., 2019) **classification model** was trained on the 6844 labeled tweets until overall accuracy reached approximately 70%, indicating good performance. The accuracy per intent category: (1) Expert Guidance 69%, (2) Reactions 74% (3) Lived Experience 80% (4) Severity 93% (5) Situational Awareness 74%. These results on real-world data indicated that our model could also be used to learn from user interactions. Current content moderation methods do not allow users to update the underlying model. However, for machine-guided social media systems to better address the changing needs of information consumers, users need to update the underlying models to match their mental model of changing information. The Intent Model allows users to update each message's intent to one that aligns more with the user's mental model or curate their intent category based on the source, content, and attitude of the messages.

5 Sensemaking during a Crisis

We present an AI-driven sensemkaing scenario during the COVID-19 pandemic to demonstrate the utility of our proposed intent classification model (ICM). We integrated ICM (a predictive model trained on our collection of tweets) with a web-based interface, which allows lay information consumers to triage credible and potentially actionable information. Let us see how this interface can be used by Mary, a parent whose child currently studies at home due to COVID-19-affected school closing. She needs to decide if she should reduce her working hours and invest in homeschooling. Figure 3 shows Mary's workflow as she interacts with our interface. From her experience, she knows that not all tweets are reliable and that topics provided by Twitter do not give her sufficient control to navigate the relevant messages. She uses the interface to triage the relevant information using *Situational Awareness* as the intent category of interest (Figure 3.a). Additionally, she selects all intent categories to learn more about the contrasting messages by reviewing the representative tweets in each category. She finds that by using the intent categories and the highlighted words she can gradually make sense of the different facets of the conversation related to the pandemic that might or might not be related to her decision-making goal. She can easily filter out messages from categories she is not interested in such as Reactions and Severity.

To find more contextually relevant messages, Mary filters by keyword *school* and chooses to summarize the overall message profile and observes the patterns across sentiment, subjectivity, expertise, and topics across each intent category (Figure 3.c). She notices that Situational Awareness tweets are from experts and lay users and that *schools* is a frequent topic. She notices that messages from experts and general users have distinct groupings which tell her they may have different opinions (Figure 3.b). Interestingly, she finds the helpful messages discussing distance learning and back-to-school efforts come from lay users because of their lived experience. It did not occur to her that online learning could be a possible solution for her child to continue learning. She found tweets suggesting reopening in-person learning once proper safety measures are determined. She saves those messages and repeats her search to find even more similar tweets (Figure 3.d).

She also decided to filter the selection criteria only to consider tweets from the *schools* topic. As she expected, applying this filter allowed her to see more messages about other parents' experiences with both homeschooling and online learning. After reviewing her saved tweets, she decided that online learning is a viable option. Since Mary has a day job she needs to report to, online learning for her child would allow Mary to continue working and her child to continue receiving an education. By leveraging the intent model, Mary could quickly reduce her exposure to irrelevant tweets. She could also assess the credibility of the messages critically and have sufficient control over the information she needed to know.

6 Conclusion and Future Work

In this paper, we present a crisis-related intent classification model and present its utility via examples and usage scenarios using the COVID-19 pandemic as an example. We developed and trained classification models for sentiment, subjectivity, and topic to further our understanding of how experts and general users communicate during the initial stages of a crisis. We use Twitter messages as the basis of our analysis to profile the information uncertainty and address the need for principled approaches towards sensemaking of socially mediated information during a crisis. We are currently developing a software prototype, demonstrated in Section 5, that allows lay users to explore messages using the intent of a message author and control their exposure to crisis-relevant information, by focusing on what they would need to address their pressing questions.

Future work would consider a user study with diverse participants to understand how users perceive the usefulness and utility of our proposed sensemaking workflow and the resulting human-machine interface in their information-seeking processes. We intend to build upon and further develop the web interface, taking into account previous studies on end-user content moderation techniques.

References

Alaa Abd-Alrazaq, Dari Alhuwail, Mowafa Househ, Mounir Hamdi, and Zubair Shah. 2020. Top concerns of tweeters during the covid-19 pandemic: infoveillance study. *Journal of medical Internet research*, 22(4):e19016.

Benigno E Aguirre and Kathleen J Tierney. 2001. Testing shibutani's prediction of information seeking behavior in rumor.

Sakun Boon-Itt, Yukolpat Skunkan, et al. 2020. Public perception of the covid-19 pandemic on twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, 6(4):e21978.

William J Brady, Joshua Conrad Jackson, Björn Lindström, and MJ Crockett. 2023. Algorithm-mediated social learning in online social networks. *Preprint at OSF preprints. https://doi. org/10.31219/osf. io/yw5ah*.

Umar Ali Bukar, Marzanah A Jabar, Fatimah Sidi, Rozi Nor Haizan Nor, Salfarina Abdullah, and Iskandar Ishak. 2022a. Mobile social media use in flooding emergencies: The impact of information seeking and sharing through social media crisis communication and resilience model. In *2022 Applied Informatics International Conference (AiIC)*, pages 51–56. IEEE.

Umar Ali Bukar, Fatimah Sidi, Marzanah A Jabar, Rozi Nor Haizan Binti Nor, Salfarina Abdullah, and Iskandar Ishak. 2022b. A multistage analysis of predicting public resilience of impactful social media crisis communication in flooding emergencies. *IEEE Access*, 10:57266–57282.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

LaShonda Louallen Eaddy and Yan Jin. 2018. Crisis history tellers matter: The effects of crisis history and crisis information source on publics' cognitive and affective responses to organizational crisis. *Corporate Communications: An International Journal*.

Elemental Editors. 2020. Six months in 50 experts to trust in a pandemic.

Michael Ehrmann and Alena Wabitsch. 2022. Central bank communication with non-experts–a road to nowhere? *Journal of Monetary Economics*, 127:69–85.

Tarleton Gillespie. 2022. Do not recommend? reduction as a form of content moderation. *Social Media+ Society*, 8(3):20563051221117552.

Melissa W Graham, Elizabeth J Avery, and Sejin Park. 2015. The role of social media in local government crisis communications. *Public Relations Review*, 41(3):386–394.

Maarten Grootendorst. 2020. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214*.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812.

Md Arafat Hossain, Steve Versteeg, Jun Han, Muhammad Ashad Kabir, Jiaojiao Jiang, and Jean-Guy Schneider. 2018. Mining accurate message formats for service apis. In *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 266–276. IEEE.

Khairul Islam, Najma Akhther, and Matthew W Seeger. 2023. Variability in media content of public health outbreak coverage: A crisis communication approach. *Communication Studies*, 74(2):113–130.

Kyungeun Jang and Young Min Baek. 2019. When information from public health officials is untrustworthy: The use of online news, interpersonal networks, and social media during the mers outbreak in south korea. *Health communication*, 34(9):991–998.

Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–33.

Hye Kyung Kim and Sung-Un Yang. 2009. Cognitive processing of crisis communication: Effects of csr and crisis response strategies on stakeholder perceptions of a racial crisis dynamics. *Public Relations Journal*, 3(1):1–39.

Xialing Lin, Patric R Spence, Timothy L Sellnow, and Kenneth A Lachlan. 2016. Crisis communication, learning and responding: Best practices in social media. *Computers in human behavior*, 65:601–605.

Brooke Fisher Liu, Julia Daisy Fraustino, and Yan Jin. 2016. Social media use during disasters: How information form and source influence intended behavioral responses. *Communication Research*, 43(5):626–646.

Brooke Fisher Liu, Yan Jin, Lucinda Austin, Erica Kuligowski, and Camila Espina Young. 2020. 14 the social-mediated crisis communication (smcc) model. *Advancing Crisis Communication Effectiveness: Integrating Public Relations Scholarship with Practice*, page 63.

Joanne Chen Lyu, Eileen Le Han, and Garving K Luli. 2021. Covid-19 vaccine–related discussion on twitter: topic modeling and sentiment analysis. *Journal of medical Internet research*, 23(6):e24435.

Melissa MacKay, Taylor Colangeli, Daniel Gillis, Jennifer McWhirter, and Andrew Papadopoulos. 2021. Examining social media crisis communication during early covid-19 from public health and news media for quality, content, and corresponding public sentiment. *International Journal of Environmental Research and Public Health*, 18(15):7986.

Charles W Meadows, Cui Zhang Meadows, Lu Tang, and Wenlin Liu. 2019. Unraveling public health crises across stages: Understanding twitter emotions and message types during the california measles outbreak. *Communication Studies*, 70(4):453–469.

Sebastião Miranda, David Nogueira, Afonso Mendes, Andreas Vlachos, Andrew Secker, Rebecca Garrett, Jeff Mitchel, and Zita Marinho. 2019. Automated fact checking in the news room. In *The World Wide Web Conference*, pages 3579–3583.

Corrado Monti, Luca Maria Aiello, Gianmarco De Francisci Morales, and Francesco Bonchi. 2022. The language of opinion change on social media under the lens of communicative action. *Scientific Reports*, 12(1):17920.

Sahiti Myneni, Paula Cuccaro, Sarah Montgomery, Vivek Pakanati, Jinni Tang, Tavleen Singh, Olivia Dominguez, Trevor Cohen, Belinda Reininger, Lara S Savas, et al. 2023. Lessons learned from interdisciplinary efforts to combat covid-19 misinformation: Development of agile integrative methods from behavioral science, data science, and implementation science. *JMIR infodemiology*, 3(1):e40156.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769*.

Usman Naseem, Imran Razzak, Matloob Khushi, Peter W Eklund, and Jinman Kim. 2021. Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis. *IEEE transactions on computational social systems*, 8(4):1003–1015.

Francisco Braulio Oliveira, Amanul Haque, Davoud Mougouei, Simon Evans, Jaime Simao Sichman, and Munindar P Singh. 2022. Investigating the emotional response to covid-19 news on twitter: a topic modeling and emotion classification approach. *IEEE Access*, 10:16883–16897.

Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 994–1009.

E Petty Richard and John T Cacioppo. 1986. Communication and persuasion: Central and peripheral routes to attitude change.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Christian Reuter and Marc-André Kaufhold. 2018. Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics. *Journal of contingencies and crisis management*, 26(1):41–57.

Constanze Rossmann, Lisa Meyer, and Peter J Schulz. 2018. The mediated amplification of a crisis: Communicating the a/h1n1 pandemic in press releases and press coverage in europe. *Risk analysis*, 38(2):357–375.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Stefan Stieglitz, Milad Mirbabaie, Lara Schwenner, Julian Marx, Janina Lehr, and Felix Brünker. 2017. Sensemaking and communication roles in social media crisis communication.

Sarah E Vaala, Matthew B Ritter, and Deepak Palakshappa. 2022. Experimental effects of tweets encouraging social distancing: Effects of source, emotional appeal, and political ideology on emotion, threat, and efficacy. *Journal of Public Health Management and Practice*, 28(2):E586–E594.

Ranjit Vijayan. 2021. Teaching and learning during the covid-19 pandemic: A topic modeling study. *Education Sciences*, 11(7):347.

Yan Wang, Haiyan Hao, and Lisa Sundahl Platt. 2021. Examining risk and crisis communications of government agencies and stakeholders during early-stages of covid-19 on twitter. *Computers in human behavior*, 114:106568.

Lisa Washburn. 2020. Understanding the health belief model. *University of Tennessee: Knoxville, TN, USA*.

Blind Spots and Biases: Exploring the Role of Annotator Cognitive Biases in NLP

Sanjana Gautam
Information Sciences and Technology
Pennsylvania State University, USA
sanjana.gautam@psu.edu

Mukund Srinath
Information Sciences and Technology
Pennsylvania State University, USA
mukund@psu.edu

Abstract

With the rapid proliferation of artificial intelligence, there is growing concern over its potential to exacerbate existing biases and societal disparities and introduce novel ones. This issue has prompted widespread attention from academia, policymakers, industry, and civil society. While evidence suggests that integrating human perspectives can mitigate bias-related issues in AI systems, it also introduces challenges associated with cognitive biases inherent in human decision-making. Our research focuses on reviewing existing methodologies and ongoing investigations aimed at understanding annotation attributes that contribute to bias.

1 Introduction

With the recent rapid expansion of generative AI models, we have witnessed their numerous benefits and the emergence of substantial ethical concerns (Thoppilan et al., 2022; Rudolph et al., 2023). There has been an influx of remarkable and noteworthy work that describes the issues of fairness, toxicity and bias in the text generation process (Bender et al., 2021; Abid et al., 2021; Seaborn and Kim, 2023). Several models are deployed as real-world solutions with a lack of informed consideration of their social implications, especially in sensitive fields such as healthcare, journalism, law, and finance (Khowaja et al., 2023). Recent research has revealed that these language models can mimic human biases present in language, perpetuating prejudiced behaviour that dehumanizes certain socio-demographic groups by deeming them more negative or toxic (Havens et al., 2022; Blodgett et al., 2020).

One of the proposed solutions to this issue has been to introduce human annotators to label the training corpora or validate pre-labelled datasets and manually remove toxic (or biased) data entries (Havens et al., 2022; Cabrera et al., 2014). It is common practice for machine learning systems to rely on crowd-sourced label data for training and evaluation (Wu et al., 2022). It is also well-known that biases present in the label data can induce biases in the trained models (Hettiachchi et al., 2021). Therefore, while humans-in-the-loop for model training may seem like an intuitive solution, it often introduces additional biases due to inherent cognitive biases in humans (Parmar et al., 2022). Crowdwork annotation studies conducted on MTurk (and other crowdwork platforms) where the participants come from a specific demographic population can potentially perpetuate populist viewpoints (Reinecke and Gajos, 2015).

Prior work has established the pitfalls in human rationality, as influenced by the lived experiences and environment, which Herbert Simon termed bounded rationality (Simon, 1957). Human biases have been identified to be the resulting gap between rational behaviour and heuristically determined behaviour (Tversky and Kahneman, 1974; Bojke et al., 2021). Over 180 cognitive biases have been identified, spawning everything from social interaction to judgment and decision-making with research spanning over 70 years (Talboy and Schneider, 2022). These tendencies or patterns can lead to faulty reasoning, irrationality, and potentially detrimental outcomes.

Bias sometimes emerges due to distractions, lack of interest, or laziness among annotators regarding the annotation task, leading them to select inaccurate labels. However, more concerning is the label bias stemming from informed and well-intentioned annotators who consistently exhibit disagreement (Hovy and Prabhumoye, 2021). Plank et al. (2014) demonstrated that this form of bias emerges when multiple correct labels are possible. For instance, the term 'social media' can be legitimately interpreted either as a noun phrase consisting of an adjective and a noun or as a noun compound comprising two nouns. For example, Sap et al. (2019) demonstrated that these biases mirror social and de-

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 82–88
June 21, 2024

mographic variances. For instance, annotators tend to evaluate utterances from various ethnic groups disparately and may misinterpret harmless banter as hate speech due to their unfamiliarity with the communication norms of the original speakers.

Merely relying on a few gold-standard corpora as training datasets or debiasing datasets is not a sustainable long-term strategy since languages undergo constant evolution. Thus, even a comprehensive sample can only encapsulate a momentary snapshot, offering at best a transient solution (Fromreide et al., 2014). We believe the design and set-up of the crowd work task plays a pivotal role in determining the goodness of data. In this work, we look at bias-diminishing strategies and identify the pressing questions in this area. Our central goal is to show that there is a need for standardized design principles when it comes to designing crowdwork studies. Specifically, we concentrate on the need for an HCI perspective in natural language processing research.

2 Bias in AI Models

Generative AI's propensity to amplify existing biases and create new ones has attracted considerable attention across a range of communities, including academics, policy-makers, industry, and civil society. Much of the initial work focused on developing quantitative definitions of fairness (Dwork et al., 2012; Hardt et al., 2016; Joseph et al., 2016; Liu et al., 2017; Verma and Rubin, 2018), and various technical methods for 'debiasing' AI models (Agarwal et al., 2018; Bolukbasi et al., 2016; Friedler et al., 2014; Zafar et al., 2017). When referring to de-biasing, we use the definition *'removing undesired skews in the data and the model outcome, such as by equalising a metric of interest between groups'*. *"Unintended bias"* is used to describe the different sources of bias that are introduced throughout an AI development life cycle (Lee and Floridi, 2021; Suresh and Guttag, 2021), focusing on not just the bias introduced, but also the harm it causes (Crawford, 2016).

Recent studies have shifted focus from merely identifying sources of bias in AI, such as flawed data collection methods, to exploring the various harms caused by these biases. This shift is supported by interdisciplinary research that highlights the contextual nature of fairness. Factors such as regional and cultural differences in lived experiences significantly influence perceptions of fairness, revealing that certain algorithmic behaviours may only be deemed harmful in specific social or cultural contexts (Green and Hu, 2018; Lee and Singh, 2021; Sambasivan et al., 2021; Selbst et al., 2019). Given these complexities, it is broadly acknowledged that eliminating bias or ensuring absolute fairness in AI systems is unfeasible (Kleinberg et al., 2016; Mehrabi et al., 2021; Pleiss et al., 2017). Instead, the objective is to minimize fairness-related harms and other adverse impacts to the greatest extent possible (Mehrabi et al., 2021; Selbst et al., 2019; Sun et al., 2019). This perspective is further enhanced by recent interdisciplinary studies (Lewicki et al., 2023), which underscore the nuanced and multifaceted nature of fairness in AI.

Identifying and acknowledging systemic biases in data collection is a crucial step in mitigating their impact on the systems that are trained using this data, and is a critical prerequisite for achieving fairness in algorithmic decision-making (Hajian et al., 2016). While humans are integral to the system, participating in data collection and various phases thereof, it is imperative to emphasize that human computation (Quinn and Bederson, 2011), the practice of harnessing human intelligence and cognitive abilities as computational elements, holds potential for addressing and mitigating these challenges.

3 Cognitive Bias among Annotators

As emphasized by Van Dis et al. (2023), ensuring human accountability is essential in scientific practice. The history of Large Language Models (LLMs) has shown that they can produce inaccurate information, or "hallucinations." To guarantee the accuracy of information, it is necessary to implement a rigorous verification and fact-checking process led by experts. Consequently, the discourse highlights the critical need for accountability in human-in-the-loop systems, particularly in response to the new challenges posed by these systems.

The importance of understanding and mitigating biases in crowd data is highly relevant to researchers, and others who rely on crowd data for creating automated systems. Prior work has explored various approaches to promoting fairness in machine learning, including the direct utilization of crowdsourced data (Balayn et al., 2018), leveraging crowds to assess perceived fairness of features (Van Berkel et al., 2019, 2021), applying pre-processing techniques such as removing sensitive attributes, resampling data to remove discrimination, and iteratively adjusting training weights for sensitive groups (Calmon et al., 2017; Kamiran and Calders, 2012; Krasanakis et al., 2018), as well as employing active learning methods (Anahideh et al., 2022).

The use of crowdsourcing for tasks such as data annotation can inadvertently introduce cognitive biases, stemming from the inherent design of the task itself. We have identified three primary reasons why annotated data can be problematic: (1) Unethical spammers submit imprecise or even arbitrary labels in order to maximize their financial advantage (Eickhoff et al., 2012) or due to external distractions. (2) Unqualified workers are, despite their best efforts, unable to produce an acceptable annotation quality (Eickhoff, 2014). (3) Malicious workers purposefully aim to undermine or influence the labelling effort (Wang et al., 2013). However, we propose that there might be some factors that have not been uncovered in prior literature. Crowd-workers have their tasks cut out for them, in cases where the nature of task design causes the propagation of bias. Research on crowd work has often focused on task accuracy whereas other factors such as biases in data have received limited attention (Hettiachchi et al., 2021).

Cognitive biases originate from individuals' own *"subjective social reality"* which is often a product of lived experiences. This makes cognitive bias a deviation from the rationality of judgement, therefore it may consist of perceptions of other people that are often illogical (Martie et al., 2005). An individual's construction of social reality, instead of the objective input, may dictate their behaviour and lead to perceptual distortion, inaccurate judgment, illogical interpretation, or irrationality (Bless and Fiedler, 2014). Past work has demonstrated that cognitive bias can affect crowdsourced labour and lead to significantly reduced result quality. This performance detriment is subsequently propagated into system ranking robustness and machine-learned ranker efficacy (Eickhoff, 2018).

The annotation instructions provided to crowdworkers can inadvertently prime them to exhibit biases towards or against specific domain information, which can be exacerbated by poorly designed instructions. Furthermore, annotators are often not fully informed about the true purpose of the research, leading to an ambiguity effect that can make the decision-making process appear more challenging and less appealing due to the limited information available (Ellsberg, 1961). Additionally, the phased revelation of information to annotators can result in an anchoring effect, where certain pieces of information are given disproportionate attention based on the timing of their disclosure. This underscores the importance of designing annotation studies that mitigate cognitive biases among workers, ensuring that the annotation process is fair, transparent, and unbiased.

4 Crowd Control

Humans in the loop bring a lot of value to generative AI and AI systems. Therefore, the solution to the issue of cognitive bias cannot be to remove the annotators from the system. Human annotators often bring expert judgements, that are valuable in creating ground truth labels. For example, annotation of medical imagery cannot be performed without the help of annotators who are medical professionals. Expert guidance, lived experiences and proximity to the problem domain make human annotators irreplaceable in the AI-training life-cycle. The common strategies of accounting for biases of annotators by employing qualification tests, demographic filters, incentives, and sophisticated worker models may not be enough to overcome this source of noise. There is therefore a need to control the annotation task design settings, to minimize the introduction of biases due to the cognitive biases of annotators. While cognitive biases and their effects on decision-making are well-known and widely studied, we note that AI-assisted decision-making presents a new decision-making paradigm. It is important to study their role in this new paradigm, both analytically and empirically.

Crowdwork platforms are often designed to position crowdworkers as interchangeable (Irani and Silberman, 2013). While some forms of digital work can be decomposed and distributed, the presumption that all crowdsourced dataset annotators exercise near-identical capacities of perception and judgement ignores the fact that social position, identity, and experience shape how annotators' actions.

Previous research has highlighted the significance of the annotator population and the power dynamics inherent in platform-mediated crowdwork, both of which can perpetuate cognitive biases (Díaz et al., 2022). Building upon this foundation, we propose a novel framework to enhance transparency and robustness in the process of designing a crowdwork task. This approach holds promise for mitigating the impact of cognitive biases in crowdwork, thereby contributing to more reliable and trustworthy outcomes.

5 Counter-measures for Biases

To minimize bias in NLP annotation tasks, several steps can be implemented. Firstly, recruiting a diverse group of annotators from various backgrounds can help balance individual biases. Providing clear and detailed guidelines ensures uniform understanding across annotators. Training sessions, followed by calibration discussions, align annotator interpretations and reveal guideline ambiguities. An iterative feedback loop allows for regular quality checks and guideline adjustments based on annotator experiences. Measuring inter-annotator agreement with metrics like Cohen's Kappa highlights discrepancies and areas needing clarification. Annotation tasks should be designed to minimize bias, such as by rotating text assignments among annotators to avoid topical biases. Finally, a post-annotation analysis can detect any remaining biases, ensuring the reliability and fairness of the annotated data.

However, biases can arise at any point in the AI lifecycle. It is therefore imperative for researchers to maintain a meticulous approach throughout the entire research process, encompassing various facets such as the selection of appropriate datasets, adherence to annotation schemes or labelling procedures, thoughtful considerations regarding data representation methodologies, judicious selection of algorithms tailored to the task at hand, and rigorous evaluation protocols for automated systems. Furthermore, researchers must consider the tangible real-world applications of their research endeavors. Particularly noteworthy is the imperative to consciously direct efforts towards leveraging technological advancements to uplift and empower marginalized communities, as underscored by Asad et al. (2019). Several studies critique existing bias mitigation algorithms for their lack of effectiveness due to inconsistent study protocols, inappropriate datasets, and over-tuning to specific test sets. To overcome these limitations, research needs to introduce robust evaluation protocol, and sensible metrics designed to evaluate algorithm robustness against various biases (Shrestha et al., 2022).

Our future work derives from the insights presented in the preceding discussion. It posits that the roots of bias within AI systems often traced back to the initial stages of the annotation process, particularly during the instruction phase. Although not all cognitive biases are inherently detrimental, a pressing need exists to advance our comprehension of how to devise annotation studies that align with the principles of human-computer interaction (HCI).

Our objective in this research endeavour is to contribute substantively to the ongoing efforts aimed at mitigating bias in crowd work. We intend to achieve this by focusing on the refinement of study design and instructional strategies. By incorporating insights from the HCI discipline, we aim to cultivate a nuanced understanding of how to create balanced annotation studies that minimize the emergence of bias. Through this work, we aspire to not only shed light on the pivotal role played by the annotation phase in propagating or mitigating bias but also to provide practical recommendations and guidelines for researchers and practitioners engaged in AI development and crowd work.

6 Conclusion

Our research highlights the critical importance of considering annotation attributes that contribute to bias in AI systems. The cognitive biases of annotators, inherent in human decision-making, can perpetuate and even amplify existing social disparities in AI models. To mitigate these issues, a multidisciplinary approach is necessary not only in deploying AI models but also in designing better systems for annotation tasks. By bringing together experts from diverse fields, including human-centered design, ethics, social sciences, law, healthcare, AI/ML, education, communication, and community representation, we can design annotation systems that are more inclusive, transparent, and fair. This collaborative framework is essential for developing annotation tasks that are free from biases, ambiguous, and unclear instructions, and that take into account the complexities of real-world data. Furthermore, a multidisciplinary approach is crucial for deploying AI models that are developed using these annotated data, ensuring that they are fair, transparent, and accountable. By acknowledging the limitations of human annotators and addressing them through a multidisciplinary approach, we can work towards a more equitable digital landscape where AI systems benefit both individuals and society as a whole.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR.

Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. 2022. Fair active learning. *Expert Systems with Applications*, 199:116981.

Mariam Asad, Lynn Dombrowski, Sasha Costanza-Chock, Sheena Erete, and Christina Harrington. 2019. Academic accomplices: Practical strategies for research justice. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, pages 353–356.

Agathe Balayn, Panagiotis Mavridis, Alessandro Bozzon, Benjamin Timmermans, and Zoltán Szlávik. 2018. Characterising and mitigating aggregationbias in crowdsourced toxicity annotations. In *Proceedings of the 1st Workshop on Disentangling the Relation between Crowdsourcing and Bias Management*. CEUR.

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*, pages 610–623.

Herbert Bless and Klaus Fiedler. 2014. *Social cognition: How individuals construct social reality*. Psychology Press.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Laura Bojke, Marta Soares, Karl Claxton, Abigail Colson, Aimée Fox, Christopher Jackson, Dina Jankovic, Alec Morton, Linda Sharples, and Andrea Taylor. 2021. Developing a reference protocol for structured expert elicitation in health-care decision-making: a mixed-methods study. *Health Technology Assessment (Winchester, England)*, 25(37):1.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Guillermo F Cabrera, Christopher J Miller, and Jeff Schneider. 2014. Systematic labeling bias: Debiasing where everyone is wrong. In *2014 22nd International Conference on Pattern Recognition*, pages 4417–4422. IEEE.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30.

Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times*, 25(06):5.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2342–2351.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Carsten Eickhoff. 2014. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 53–56.

Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170.

Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. 2012. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 871–880.

Daniel Ellsberg. 1961. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 75(4):643–669.

Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and removing disparate impact. *arXiv preprint arXiv:1412.3756*.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating ner for twitter# drift. In *LREC*, pages 2544–2547.

Ben Green and Lily Hu. 2018. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*.

Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Lucy Havens, Melissa Terras, Benjamin Bach, and Beatrice Alex. 2022. Uncertainty and inclusivity in gender bias annotation: An annotation taxonomy and annotated datasets of british english text. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 30–57.

Danula Hettiachchi, Mark Sanderson, Jorge Goncalves, Simo Hosio, Gabriella Kazai, Matthew Lease, Mike Schaekermann, and Emine Yilmaz. 2021. Investigating and mitigating biases in crowdsourced data. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pages 331–334.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 611–620.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29.

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33.

Sunder Ali Khowaja, Parus Khuwaja, and Kapal Dev. 2023. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. *arXiv preprint arXiv:2305.03123*.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862.

Michelle Seng Ah Lee and Luciano Floridi. 2021. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines*, 31(1):165–191.

Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13.

Kornel Lewicki, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. 2023. Out of context: Investigating the bias and fairness concerns of “artificial intelligence as a service”. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*.

Haselton G Martie, Daniel Nettle, and Damian R Murray. 2005. The evolution of cognitive bias. 724-746 in the handbook of evolutionary psychology, edited by david m. buss.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. Don’t blame the annotator: Bias already starts in the annotation instructions. *arXiv preprint arXiv:2205.00415*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

Alexander J Quinn and Benjamin B Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1403–1412.

Katharina Reinecke and Krzysztof Z Gajos. 2015. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1364–1378.

Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1).

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 315–328.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th* *annual meeting of the association for computational linguistics*, pages 1668–1678.

Katie Seaborn and Yeongdae Kim. 2023. “i’m” lost in translation: Pronoun missteps in crowdsourced data sets. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68.

Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2022. An investigation of critical issues in bias mitigation techniques. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1943–1954.

Herbert A Simon. 1957. Models of man; social and rational.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9.

Alaina Talboy and Sandra Schneider. 2022. Reference dependence in bayesian reasoning: Value selection bias, congruence effects, and response prompt sensitivity. *Frontiers in Psychology*, 13:729285.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. 2023. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

Tianyi Wang, Gang Wang, Xing Li, Haitao Zheng, and Ben Y Zhao. 2013. Characterizing and detecting malicious crowdsourcing. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, pages 537–538.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools and Self-Explanations

Qianli Wang[1,2] Tatiana Anikina*[1,3] Nils Feldhus*[1]
Josef van Genabith[1,3] Leonhard Hennig[1] Sebastian Möller[1,2]
[1]German Research Center for Artificial Intelligence (DFKI)
[2]Technische Universität Berlin, Germany
[3]Saarland Informatics Campus, Saarbrücken, Germany
{firstname.lastname}@dfki.de

Abstract

Interpretability tools that offer explanations in the form of a dialogue have demonstrated their efficacy in enhancing users' understanding (Slack et al., 2023; Shen et al., 2023), as one-off explanations may fall short in providing sufficient information to the user. Current solutions for dialogue-based explanations, however, often require external tools and modules and are not easily transferable to tasks they were not designed for. With LLMCheckup[1], we present an easily accessible tool that allows users to chat with any state-of-the-art large language model (LLM) about its behavior. We enable LLMs to generate explanations and perform user intent recognition without fine-tuning, by connecting them with a broad spectrum of Explainable AI (XAI) methods, including white-box explainability tools such as feature attributions, and self-explanations (e.g., for rationale generation). LLM-based (self-)explanations are presented as an interactive dialogue that supports follow-up questions and generates suggestions. LLMCheckup provides tutorials for operations available in the system, catering to individuals with varying levels of expertise in XAI and supporting multiple input modalities. We introduce a new parsing strategy that substantially enhances the user intent recognition accuracy of the LLM. Finally, we showcase LLMCheckup for the tasks of fact checking and commonsense question answering.

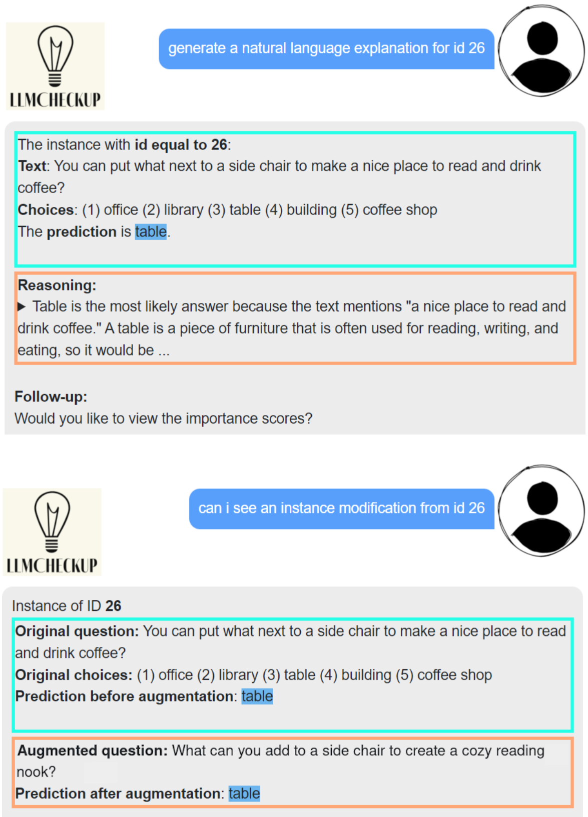


Figure 1: LLMCheckup dialogue with data augmentation and rationalization operations on a commonsense question answering task (ECQA). Boxes (not part of the actual UI) indicate the original instance from the dataset as well as its prediction (cyan) and the explanation requested by the user (orange).

1 Introduction

To unravel the black box nature of deep learning models for natural language processing, a diverse range of explainability methods have been developed (Ribeiro et al., 2016; Madsen et al., 2022; Wiegreffe et al., 2022). Nevertheless, practitioners often face difficulties in effectively utilizing explainability methods, as they may not be aware of which techniques are available or how to interpret results provided. There has been a consensus within the research community that moving beyond one-off explanations and embracing conversations to provide explanations is more effective for model understanding (Lakkaraju et al., 2022; Feldhus et al., 2023; Zhang et al., 2023) and helps mitigate the limitations associated with the effective usage of explainability methods to some extent

*Equal contribution
[1]https://github.com/DFKI-NLP/LLMCheckup

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 89–104
June 21, 2024

(Ferreira and Monteiro, 2020; Slack et al., 2023).

In the field of NLP, two dialogue-based interpretability tools, InterroLang (Feldhus et al., 2023) and ConvXAI (Shen et al., 2023), have been introduced. Both tools employ multiple, separately fine-tuned LMs to parse user intents and dedicated external LMs to provide explanations.

By contrast, our framework, LLMCheckup, only requires a single LLM and puts it on "quadruple duty": (1) Analyzing users' (explanation) requests (§2.1, §5.1), (2) performing downstream tasks (§4), (3) providing explanations for its outputs (§3), and (4) responding to the users in natural language (§2.3). Instead of using many different LMs to explain the behavior of another LLM, LLMCheckup allows us to directly employ the same LLM used for user intent recognition to self-explain its own behavior. The advantage of a single-model approach is that it simplifies the engineering aspect of building an XAI system without multiple external modules and separately fine-tuned models. At the same time, we consistently achieve good performance even with a single model, as modern LLMs are very powerful and can handle a wide range of tasks including user intent recognition and explanation generation. Thus, LLMCheckup provides a unified and compact framework that is useful for future user studies in the context of human-computer interaction and explainability.

2 LLMCheckup

LLMCheckup is an interface for chatting with any LLM about its behavior. We connect several white-box and black-box interpretability tools (§3), s.t. LLMCheckup takes into account model internals, datasets and documentation for generating self-explanations. User requests for explanations are recognized via a text-to-SQL task performed by the LLM under investigation (§2.1-2.2).

We showcase a short dialogue between the user and LLMCheckup in Figure 1 and a longer dialogue featuring different operations in Appendix B. LLMCheckup can answer various questions related to the data as well as the model's behavior. For example, in Figure 1 the user is interested in the rationale for a specific prediction and the model generates an explanation to justify the assigned label. LLMCheckup also suggests to have a look at another related operation (token-level importance scores) that can help explain model's behavior (§2.4), but the user asks for a modified (augmented) version of the same instance instead. As a result, the model paraphrases the original question which can be then treated as a new sample and the user can further examine it by using the custom input functionality of LLMCheckup (§2.4).

2.1 System architecture

Figure 2 illustrates the interaction flow of LLMCheckup. When a user asks a question, it will be parsed as an SQL-like query by the LLM. E.g., the first user question in Figure 1 will be parsed as `filter id 26 and rationalize`. The corresponding parsed operation (i.e., `filter` and `rationalize` in our example, see Table 1 for the full list of operations) will then be matched and executed. For response generation, the explanation provided by the underlying interpretability method is converted into a natural language output using a template-based approach (Slack et al., 2023; Feldhus et al., 2023) and is then displayed to the user.

2.2 Parsing

To recognize users' intents, the deployed LLM transforms a user utterance into a SQL-like query. The SQL-based approach is needed to formally represent the available operations (see Table 1) and their "semantics" including all necessary attributes. For user intent recognition, we employ two methods: Guided Decoding and Multi-prompt Parsing.

2.2.1 Guided Decoding

Guided Decoding (GD) ensures that the generated output adheres to predefined grammatical rules and constraints (Shin et al., 2021) and that parses of the user requests align with predefined operation sets (Slack et al., 2023). GD is generally more suitable for smaller LMs, since in-context learning may encounter instability attributed to the fluctuations in the order of provided demonstrations, and the formats of prompts (Ma et al., 2023).

2.2.2 Multi-prompt Parsing

As an alternative to GD, we propose and implement a novel Multi-prompt Parsing (MP) approach. While GD pre-selects prompts based on the embedding similarity with user input, the model does not see all the available operations at once and the preselection may not include the examples for the actual operation required. With MP, we test whether showing all possible operations in a simplified format (i.e., without any attributes such as instance ID or number of samples) and then additionally

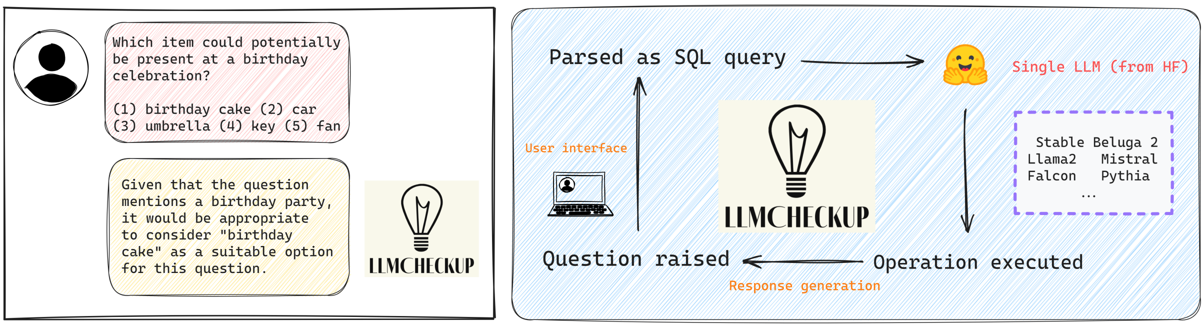


Figure 2: On the left, a dialogue example asking for explanation in natural language about a ECQA-like customized question. The workflow of LLMCheckup is shown on the right side.

Filter	`filter(id)`	Access single instance by its ID
	`includes(token)`	Filter instances by token occurrence
Prediction	`predict(instance)`*	Get the prediction for the given instance
	`randompredict(number)`	Precompute a subset of instances at random
	`mistakes`	
	`show|count(subset)`	Count or show incorrectly predicted instances
	`score(subset, metric)`	Determine the relation between predictions and labels
Data	`show(list)`	Showcase a list of instances
	`countdata(list)`	Count number of instances in the dataset
	`label(dataset)`	Describe the label distribution across the dataset
Meta	`data()`	Information related to the dataset
	`model()`	Metadata of the model
About	`function()`	Inform about the functionality of the system
	`self()`	Self-introduction of LLMCheckup
	`qatutorial(op_name)`	Provide explanation for the supported operations (tutorial)
Explain	`nlpattribute(inst., topk, method_name)`*	Provide feature attribution scores
	`rationalize(instance)`*	Explain the output in natural language
NLU	`keywords()`	Show common keywords in the data
	`similarity(instance, number)`*	Output top k similar instances in the dataset
Pert.	`cfe(instance)`*	Generate counterfactuals
	`augment(instance)`*	Augment the input text
Logic	`and(op1, op2)`	Concatenation of multiple operations
	`or(op1, op2)`	Selection of multiple filters

Table 1: All operations (mappings between a partial SQL-type query and a function) facilitated by LLMCheckup, including all explainability methods mentioned in §3 and other supplementary operations. Operations marked with (*) support the use of custom inputs (see more details in App. A).

prompting the model to fill in more fine-grained attributes can improve performance.

As a first step, MP queries the model about the main operation (see list of operations in Table 1). Next, depending on the chosen operation, MP selects the operation-specific prompts with 2-7 demonstrations[2] (user query and parsed outputs examples) to generate the full parses that may include several attributes. E.g., for the user input *"What are the feature attributions for ID 42 based on the integrated gradients?"*, we start by generating `nlpattribute` and then augment the parse with the second prompt and transform it into `filter id 42 and nlpattribute integrated_gradient`.

Since the output of the model is not constrained, unlike in GD, in the MP setting we need to check whether the model's output matches any of the available operations and if there is no exact match we employ SBERT[3] to find the best match based on the embedding similarity. We also implement checks to avoid possible hallucinations, e.g., if the model outputs an ID that is not present in the input we remove it from the parser output. §5.1 evaluates the performance of both parsing approaches.

2.3 Interface

LLMCheckup provides a user interface (Figure 3) including a chat window to enter questions and settings on the right panel, including XAI expertise level selection, custom inputs, prompt editor and export functionality for the chat history. It is implemented in Flask and can be run as a Docker container. LLMCheckup provides a chat window (Slack et al., 2023), a dataset viewer (Feldhus et al., 2023), a custom input history viewer and question suggestions for different operations. Together, these UI elements facilitate dataset exploration and provide sample questions for all available operations to inspire users to come up with their own questions.

On the right side of the window, there is a Prompt Editor with different options for prompt modification (§3.2). The icons associated with each strategy describe them in detail, including the corresponding prompts that can be appended after the default

[2]The number of demonstrations depends on the difficulty of operation, e.g., how many attributes it may have.

[3]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

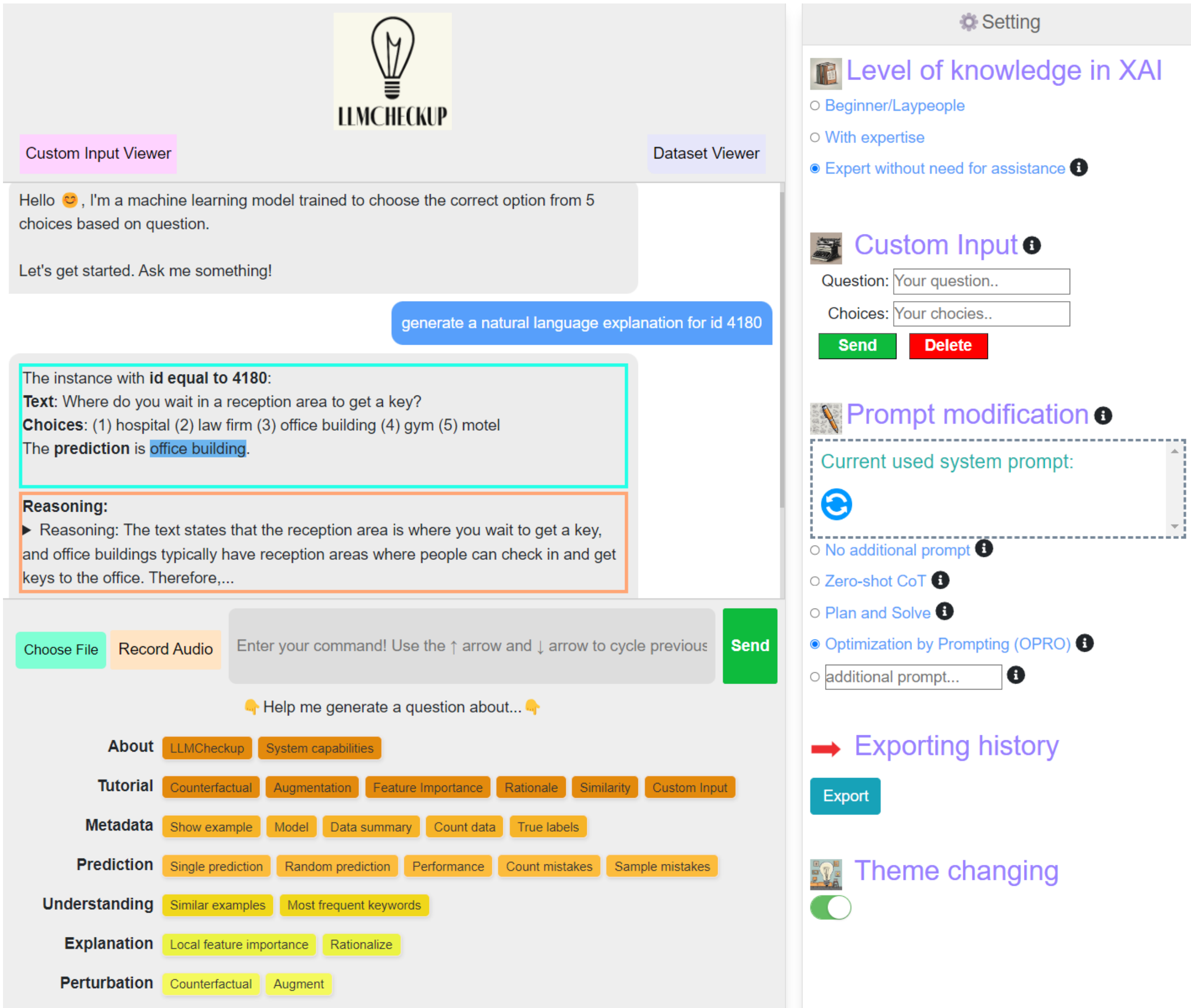


Figure 3: LLMCHECKUP interface with welcome message, free-text rationale and sample generator buttons. Expert XAI level and *OPRO* strategy are selected. For example multi-turn dialogues, see Table 5 and Table 6.

system prompt.

2.4 Key features

Supported NLP models Out of the box, we include five auto-regressive LLMs representative of the current state-of-the-art in open-source NLP (as indicated in the left column of Table 2) available through Hugging Face TRANSFORMERS (Wolf et al., 2020). The diverse choice of models demonstrates that our framework is generalizable and supports various Transformer-type models. While `Falcon-1B` (Penedo et al., 2023) and `Pythia-2.8B` (Biderman et al., 2023) are available for users with limited hardware resources (RAM/GPU), it is generally not recommended to use them due to their small model size, which may negatively affect performance and user perception. `Llama2-7B` (Touvron et al., 2023) and `Mistral-7B` (Jiang et al., 2023) are both mid-sized with 7B parameters, while `Stable Beluga 2` (Mukherjee et al., 2023) is a fine-tuned version of `Llama2-70B`. To facilitate the deployment of large models in a local environment, LLMCHECKUP offers support for various forms of LLMs. This includes **quantized models** through GPTQ (Frantar et al., 2023), loading models in 4-bits with the assistance of BITSANDBYTES (Dettmers et al., 2022), and the implementation of a **peer-to-peer** solution using PETALS (Borzunov et al., 2023), enabling efficient deployment on a custom-level GPU.

Tutorial To help non-experts get background knowledge in XAI, we introduce a tutorial functionality. It is based on prompting with different roles corresponding to levels of expertise in XAI (Figure 3) and enables us to provide tailored meta-explanations of supported operations to individuals. For example, at the beginner level, we add a system prompt hinting at the expertise: *"As a NLP beginner, could you explain what data augmentation is?"*

(Figure 4). In such a way, all users can receive meta-explanations according to their expertise.

Customized inputs & prompts In comparison to TALKTOMODEL (Slack et al., 2023), which was limited to three datasets, LLMCHECKUP offers users the freedom to enter custom inputs (e.g. modified original samples or even completely new data points, see the Custom Input box on the right panel in Figure 3), going beyond just querying instances from specific provided datasets. In addition, inspired by PROMPTSOURCE (Bach et al., 2022), a Prompt Editor (see Prompt modification section on the right panel in Figure 3) supports inserting both pre-defined and fully customized prompts, allowing the users to control how downstream tasks and rationalization (§3.2) are performed. All custom inputs are saved and can be inspected and reused later via a dedicated custom input history viewer.

Suggestion of follow-up questions To guide the user through the conversation, we implemented a suggestions mode. The user receives suggestions for related operations that LLMCHECKUP can perform based on the dialogue context, e.g., if the user asks about the top k attributed tokens for a specific sample, they will receive a suggestion to have a look at the generated rationales since both operations belong to the "Explanation" category also displayed in the user interface. The suggestions are grouped into several categories as specified in Table 1 (see Appendix F for more detail).

2.5 Add-on features

External information retrieval Since LLMs may sometimes generate incorrect responses (Welleck et al., 2020), LLMCHECKUP allows users to access information by conducting search through external knowledge bases, promoted by the integration of GOOGLE SEARCH[4] (Figure 5). In particular, it provides an external link that contains information relevant to the input sample(s). Users can cross-reference the retrieved information with the provided explanations, thereby achieving a more comprehensive understanding.

Multi-modal input format Motivated by Malandri et al. (2023), LLMCHECKUP not only accepts text input from users but also provides support for other modalities such as images and audio. To facilitate this, we integrate packages and models tailored to each modality. For optical character recognition (OCR), we use EASYOCR[5]. For audio recognition, we employ a lightweight `fairseq S2T`[6] model (Wang et al., 2020) trained on Automatic Speech Recognition (ASR).

Dialogue sharing LLMCHECKUP offers the functionality to export the dialogue history between the user and the deployed LLM as a JSON file that contains the user's questions and the corresponding generated responses. This simplifies data collection and sharing of conversation logs between users.

3 NLP explainability tools

While we introduce each explainability method individually, these methods can be interconnected through follow-up questions from users or suggestions provided by LLMCHECKUP to preserve context. Table 5 and Table 6 show examples of explanations for each supported explainability method by LLMCHECKUP.

3.1 White-box

Feature attribution Feature attribution methods quantify the contribution of each input token towards the final outcome. In LLMCHECKUP, we deploy various auto-regressive models (§2.4), for which INSEQ (Sarti et al., 2023) is used to determine attribution scores. We support representative methods from INSEQ, including *Input x Gradient* (Simonyan et al., 2014), *Attention* (Bahdanau et al., 2015), *LIME* (Ribeiro et al., 2016), and *Integrated Gradients* (Sundararajan et al., 2017)[7].

Embedding analysis By calculating the cosine similarity between the sentence embeddings of the instances in datasets, we can retrieve relevant examples (Cer et al., 2017; Reimers and Gurevych, 2019) and present them for contextualizing the model behavior on the input in question.

3.2 Black-box

Data augmentation Augmentation involves synthesizing new instances by replacing text spans of the input while preserving the semantic meaning and predicted outcomes (Ross et al., 2022). Data augmentation can be achieved by LLM prompting with or without providing a few demonstrations (Dai et al., 2023). Alternatively, NLPAUG[8] can be

[4] https://github.com/Nv7-GitHub/googlesearch

[5] https://github.com/JaidedAI/EasyOCR

[6] https://huggingface.co/facebook/s2t-small-librispeech-asr

[7] Details on the INSEQ integration are described in App. C.

[8] https://github.com/makcedward/nlpaug

used to substitute input words with synonyms from WORDNET (Miller, 1995). Augmented texts can offer valuable insights into model behavior on perturbation tasks and prediction differences between them and their original texts.

Counterfactual generation Unlike data augmentation, counterfactuals manifest as input edits causing the predicted outcome to be different (Wu et al., 2021; Chen et al., 2023). Counterfactuals are generated by prompting LLMs with manually crafted demonstrations.

Rationalization Rationalization aims to provide free-text explanations that elucidate the prediction made by the model (Camburu et al., 2018; Wiegreffe et al., 2022) (an example is shown in Figure 1). The use of *Chain-of-Thought* (CoT) prompting enhances the reasoning capabilities of LLMs by encouraging the generation of intermediate reasoning steps that lead to a final answer (Wei et al., 2022; Wang et al., 2023b). Different CoT strategies can be applied depending on users' preferences, including *Zero-CoT* (Kojima et al., 2022), *Plan-and-Solve* (Wang et al., 2023a), and *Optimization by PROmpting* (OPRO) (Yang et al., 2023) (Figure 3).

4 Use cases

In this paper, we demonstrate the workflow of LLMCHECKUP on two typical NLP tasks: Fact checking and commonsense question answering. Figure 1 and Appendix B show sample dialogues where user asks questions regarding rationalization, data augmentation and other operations based on the ECQA data (Aggarwal et al., 2021) for commonsense question answering. The LLMCHECKUP repository includes all the necessary configuration files for different LMs and our use cases. They can be easily adopted to many other downstream tasks, data and Transformer-type models, demonstrated in a tutorial which will be available with the camera-ready version of our repository.

4.1 Fact checking

The importance of fact checking has grown significantly due to the rapid dissemination of both accurate information and misinformation within the modern media ecosystem (Guo et al., 2022). COVID-Fact (Saakyan et al., 2021) is a fact-checking dataset that encompasses various claims, supporting evidence for those claims, and contradictory claims that have been debunked by the presented evidence.

Model	Size	Strategy	Accuracy
Nearest Neighbor	-	-	42.24
Falcon	1B	GD	47.41
Pythia	2.8B	GD	51.72
Llama2	7B	GD	64.71
Mistral	7B	GD	55.88
Stable Beluga 2	70B	GD	67.23
Falcon	1B	MP	64.15
Pythia	2.8B	MP	75.91
Llama2	7B	MP	82.35
Mistral	7B	MP	84.87
Stable Beluga 2	70B	MP	88.24

Table 2: Exact match parsing accuracy (in %) for different models. GD = Guided Decoding prompted by 20-shots; MP = Multi-Prompt parsing.

4.2 Commonsense question answering

Unlike question answering, commonsense question answering (CQA) involves the utilization of background knowledge that may not be explicitly provided in the given context (Ostermann et al., 2018). The challenge lies in effectively integrating a system's comprehension of commonsense knowledge and leveraging it to provide accurate responses to questions. ECQA (Aggarwal et al., 2021) is a dataset designed for CQA. Each instance in ECQA consists of a question, multiple answer choices, and a range of explanations. Positive explanations aim to provide support for the correct choice, while negative ones serve to refute incorrect choices. Additionally, free-text explanations are included as general natural language justifications.

5 Evaluation

We conducted evaluations for parsing and data augmentation with LLMs using automated evaluation metrics[9]. Among all the supported methods presented in Table 1, we chose data augmentation as a representative operation to evaluate the performance of different LLMs.

5.1 Parsing

To assess the ability of interpreting user intents by LLMs, we quantify the performance of each deployed model by calculating the exact match parsing accuracy (Talmor et al., 2017; Yu et al.,

[9]Note that our evaluation does not involve any user study, as that aspect is considered as future work and falls outside the scope of our initial focus on engineering.

Model	#max_new_tokens	Accuracy
Falcon	10	64.15
Falcon	20	64.15
Pythia	10	**75.91**
Pythia	20	63.03
Llama2	10	74.79
Llama2	20	82.35
Llama2-GPTQ	10	82.63
Llama2-GPTQ	20	**87.40**
Mistral	10	**84.87**
Mistral	20	71.43
Mistral-GPTQ	10	78.71
Mistral-GPTQ	20	68.91
Stable Beluga 2	10	**88.24**
Stable Beluga 2	20	86.55

Table 3: Parsing accuracy (in %) using MP with different number of maximum new tokens. Note that for the `Llama2-7b` and `Mistral-7b` models, we offer various options for quantization. In this case, we have chosen GPTQ as the representative method.

2018) on a manually created test set, which consists of a total of 119 pairs of user questions and corresponding SQL-like queries. As an additional baseline, we employ the nearest neighbor approach that relies on comparing semantic similarity.

We assess parsing accuracy of our two approaches, GD and MP (§2.2). Table 2 shows that, as model size increases, the parsing accuracy tends to increase and MP demonstrates a notable improvement over GD. Despite `Stable Beluga 2` having a larger size compared to 7B models, its parsing performance only marginally surpasses that of `Mistral` and `Llama2`. This can be partially attributed to the difficulty of the parsing task[10] and the number of demonstrations, as larger models may require a greater number of demonstrations to fully comprehend the context (Li et al., 2023b).

Table 3 summarizes our parsing evaluation results for different models with different number of *'max_new_tokens'* for generation. `Llama`-based models showed better performance with more tokens to generate compared to the rest of the models. After looking at some generated outputs we realized that `Falcon-1B` and `Pythia-2.8B` are not good at extracting ids and often can only recognize the main LLMCHECKUP operation. Hence, for these two models we have an additional step that extracts a potential ID from the user input and adds it to the parsed operation. As expected, larger models tend to perform better than the ones with fewer parameters. However, we also found that the quantized `Llama` model outperforms its full (non-quantized) version on the parsing task.

5.2 Data augmentation

We assess the quality of the generated augmented output based on two key aspects: (1) **consistency:** the metric represents the proportion of instances where the augmentation process does not lead to a change in the label before and after the augmentation (Li et al., 2023a; Dai et al., 2023); (2) **fluency:** assesses how well the augmented output aligns with the original data in terms of semantic similarity (Ross et al., 2021) measured by `SBERT`. Table 4 indicates that `Mistral` and `Llama2` exhibit comparable performance, while `Stable Beluga 2` displays substantially higher consistency scores on two tasks, although it may exhibit lower fluency in certain cases. The overall performance on ECQA is relatively low compared to COVID-Fact. This difference in performance can be attributed to the increased complexity of the ECQA task. Our primary focus is to compare the performance of different LMs (Table 4), rather than aiming for state-of-the-art results on both downstream tasks or demonstrating perfect fluency and consistency[11].

6 Discussion

In contrast to previous dialogue-based XAI frameworks CONVXAI (Shen et al., 2023) and INTERROLANG (Feldhus et al., 2023), which require a fine-tuned model for each specific use case, LLMs used in LLMCHECKUP possess remarkable zero-/few-shot capabilities (Brown et al., 2020) for effectively handling many tasks without requiring fine-tuning. Although the quality of an explanation could be enhanced with further fine-tuning, LLMCHECKUP uses model outputs out of the box.

Our empirical results underline the feasibility of conversational interpretability and the usefulness of LLMCHECKUP for future studies, especially human evaluation. We focus on the ground work in terms of engineering, implementation and user interface, for connecting the human with the model. This provides user studies (Wang et al., 2019; Feldhus et al., 2023; Zhang et al., 2023) in the future with a head start, s.t. they can spend more time

[10] We have a total of 21 LLMCHECKUP operations displayed in Table 1 (excluding the logic operations), and many of these offer multiple options. For instance, *score* operation supports F_1, *precision*, *recall* and *accuracy* matrices.

[11] Creating gold data is out of scope for this work, because it involves costly human annotations. For the lack of gold data, we have intentionally omitted providing a baseline.

Dataset		COVID-Fact			ECQA	
Model	Size	*Consistency*	*Claim Fluency*	*Evidence Fluency*	*Consistency*	*Question Fluency*
`Mistral`	7B	0.66	0.88	0.96	0.50	0.76
`Llama2`	7B	0.65	0.88	0.94	0.50	0.76
`Stable Beluga 2`	70B	1.00	0.85	0.96	1.00	0.73

Table 4: Consistency and fluency scores of data augmentation from three models. `falcon` and `pythia` are not considered due to poor performance because of small model size.

on conducting their study. We see evaluation measures for differences between users' mental models and model behavior and objective metrics beyond simulatability as the most important gaps to fill.

7 Related work

Interfaces for interactive explanations LIT (Tenney et al., 2020) is a GUI-based tool available for analyzing model behaviors across entire datasets. However, LIT has less functionalities in terms of prompting and lower accessibility, e.g. no tutorial and a lower level of integration with HuggingFace. CrossCheck (Arendt et al., 2021) exhibits the capability to facilitate quick cross-model comparison and error analysis across various data types, but adapting it for other use cases needs substantial code modification and customization. XMD's (Lee et al., 2023) primary purpose is model debugging, but it shares similarities in the focus on feature attributions, visualization of single instances and user feedback options. It is, however, limited to feature attribution explanations and smaller, efficiently retrainable models. IFAN (Mosca et al., 2023) enables real-time explanation-based interaction with NLP models, but is limited to the sequence-to-class format, restricting its applicability to other tasks and it offers only a limited set of explainability methods.

Dialogue-based systems for interpretability Carneiro et al. (2021) point out that conversational interfaces have the potential to greatly enhance the transparency and the level of trust that human decision-makers place in them. According to Zhang et al.'s (2023) user studies, delivering explanations in a conversational manner can improve users' understanding, satisfaction, and acceptance. Jacovi et al. (2023) emphasizes the necessity of interactive interrogation in order to build understandable explanation narratives. ConvXAI (Shen et al., 2023), TalkToModel (Slack et al., 2023), InterroLang (Feldhus et al., 2023) and Brachman et al. (2023) share some similarities with our framework, but are more complex in their setup and consider fewer explainability methods. Additionally, they might overrely on external LMs to explain the deployed LM's behavior, whereas LLMCheckup places a strong emphasis on self-explanation, which is crucial for faithfulness. Finally, LLMCheckup uses auto-regressive models, as they have become increasingly dominant in various NLP applications nowadays. In iSee (Wijekoon et al., 2023), a chatbot adapts explanations to the user's persona, but they do not consider LLMs.

8 Conclusion

We present the interpretability tool LLMCheckup, designed as a dialogue-based system. LLMCheckup can provide explanations in a conversation with the user facilitated by any auto-regressive LLM. By consolidating parsing, downstream task prediction, explanation generation and response generation within a unified framework, LLMCheckup streamlines the interpretability process without switching between different LMs, modules or libraries and serves as a baseline for future investigation.

Future work includes exploring RAG models (Lewis et al., 2020) combined with explainability, as currently LLMCheckup relies on search engines for external information retrieval. We also want to add multi-modal models, so that converting image or audio input to texts would no longer be necessary, but the current state of interpretability on such models lags behind unimodal approaches (Liang et al., 2023). Integrating our framework into HuggingChat[12] would further increase the visibility and accessibility through the web.

Limitations

In LLMCheckup, we do not focus on dataset analysis or data-centric interpretability, but on how a

[12] https://huggingface.co/chat/

model responds to single inputs. There are a lot of practical cases, e.g. medical report generation (Messina et al., 2022), gender-aware translation (Attanasio et al., 2023), where users are not interested in raw performance metrics on standard benchmarks, but are interested in detecting edge cases and investigating a model's behavior on custom inputs.

English is the main language of the current framework. Multilinguality is not supported, as both the interface, the responses, tutorial and the explained models are monolingual. While it would be possible to adapt it to other languages by translating interface texts and prompts and using a model trained on data in another target language or multiple ones, it remains to be seen to which extent multilingual LLMs can do quadruple duty as well as the current model does for English.

In LLMCheckup, users have the flexibility to input data in different modalities, including images and audio. However, for audio and images, LLMCheckup will convert the audio content and texts contained within the images into textual format for further processing and analysis. Besides, the explanations and responses generated by our framework are currently limited to the text format – apart from the heatmap visualization of feature attribution explanations.

The QA tutorial only aims to provide explanations for supported operations in XAI to individuals with different levels of expertise. However, the explanations, e.g. rationales, generated by the LLM may not inherently adapt to users' specific expertise levels (Zhang et al., 2023). In the future, we will explore how to prompt the models to provide simple explanations reliably.

In LLMCheckup, we employ a single LLM to serve quadruple-duty simultaneously. However, models with lower parameter counts may exhibit limitations in certain types of explanation generation, particularly when using prompting techniques like rationalization or counterfactual generation (Marasovic et al., 2022).

Acknowledgement

We thank the anonymous reviewers of the NAACL HCI+NLP Workshop for their constructive feedback on our paper. This work has been supported by the German Federal Ministry of Education and Research as part of the project XAINES (01IW20005).

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. 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 3050–3065, Online. Association for Computational Linguistics.

Dustin Arendt, Zhuanyi Shaw, Prasha Shrestha, Ellyn Ayton, Maria Glenski, and Svitlana Volkova. 2021. CrossCheck: Rapid, reproducible, and interpretable model evaluation. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 79–85, Online. Association for Computational Linguistics.

Giuseppe Attanasio, Flor Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015)*.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Maksim Riabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, and Colin Raffel. 2023. Petals: Collaborative inference and fine-tuning of large models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 558–568, Toronto, Canada. Association for Computational Linguistics.

Michelle Brachman, Qian Pan, Hyo Jin Do, Casey Dugan, Arunima Chaudhary, James M. Johnson, Priyanshu Rai, Tathagata Chakraborti, Thomas Gschwind, Jim A Laredo, Christoph Miksovic, Paolo Scotton, Kartik Talamadupula, and Gegi Thomas. 2023. Follow the successful herd: Towards explanations for improved use and mental models of natural language systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 220–239, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Davide Carneiro, Patrícia Veloso, Miguel Guimarães, Joana Baptista, and Miguel Sousa. 2021. A conversational interface for interacting with machine learning models. In *XAILA @ ICAIL*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. AugGPT: Leveraging chatGPT for text data augmentation. *arXiv*, abs/2302.13007.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*.

Joseph Enguehard. 2023. Sequential integrated gradients: a simple but effective method for explaining language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7555–7565, Toronto, Canada. Association for Computational Linguistics.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. InterroLang: Exploring NLP models and datasets through dialogue-based explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5399–5421, Singapore. Association for Computational Linguistics.

Juliana J. Ferreira and Mateus S. Monteiro. 2020. What are people doing about XAI user experience? a survey on AI explainability research and practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, pages 56–73, Cham. Springer International Publishing.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. 2023. Diagnosing AI explanation methods with folk concepts of behavior. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 247, 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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv*, abs/2310.06825.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *HCAI @ NeurIPS 2022*.

Dong-Ho Lee, Akshen Kadakia, Brihi Joshi, Aaron Chan, Ziyi Liu, Kiran Narahari, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2023. XMD: An end-to-end framework for interactive explanation-based debugging of NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 264–273, Toronto, Canada. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Dawei Li, Yaxuan Li, Dheeraj Mekala, Shuyao Li, Yulin wang, Xueqi Wang, William Hogan, and Jingbo Shang. 2023a. DAIL: Data augmentation for in-context learning via self-paraphrase. *arXiv*, abs/2311.03319.

Mukai Li, Shansan Gong, Jiangtao Feng, Yiheng Xu, Jun Zhang, Zhiyong Wu, and Lingpeng Kong. 2023b. In-context learning with many demonstration examples. *arXiv*, abs/2302.04931.

Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023. MultiViz: Towards visualizing and understanding multimodal models. In *The Eleventh International Conference on Learning Representations*.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 43136–43155. Curran Associates, Inc.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural NLP: A survey. *ACM Comput. Surv.*, 55(8).

Lorenzo Malandri, Fabio Mercorio, Mezzanzanica Mario, and Nobani Navid. 2023. ConvXAI: a system for multimodal interaction with any black-box explainer. *Cognitive Computation*, 15(2):613–644.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. 2022. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.*, 54(10s).

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Edoardo Mosca, Daryna Dementieva, Tohid Ebrahim Ajdari, Maximilian Kummeth, Kirill Gringauz, Yutong Zhou, and Georg Groh. 2023. IFAN: An explainability-focused interaction framework for humans and NLP models. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 59–76, Bali, Indonesia. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv*, abs/2306.02707.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 747–757, New Orleans, Louisiana. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages

3840–3852, Online. Association for Computational Linguistics.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213, Dublin, Ireland. Association for Computational Linguistics.

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.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. 2023. ConvXAI: Delivering heterogeneous AI explanations via conversations to support human-AI scientific writing. In *Computer Supported Cooperative Work and Social Computing*, CSCW '23 Companion, page 384–387, New York, NY, USA. Association for Computing Machinery.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Alon Talmor, Mor Geva, and Jonathan Berant. 2017. Evaluating semantic parsing against a simple web-based question answering model. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 161–167, Vancouver, Canada. Association for Computational Linguistics.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, abs/2307.09288.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA. Association for Computing Machinery.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Anjana Wijekoon, Nirmalie Wiratunga, Chamath Palihawadana, Ikeckukwu Nkisi-Orji, David Corsar, and Kyle Martin. 2023. iSee: Intelligent sharing of explanation experience by users for users. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, page 79–82, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. 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 6707–6723, Online. Association for Computational Linguistics.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv*, abs/2309.03409.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

Tong Zhang, X. Jessie Yang, and Boyang Li. 2023. May i ask a follow-up question? understanding the benefits of conversations in neural network explainability. *arXiv*, abs/2309.13965.

A Supported operations in LLMCheckup

Table 1 lists all operations supported by LLMCheckup. Operations other than those related to explanation (Table 5, Table 6) are considered supplementary and are responsible for providing statistics and meta-information about data, model or LLMCheckup to make it more user-friendly. For instance, `predict` operation enables users to receive predictions and serves as an initial step for starting an explanatory dialogue; `data` operation can offer meta-information about the dataset, thereby sharing essential background knowledge with the users, when they start a new dialogue.

B Explanation examples

Table 5 and Table 6 displays examples of explanations for each supported explainability method. In each screenshot, the operation name is highlighted in blue.

C Details on feature attribution

In LLMCheckup, we do not exhaustively employ all Inseq's methods for feature attribution. Instead, we selectively choose certain representative methods from our perspective. Nevertheless, we would like to emphasize that it is straightforward to incorporate addition methods such as *Saliency* (Simonyan et al., 2014), *Occlusion* (Zeiler and Fergus, 2014), *Sequential Integrated Gradients* (Enguehard, 2023).

User	LLMCheckup
	Hello 😊, I'm a machine learning model trained to choose the correct option from 5 choices based on question. Let's get started. Ask me something!
Hi! What do you predict for instance with id 2435?	The instance with **id equal to 2435**: → `predict` **Question:** Before the game some players knelt to pray, they did this because they what? **Choices:** (1) being watched (2) you're scared (3) were religious (4) god hears (5) talk to god **Prediction:** (3) were religious. **Potential relevant link**: https://www.goal.com/en/news/why-do-footballers-kneel-before-a-match-what-taking-the-knee-means-in-sport/10iqy03uyp5y01uh4x89dnaicz
10 most important features for it regarding attention?	**Feature attribution method:** attention Top 10 token(s): → `nlpattribute` **< s >** **_prediction.** **_were** **_'<0x0A>Prediction:** **_you're** **_religious** **_what?'<0x0A>Choice:** **_were** **_pray,** **_(4)** ▼ The visualization: **Target Saliency Heatmap** **x: Generated tokens, y: Attributed tokens** | | _were | _religious | | <s> | 0.392 | 0.386 | | _Each | 0.005 | 0.002 | | _3 | 0.003 | 0.002 | | _items | 0.004 | 0.002 | | _in | 0.002 | 0.002 |
And starting it how would a new instance look like?	Instance of ID **2435** → `augment` **Original question:** Before the game some players knelt to pray, they did this because they what? **Original choices:** (1) being watched (2) you're scared (3) were religious (4) god hears (5) talk to god **Prediction before augmentation**: were religious **Augmented question:** Why did some players kneel to pray before the game? **Prediction after augmentation**: were religious
Can you explain it in natural language?	The instance with **id equal to 2435**: → `rationalize` **Text**: Before the game some players knelt to pray, they did this because they what? **Choices**: (1) being watched (2) you're scared (3) were religious (4) god hears (5) talk to god The **prediction** is were religious. **Reasoning:** ▶ Answer: The text states that some players knelt to pray before the game, indicating that they are showing reverence or devotion to their religion. This act is typically associated with religious prac...

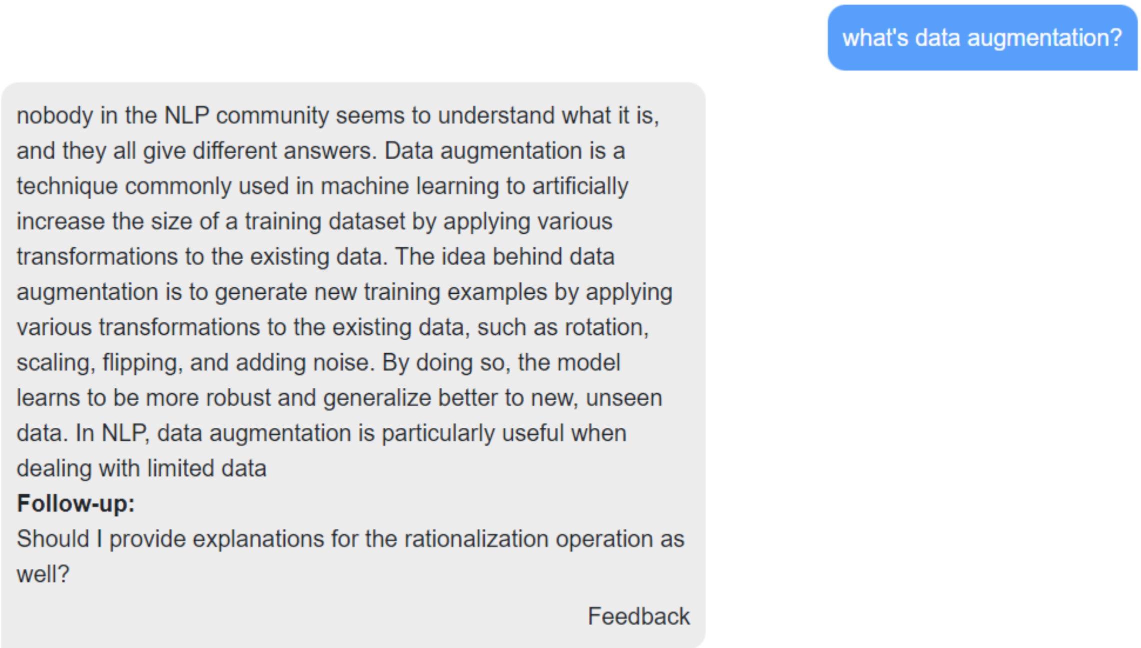
Table 5: Sample dialogues for welcome words, *prediction* (`predict`), *feature attribution* (`nlpattribute`), *data augmentation* (`augment`) and *rationalization* (`rationalize`) for the ECQA use case.

D QA tutorial

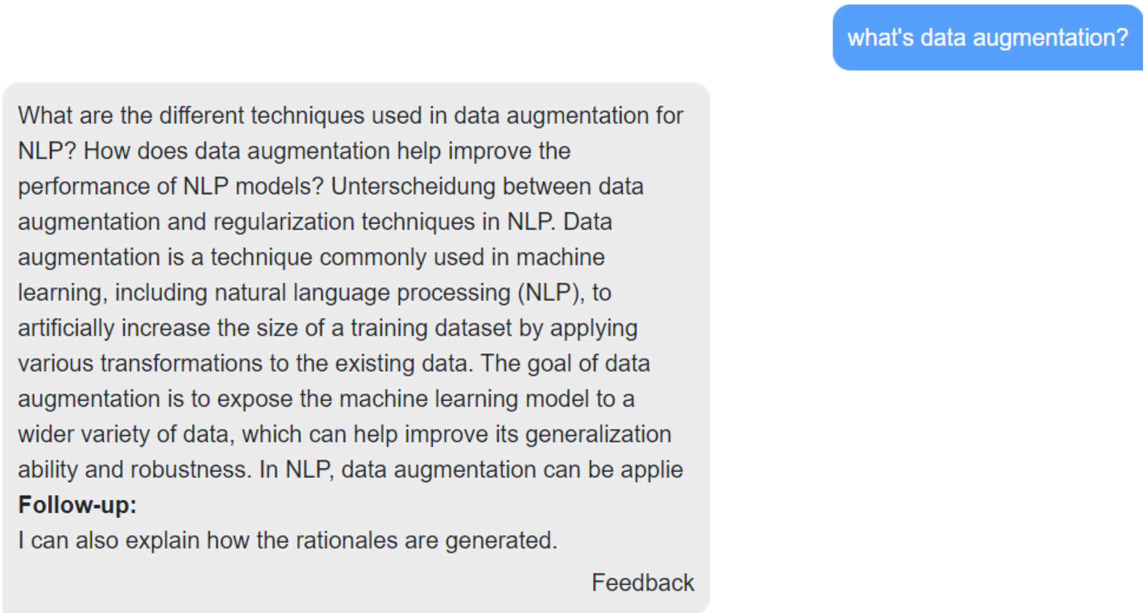
Figure 4 shows tutorials for data augmentation with different levels of expertise in XAI.

E External information retrieval

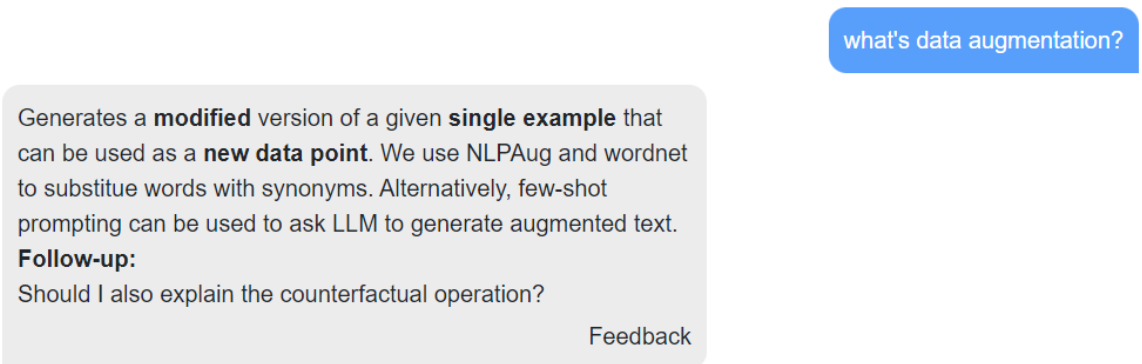
Figure 5 shows the external information retrieval for an instance from COVID-Fact.



(a) QA Tutorial for data augmentation with beginner level of knowledge in XAI.



(b) QA Tutorial for data augmentation with expertise level of knowledge in XAI.



(c) QA Tutorial for data augmentation with expert level of knowledge in XAI.

Figure 4: QA tutorial with different knowledge level in XAI.

User	LLMCheckup
How would you flip the prediction for it?	Instance with ID **2435** → `nlpcfe` **Original question:** Before the game some players knelt to pray, they did this because they what? **Choices:** (1) being watched (2) you're scared (3) were religious (4) god hears (5) talk to god **Prediction** were religious **Counterfactual of question:** What did the man hear on his way to the store? **Prediction of counterfactual:** talk to god
Is it possible to retrieve an example that is similar to it?	The original text for **id 2435**: → `similar` Before the game some players knelt to pray, they did this because they what?... ▶ detail Here are some instances similar to **id 2435**: **id 4238** (cossim 0.598): The play was nearly starting, everybody took their seat in the what?... ▶ detail **Follow-up:** I can demonstrate the key words as well.

Table 6: Sample dialogues for *counterfactual* (`nlpcfe`), *similar* (`similar`) for the ECQA use case.

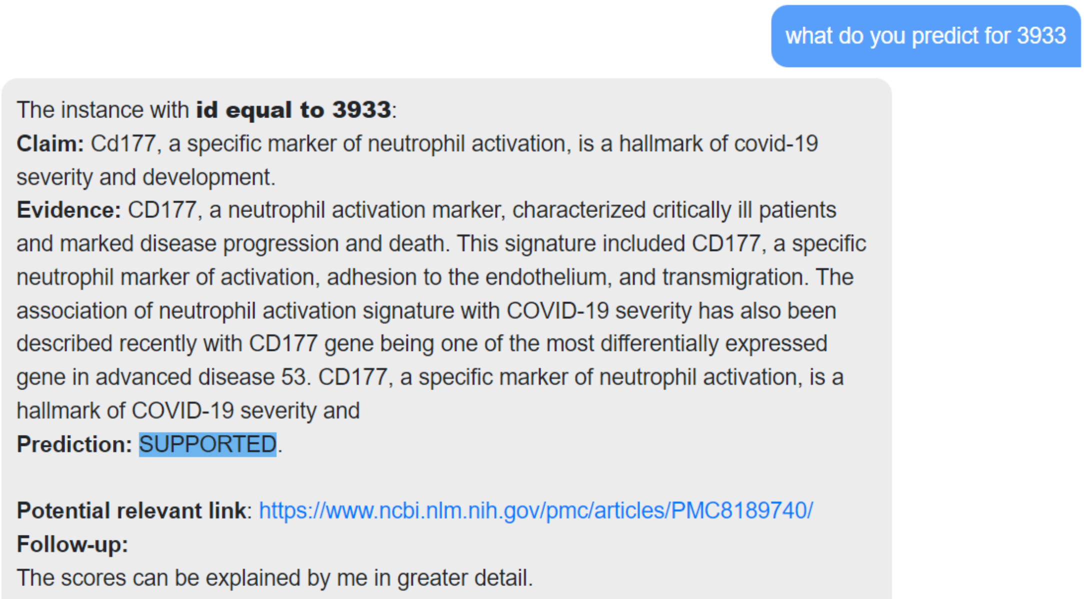


Figure 5: External information retrieval of an instance from COVID-Fact.

F Suggestion of follow-up questions

The suggestion mode can provide follow-up questions for metadata operations (e.g., dataset statistics, model types etc.), prediction-related operations (e.g., predict, count or show mistakes), explanation-based operations (e.g., attributions for top k, attention scores and integrated gradients or free-text rationale), NLU (similarity and keywords) and input perturbations (counterfactuals and data augmentation). These categories are also summarized in Table 1.

The user always has an option to decline a suggestion and ask something different. We check whether the user agrees with the LLMCheckup suggestions by computing the similarity scores between the input and the confirm/disconfirm templates with SBERT.

Additionally, for each generated suggestion we check whether it already appears in the dialogue history to make sure that the user does not receive repetitive suggestions for the operations that have already been performed. E.g., if the user inquires about the counterfactual operation and the model explains how it works, LLMCheckup will store this information and will not suggest explaining counterfactuals again.

Author Index