Language Understanding in the Human Machine Era 2025

Proceedings of 2nd LUHME Workshop

Edited by

Henrique Lopes Cardoso

Rui Sousa-Silva

Maarit Koponen

Antonio Pareja-Lora

LUHME @ ECAI 2025

Bologna, Italy

October 26, 2025

Published by
UP - Universidade do Porto (https://www.up.pt)
LIACC - Laboratório de Inteligência Artificial e Ciência de Computadores da Universidade do Porto
CLUP - Centro de Linguística da Universidade do Porto
UEF - The University of Eastern Finland
UAH - Universidad de Alcalá

ISBN 978-989-9193-73-4
DOI https://doi.org/10.21747/978-989-9193-73-4/lan2

This publication was supported by FCT – Fundação para a Ciência e a Tecnologia, I.P., under projects UID/00027 (LIACC) and UID/00022 (CLUP).

October, 2025

Introduction: Language Understanding in the Human-Machine Era

Large language models (LLMs) have revolutionized the way interactional artificial intelligence (AI) systems are developed by making them accessible to the general public. Significant advances have been observed in fields such as conversational AI and machine translation, and their widespread use in the so-called human-machine era Sayers et al., 2021 is undeniable; those models have produced remarkable achievements in several benchmarks Gao et al., 2021; Hendrycks et al., 2021; Srivastava et al., 2023; A. Wang et al., 2019; Zhou et al., 2020, and the scientific community has discussed emergent properties Wei et al., 2022 that result from scaling laws Kaplan et al., 2020. Nevertheless, state-of-the-art systems are still prone to brittleness in language understanding, which raises doubts about the extent to which such systems can truly *understand* human language(s) Mitchell and Krakauer, 2023.

The concept of language understanding has always been controversial Lyons, 1990; Michael et al., 2023. As contemporary linguistic theories have shown, meaning-making relies not only on form and (immediate) semantic meaning, but also on context. Thus, understanding natural language entails more than observing the form and the meaning withdrawn from that form; instead, harnessing meaning Bender and Koller, 2020 requires access to grounding of some sort Allein et al., 2025. Therefore, understanding language is unsurprisingly a very complex task, even for humans Lyons, 1990. As discourse, pragmatics, and (social) context are particularly relevant for understanding language, how to equip language models with such linguistics-grounded capabilities is yet to be fully understood Mao et al., 2025. Nevertheless, language models are seemingly capable of generalising concepts, which could be seen as some kind of meaning understanding Piantadosi and Hill, 2022, even if modest.

Consequently, understanding language is a doubly challenging task. Besides understanding the intrinsic capabilities of LLMs, it is increasingly important to investigate the requirements and impact of using such systems in real-world applications. As has been empirically demonstrated, LLMs can be used effectively in various applications, even without sophisticated language understanding skills, but the absence of solid theories that support these findings raises concerns about which kinds of applications pose greater risks and ethical concerns, such as bias X. Wang et al., 2025, particularly those dealing directly with human interaction. Notable examples of such risks and concerns include the impact of language technology on teaching and language work. For instance, research is underway on the use of language models in educational settings Garcia-Varela et al., 2025; S. Wang et al., 2024.

Machine translation (MT) is increasingly ubiquitous, as it is used by both language professionals and general speakers at (apparently) no cost. Causal language modeling is now becoming the new standard for MT Xu et al., 2024. Yet, as MT systems can take in a limited amount of context, they tend to make mistakes similar to what may happen to inattentive human translators if they rely on the source text alone. To understand and convey the intended meaning, human translators also need to rely on their own text-external knowledge. More broadly, recent research has called increasing attention to the role of situated and embodied cognition in translation Risku and Rogl, 2020.

As the way AI systems are intertwined with human expertise in language understanding is quickly changing, some have raised the question of the role played by language professionals in tasks such as translation. These professionals systematically add value to building next-generation language models that use linguistic and commonsense knowledge to provide more robust systems. Furthermore, it is important to understand how increasing human-machine interaction impacts the work of language professionals.

The "Language Understanding in the Human-Machine Era" (LUHME) workshop aims to reignite, retrieve, resume, and refocus the enduring debate about the role of understanding in natural language use and related applications. Specifically, it seeks to elucidate the nature of language understanding and ascertain whether it is indispensable for computational natural language tasks such as automated translation and natural language generation. Furthermore, it aims to provide insight into the role played by language professionals (e.g., linguists, professional translators, interpreters, language educators) in computational natural language understanding. It will, therefore, convene researchers interested in the intersection of language understanding and the effective use of language technologies in human-machine interaction.

The workshop's call for papers included the following topics: Language understanding in LLMs; Language grounding; Psycholinguistic approaches to language understanding; Discourse, pragmatics and language understanding; Intent detection; Evaluation of language understanding; Human vs. machine language understanding; Machine translation/interpreting and language understanding; Multimodality and language understanding; Socio-cultural aspects in understanding language; Effects and risks of language misunderstanding; Manifestations of language (mis)understanding; Natural language understanding and toxic content; Ethical issues in language misunderstanding; Distributional semantics and language understanding; Linguistic theory and language understanding by machines; Linguistic, world, and commonsense knowledge in language understanding; Role of language professionals in the LLMs era; Understanding language and explainable AI.

Each of the 15 papers submitted to the workshop was carefully revised by three PC members, and 10 papers were accepted. The program also includes a keynote by Chloé Clavel (INRIA Paris). The workshop's program is organized into four thematic sessions:

- Transparency and Social Dynamics in LLMs
- Cultural and Ethical Perspectives
- Extending the Capabilities of Language Models
- Evaluation, Judgment, and Public Discourse

Each session is composed of paper presentations and a discussion.

We thank all the authors and members of the PC for their invaluable contributions to make LUHME a very successful workshop. We also thank our keynote speaker. Finally, we thank the ECAI 2025 organizers for their support.

November 11, 2025

Henrique Lopes Cardoso
Rui Sousa-Silva
Maarit Koponen
Antonio Pareja-Lora

References

Allein, L., Trușcă, M. M., & Moens, M.-F. (2025). Interpretation modeling: Social grounding of sentences by reasoning over their implicit moral judgments. *Artificial Intelligence*, *338*, 104234. https://doi.org/10.1016/j.artint.2024.104234

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Gao, L., Tow, J., Biderman, S., & et al. (2021, September). A framework for few-shot language model evaluation. https://doi.org/10.5281/zenodo.5371628

Garcia-Varela, F., Bekerman, Z., Nussbaum, M., Mendoza, M., & Montero, J. (2025). Reducing interpretative ambiguity in an educational environment with chatgpt. *Computers & Education*, *225*, 105182. https://doi.org/10.1016/j.compedu.2024.105182

Hendrycks, D., Burns, C., Basart, S., & et al. (2021). Measuring massive multitask language understanding. *International Conference on Learning Representations*.

Kaplan, J., McCandlish, S., Henighan, T., & et al. (2020). Scaling laws for neural language models. *CoRR*, *abs/2001.08361*.

Lyons, J. (1990). *Language and linguistics: An introduction*. Cambridge University Press.

Mao, R., Ge, M., Han, S., Li, W., He, K., Zhu, L., & Cambria, E. (2025). A survey on pragmatic processing techniques. *Information Fusion*, *114*, 102712. https://doi.org/10.1016/j.inffus.2024.102712

Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., & Bowman, S. R. (2023). What do NLP researchers believe? results of the NLP community metasurvey. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16334–16368. https://doi.org/10.18653/v1/2023.acl-long.903

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Piantadosi, S., & Hill, F. (2022). Meaning without reference in large language models. *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Risku, H., & Rogl, R. (2020). Translation and situated, embodied, distributed, embedded and extended cognition. In F. Alves & A. L. Jakobsen (Eds.), *The Routledge handbook of translation and cognition* (pp. 478–499). Routledge.

Sayers, D., Sousa-Silva, R., Höhn, S., & et al. (2021). *The dawn of the human–machine era: A forecast of new and emerging language technologies* (tech. rep.). EU COST Action CA19102 'Language In The Human–Machine Era'. https://doi.org/10.17011/jyx/reports/20210518/1

Srivastava, A., Rastogi, A., Rao, A., & et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024). Large language models for education: A survey and outlook. https://arxiv.org/abs/2403.18105

Wang, X., Liu, X., Wang, L., Wu, S., Su, J., & Wu, H. (2025). A simple yet effective self-debiasing framework for transformer models. *Artificial Intelligence*, *339*, 104258. https://doi.org/10.1016/j.artint.2024.104258

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models [Survey Certification]. *Transactions on Machine Learning Research*.

Xu, H., Kim, Y. J., Sharaf, A., & Awadalla, H. H. (2024). A paradigm shift in machine translation: Boosting translation performance of large language models. *The Twelfth International Conference on Learning Representations*.

Zhou, X., Zhang, Y., Cui, L., & Huang, D. (2020). Evaluating commonsense in pre-trained language models. *Procs. 34th AAAI, New York, USA, February 7-12, 2020*, 9733–9740. https://doi.org/10.1609/aaai.v34i05.6523

Workshop Organizers

Henrique Lopes Cardoso
University of Porto, Portugal

Rui Sousa-Silva
University of Porto, Portugal

Maarit Koponen
University of Eastern Finland, Finland

Antonio Pareja-Lora
Universidad de Alcalá, Spain

Web Master

Felermino Ali
University of Porto, Portugal

Assistant

Karen de Souza
University of Eastern Finland, Finland

Programme Committee

Aida Kostikova, *Bielefeld University, Germany*
António Branco, *University of Lisbon, Portugal*
Barbara Lewandowska-Tomaszczyk , *University of Applied Sciences in Konin, Poland*
Belinda Maia, *University of Porto, Portugal*
Alípio Jorge, *University of Porto, Portugal*
Bram van Dijk, *Leiden University, Netherlands*
Chaya Liebeskind, *Jerusalem College of Technology, Israel*
Efstathios Stamatatos, *University of the Aegean, Greece*
Ekaterina Lapshinova-Koltunski, *University of Hildesheim, Germany*
Eliot Bytyçi, *Universiteti i Prishtinës "Hasan Prishtina", Kosovo*
Federico Ruggeri, *University of Bologna, Italy*
Lynne Bowker, *University of Ottawa, Canada*
Nataša Pavlović, *University of Zagreb, Croatia*
Sule Yildirim Yayilgan, *Norwegian University of Science and Technology, Norway*
Tharindu Ranasinghe, *Lancaster University, UK*

Keynote Speaker

Chloé Clavel, *Inria Paris*

Contents

Transparency and Social Dynamics in LLMs

Keynote Speaker

Understanding Social Interactions in the Era of LLMs the Challenges of Transparency

Chloé Clavel

Inria Paris

Abstract: Research on AI and social interaction is not entirely new it falls within the field of social and affective computing, which emerged in the late 1990s. To understand social interactions, the research community has long drawn on both artificial intelligence and social science. In recent years, however, the field has shifted toward a dominant focus on generative large language models (LLMs). These models are undeniably powerful but often opaque. In this talk, I will present our current work on developing machine learning approaches from classical methods to LLMs for modeling the socio-emotional layer of interaction, with a particular focus on improving model transparency. I will also briefly present some of the applications we are developing to support human skill development, particularly in the fields of education and health.

Bio: Chloé Clavel is a Senior Researcher in the ALMAnaCH team at Inria Paris, the French national research institute for digital science and technology. Her research interests lie in the areas of Affective Computing and Artificial Intelligence, at the crossroads of multiple disciplines including speech and natural language processing, machine learning, and social robotics. She works on computational models of socio-emotional behaviors such as sentiment, social stance, engagement, and trust in both humanhuman interactions (e.g., conversations in social networks or face-to-face settings) and humanagent interactions (e.g., conversational agents and social robots).

Building Common Ground in Dialogue: A Survey

Tatiana Anikina[a,*], **Alina Leippert**[a] **and Simon Ostermann**[a]

[a]German Research Center for Artificial Intelligence, Saarland Informatics Campus, Germany

Abstract. *Common ground* plays a crucial role in human communication and the grounding process helps to establish shared knowledge. However, common ground is also a heavily loaded term that may be interpreted in different ways depending on the context. The scope of common ground ranges from domain-specific and personal shared experiences to common sense knowledge. Representationally, common ground can be uni- or multi-modal, and static or dynamic.

In this survey, we attempt to systematize different facets of common ground in dialogue and position it within the current landscape of NLP research that often relies on the usage of language models (LMs) and task-specific short-term interactions. We outline different dimensions of common ground and describe modeling approaches for several grounding tasks, discuss issues caused by the lack of common ground in human-LM interactions, and suggest future research directions. This survey serves as a roadmap of what to pay attention to when equipping a dialogue system with grounding capabilities and provides a summary of current research on grounding in dialogue, categorizing 448 papers and compiling a list of the available datasets.

1 Introduction

Common ground has been studied in a variety of settings by linguists, computer scientists, and philosophers alike (see e.g. foundational work by Clark and Brennan [33], Stalnaker [160], Benotti and Blackburn [10]). Common ground in dialogue can be defined as a set of shared beliefs between the interlocutors. However, as pointed out in Markowska et al. [120], a more complete definition should also include other components such as shared desires, intentions, and goals. According to Larsson [92], grounding has two core "meanings". The first meaning is called *symbol grounding* and it is concerned with the process of mapping symbols (e.g., words) to perception and the world (e.g., objects, documents, images etc.). The second meaning refers to the process of interactively adding to common ground in dialogue, i.e. *communicative grounding*. In Natural Language Processing (NLP), the former meaning of symbol grounding is more commonly used, whereas in fields such as Cognitive Science and Dialogue Theory, grounding typically refers to the process of achieving mutual understanding, i.e. the latter meaning of communicative grounding. In this survey we consider both meanings and cover some relevant works on communicative grounding in Section 3 and focus mostly on symbol grounding in Section 4. Furthermore, Section 5 describes the process of building common ground (both communicative and symbolic) in the context of language models (LMs).

From a Natural Language Processing perspective, conversational grounding is important for building trustworthy dialogue systems that can reliably use shared knowledge in conversation [126]. Despite the widespread usage of chat-based LMs, common ground is still often overlooked and not evaluated when comparing the performance of different models. The question arises: **Can we trust LMs and their generated outputs without building some common ground first?** Ideally, interactions should happen in such a way that common ground is built between users and LMs for more efficient and trustworthy communication. In order to address this question, it is necessary to clearly define and separate different dimensions of common ground and also to reflect on how these can be modeled and evaluated. To that end, the contributions of this survey are as follows:

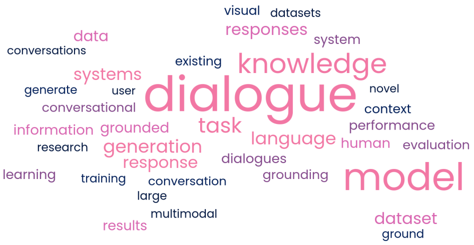


Figure 1. Word cloud based on the term frequency and abstracts of the papers mentioning "common ground" and "dialog" in the ACL Anthology.

(1) We describe different **dimensions of common ground** in dialogue that capture modality, type, and scope (Section 3);
(2) We survey **approaches towards modeling** common ground based on several grounding tasks (Section 4);
(3) We identify **potential problems** caused by the lack of common ground in LM-based dialogues (Section 5) and propose **future research directions** (Section 6).

2 Methodology

This survey focuses less on the theory of common ground (see Section 3 for definitions and some examples) and more on how it has been realized in NLP models (see Section 4 and 5). In order to survey current research and also cover a variety of common ground definitions and modeling approaches we started by collecting a list of all papers published in the ACL Anthology since 2015 that mention "common ground" or "grounding" in their abstracts together with "dialogue", and then examined them in terms of the definition of common ground and which approaches were used to model it. We started with 448 papers and focused on those that address different dimensions of common ground relevant for dialogue processing. Specifically, we focused on the papers that discuss modality (e.g., textual, visual, multimodal grounding), type (static vs. dynamic),

[*] Corresponding Author. Email: tatiana.anikina@dfki.de

and scope of grounding (commonsense, domain or contextual knowledge). Some papers were removed from the initial selection if their content was deemed irrelevant and some papers were added based on citation chaining. According to the classification of literature reviews outlined in Paré and Kitsiou [138], this survey can thus be considered as both *descriptive* and *narrative* since it aims to provide an overview of the available work and identify some trends for grounding in dialogue. Simultaneously, it is more focused on a qualitative interpretation of prior knowledge. The survey was conducted by a small team of researchers with experience in natural language processing who selected and judged papers independently. The inclusion of papers in the survey was determined based on the relevance of their topics to both grounding and dialogue as reflected in the abstract and assessed by the authors (see Appendix A for more detail and statistics). Figure 1 shows a word cloud based on the word frequencies of the most common terms from the abstracts of the surveyed papers. Unsurprisingly, *dialogue*, *model*, and *knowledge* are the most frequent terms; although terms such as *generation*, *context*, *visual*, and *multimodal* are also commonly used, which reflects the challenging and multifaceted nature of common ground research.

3 Common Ground Dimensions

Common ground contributes to successful communication through dialogue. The concept is thereby easy to define in terms of its relevance, but what information is needed to form common ground is fleeting. As noted by Chandu et al. [19], there is often no such thing as an "axiomatic common ground". Successful grounding in human-machine-communication is usually only evident in whether the goal of the task is reached; or through assessment of the quality of the conversation. While the common ground does consist of prior knowledge (such as world, commonsense or knowledge about previous events), much of the common ground of a conversation is built and established *during* communication. A system must thereby adapt to the evolving context of the conversation and the newly acquired knowledge.

There is not one way to establish a grounded dialogue between human and machine. As Chandu et al. [20] note, grounding is often performed with the goal of supporting a more defined end purpose task. How grounding is achieved heavily depends on the purpose of the conversation, e.g., whether the goal is to find a an object in a shared environment or to enrich a dull chit-chat with more interesting or personal, user-targeted facts. Researchers have proposed different methods for establishing common ground up to now. We aim to systematize the existing approaches to symbol grounding and categorize them in terms of the following dimensions (see Figure 2):

1. The **modality** through which the conversation is grounded (e.g. textual, visual, multimodal)
2. The **scope** of the grounded information (e.g. commonsense, domain-specific or contextual knowledge)
3. The **type** of grounding (static or dynamic)

Building on this classification, we present a roadmap for incorporating common ground into dialogue systems.

3.1 Modality

Dialogue participants often integrate external contexts into the conversation, and these become part of the common ground [162]. Grounding can thereby connect the conversation to the environment: Grounding utterances in the real world allows models to account for what is missing or cannot be learned from conversational data. There exist many forms of external contexts. Strub et al. [162] for example mention the physical environment, a collaborative task the participants work on, a map they use for coordination or a database they want to access. Real world contexts that ground a conversation can thus be derived from different modalities, which Chandu et al. [20] classify into:

- **Textual modality**: e.g. plain text, entities/events, knowledge bases
- **Non-textual modality**: e.g. images, speech, videos

In recent years, many tasks that go beyond a single modality (in NLP: the textual one) have been proposed with the help of neural architectures [137]. Parcalabescu et al. [137] address the need for an appropriate definition of multimodality when the information receiver and processor is a machine learning system. The authors propose a *task-relative* definition: The task determines what information is relevant and how it can be stored, thereby indicating under which circumstances multiple modalities are necessary. Only in cases where different language representations (e.g. speech and image of a text) cannot be converted into one another without losing task-relevant information, they depict multiple modalities.

3.1.1 Grounding in the Textual Modality

Conversations between user and agent can be grounded in additional textual input that goes beyond the conversation history. This could be an external knowledge graph or other textual sources, providing world or domain knowledge.

Textual resources can be used to incorporate knowledge from the human world into the conversation between human and machine. As an example, Ghazvininejad et al. [57] model knowledge-grounded conversations with the goal to produce more contentful utterances grounded in the real world, i.e. taking into account not only the conversation history, but also external facts. To achieve this they retrieve various facts from textual sources such as Wikipedia and Foursquare, selecting the facts relevant to the conversation context. Similarly, language understanding can be improved by injecting commonsense knowledge into a conversation via knowledge graphs, providing background knowledge that machines otherwise lack as this information cannot be learned merely from conversational data [218, 148, 198, 139]. The additional knowledge can also come in the form of domain knowledge, as in Zhu et al. [222]. Focusing on the example of the music domain, their system uses structural background knowledge represented in the knowledge base to discuss and recommend songs to a user.

Other works focusing on knowledge base (KB) integration and LM-based knowledge generation include, e.g. the work by Chen et al. [23] that introduces a neural agent who can interact with KBs via generated SQL queries. The agent learns to infer and confirm user intent, dynamically deciding when to ground user constraints into SQL queries to retrieve relevant information from KBs. Liu et al. [106] explore knowledge-grounded dialogue generation under low-resource conditions, introducing a knowledge-aware transformer and a three-stage learning framework that leverages large-scale dialogues and unstructured KBs. Liu et al. [109] introduce a multi-stage prompting approach for knowledgeable dialogue generation, which first generates knowledge from the dialogue context and then produces responses based on both the context and the previously generated knowledge. Li et al. [102] work on eliciting knowledge from LMs for unsupervised, knowledge-grounded conversations, demonstrat-

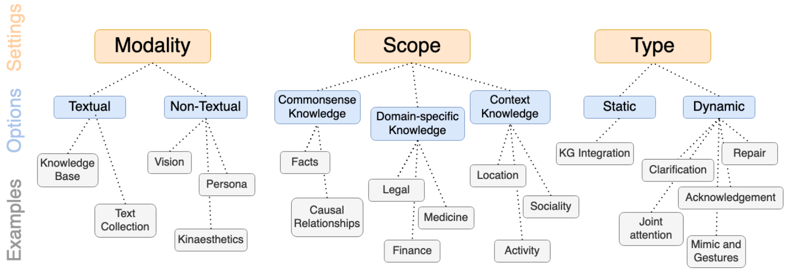


Figure 2. Common ground dimensions.

ing that LMs are capable of generating commonsense knowledge and summarizing facts not directly accessible via search engines.

3.1.2 Multimodal Grounding

Besides the textual modality, an increasing number of grounding-related tasks is multimodal, for example modeling an interplay between language and vision. Another possibility is to ground a dialogue in what is specific to the user, e.g. emotion or persona [51]. Pustejovsky and Krishnaswamy [143] argue that multimodal dialogue includes multiple aspects: (1) *Co-situatedness* and *co-perception* of the agents (i.e. how they perceive the environment and interpret the situation), (2) *Co-attention* of a shared situated reference (e.g. referring to the objects in the environment through language, gestures, visual clues), and (3) *Co-intent* of a shared goal which is especially relevant for problem-solving and collaborative tasks.

Grounding in Vision Possible multimodal tasks are visual games [162] or holding a dialogue about visual scenes [36], where shared representations help to ground meaning. Referring expression comprehension with visual features [67] is another important aspect of building common ground. A range of work where multimodal grounding is relevant focuses on dialogue applications. Such applications of visually-grounded systems can reach beyond interacting with a smart assistant, e.g. by helping visually impaired users to understand their surroundings or online content [22, 60]. They can also help to quickly gain an overview in search and rescue missions where an operator is 'situationally blind' but can interact via language [36] and contribute to grounding in the shared physical context [200].

Grounding in Persona More personalized and engaging conversation can be approached by grounding a dialogue in user-specific attributes such as persona or emotion [119, 186, 24, 203]. These attributes can be represented both through a textual and non-textual modality. A persona can be formed through a textual profile description of the user, including e.g. their personal interests and occupation [208]. Distributed representations for a persona can also be learned through conversational content such as a user's speaking style [99]. A multimodal approach for persona grounding is introduced by Ahn et al. [3], who enrich facts about user's personality with their pictures posted on social media and the corresponding comments.

Grounding in Kinaesthetics Another modality that can ground a dialogue by incorporation of the environment of user and system is kinaesthetics. The perception of one's own body position and movements is an important dimension for embodied systems in human-computer interaction [48]. Navigation agents that ground the conversation in the environment can help a user navigate through space [111], helping for instance persons with visual impairments, by retrieving landmark destinations and providing visual information to the user [108]. Communication grounded in spatial dimensions and actions can also be used to help a tourist reach a target location [39]. For such tasks, landmark recognition, user localization and natural language instructions are needed. Navigation instructions further require grounding in visual objects (e.g. "stop at the door") and geometric structure and directions (e.g. “turn left”) [65].

Other modalities to ground the conversation in include e.g. gaze and nodding [5]. As early as in 1996, Dillenbourg et al. made the observation that the process of grounding is not bound inside one interaction mode but instead crosses different modalities. They suggest that agents should instead be capable of *modality-independent* grounding mechanisms, flexibly adjusting to a conversation's interaction-style. For instance, speech can serve as the primary source of new information, while visual cues like head movements, facial expressions, and hand gestures act as a *backchannel* to support grounding. In a similar vein, text can function as the main communication channel in text-based interactions while accompanying illustrations can also help establish common ground.

3.2 Scope

Communicating on human terms requires more than just knowing the meanings of words; It demands a deep, integrated understanding of how, when, and why to use them. An important factor for successful communication is an “understanding of the shared world” [9]. Clark [32] proposes to define different types of scope of common ground, namely **communal** common ground (speaking the same language, sharing a hobby or profession, leading to a communal lexicon of technical jargon and naming conventions) and **personal** common ground, i.e. joint (linguistic) experiences, leading to a private lexicon.

Based on our survey of recent publications and how they define the scope of common ground, we widen this classification and propose to distinguish between commonsense knowledge, domain-specific, and contextual knowledge.

Commonsense Knowledge. It is fundamental to everyday conversations and therefore plays a substantial role for the grounding process of a conversation. People use commonsense knowledge to understand and enrich conversations with related information or to picture what they do not understand. Moreover, commonsense knowledge enables reasoning about previously unseen events [154, 135]. Building a model grounded in commonsense knowledge is a challenging task, largely because there is no clear definition and there are certain aspects of commonsense knowledge that only come to light in the corresponding situations. As commonsense knowledge is rarely made explicit in natural conversation [37, 58] since it is assumed to be universally shared, it is often not represented in conversational datasets and dialogue systems lack grounding in the real world [57].

Currently, no universally agreed upon strategy for encoding commonsense knowledge exists [149]. Commonsense knowledge can either be **implicit** in the training data or **explicit** in external knowledge sources. A common approach for adding explicit commonsense knowledge to a dialogue system is to harness an external KG, such as ConceptNet [159] or Atomic [154]. To enrich the conversation, the response generation is conditioned on the previous utterance and the related external facts (as in Young et al. [197] or Zhou et al. [218]). While grounding in external commonsense knowledge is helpful especially for concepts which are poorly represented in conversational training data [204], Davison et al. [37] point to the drawback that commonsense knowledge bases do contain high-quality information, but their coverage would be low. Richardson and Heck [149] observe a shift from grounding with KGs to using neural models for learning commonsense knowledge implicit in text data. Importantly, popular LMs produce natural sounding text but the responses may fail to integrate correct knowledge and the facts can be distorted [66]. Moreover, Bian et al. [12] observe that while GPT performs well on many commonsense benchmark tasks, it has drawbacks in domains which require a deeper understanding of human behaviour, such as social or temporal commonsense knowledge. Commonsense can not be learned from descriptions only, it requires both reasoning and inference abilities to uncover implicit and situational knowledge.

Domain Knowledge. Domain knowledge plays a crucial role in task-oriented dialogues and the required knowledge goes beyond commonsense or what can be inferred based on the conversational history, such knowledge varies a lot across different domains and often requires an access to some ontology or specialized knowledge base. For instance, when booking a hotel, it is important to know about the details like available room types, dates, and services. In the context of emergency response [62, 6], domain knowledge may include information about the responder roles, their responsibilities and equipment. The dialogue system should be able to correctly interpret and utilize the specialized terminology, e.g. it must know that UGV stands for an Unmanned Ground Vehicle in this domain.

There are different ways of integrating domain knowledge in dialogue systems apart from the direct fine-tuning on the domain data. Qian and Yu [144] propose a domain-adaptive dialogue generation approach called *DAML* (*Domain-Adaptive Meta-Learning*) that employs a two-step gradient update during training, allowing the dialogue system to capture general features across various tasks while enhancing its sensitivity to new domains, so that it can efficiently adapt with just a few training samples. Pryor et al. [142] employ a neural-symbolic approach to incorporate symbolic knowledge into the latent space of a neural model, effectively integrating domain knowledge and guiding the induction of dialogue structure. Suresh et al. [167] introduce a dialogue generation framework that can generate high-quality dialogue data for different domains using LMs and Chain-Of-Thought approach [181].

Contextual Knowledge. Situational grounding requires linking the content of an utterance to its *meaning in the specific context* [143]. The context heavily determines the utterance's interpretation. In situated dialogue, where conversation partners share time and space, grounding can take the form of links to entities in this shared space [76]. However, context has many dimensions which go beyond just a shared space in which an utterance is voiced. Context carries memories of previous utterances, the background or purpose of the conversation, the interrelation and dynamics between the conversation partners, including social and emotional connotations [64].

A reliable and efficient conversational system should adapt to the former context dimensions, along with interactions users feel comfortable with in a specific context. Moreover, different communication styles can be preferred depending on culture, e.g. regarding the expressiveness of emotions, rhetorical style and directness [122].

Katayama et al. [75] define the following contextual dimensions: Location (e.g. home or public), Sociality (e.g. alone or group), Activity (e.g. walking or driving) and Emotion (e.g. neutral or happy). Whether a user feels excited or annoyed, is busy or has time for chit-chat, should receive consideration in finding a suitable conversation strategy, e.g. by eliciting a more discreet as opposed to an entertaining continuation [105]. Kola et al. [86] encourage situation-awareness in agent development such that "agents should provide support that is consistent with the user's goals and preferences", taking into account *situation cues* and *social relationship features*. The four context dimensions proposed by Katayama et al. [75] provide a starting point for assessing a system's context considerations and can be expanded according to task and goal.

3.3 Type

Regarding the type of grounding, we distinguish between **static-symbolic grounding** and **dynamic-collaborative grounding**, following findings in literature including [93, 11, 19] (Table 1).

- **Static-symbolic grounding**: The common ground is the ground truth external data, e.g. a KG or the shared perceptual environment.
- **Dynamic-collaborative grounding**: The common ground is formed interactively, e.g. through clarification and negotiation between user and agent.

A static-symbolic approach is used in Ji et al. [70], who use knowledge graph grounding to reduce hallucinated responses. What is missing from static-symbolic approaches, as emphasized by Benotti and Blackburn [11], is the aspect of *error recovery* through negotiation of meaning, which becomes relevant in dynamic-collaborative grounding. In the dataset *GrounDialog*, Zhang et al. [211] focus on dialogues where participants are provided with dissenting information. The naturally arising need to negotiate and clarify therein automatically leads to dynamic-collaborative grounding. Grounding success in a dynamic setting depends on effectively communicating the mutually shared information until a common ground between user and agent is established, while in static grounding it relates to the ability of the agent to successfully link the query to the data [19].

Table 1. Static vs. dynamic types of grounding.

	Static-symbolic	Dynamic-collaborative	Grounding Motivation
Larsson [93]	*Symbol Grounding:* Connect symbols (e.g. words) to world via perception.	*Communicative Grounding:* Interactively update CG in dialogue	Speakers need to converge on shared meaning.
Chandu et al. [19]	*Static Grounding:* CG is the external data, assuming its universality.	*Dynamic Grounding:* CG is built via interaction and clarification.	Axiomatic CG does not exist and needs to be established in real world.
Benotti and Blackburn [11]	*Symbol Grounding:* Link symbols with perception, e.g. language grounded in vision.	*Collaborative Grounding:* Reach mutual understanding incrementally through dialog.	Human perception is unstable and depends on memory, capabilities, perspective.

4 Modeling Approaches

In this section, we provide an overview of recent modeling approaches defined by several common grounding tasks: knowledge (static and dynamic), vision, and persona grounding. We acknowledge that this categorization is not exhaustive, and there are more grounding-related tasks that can be considered (e.g., kinaesthetics and multimodal grounding). However, for the purpose of this survey, we focus on three distinct categories of tasks that are prominent in the current scientific literature on grounding.

4.1 Knowledge Grounding Tasks

Knowledge-based grounding can be based on static or dynamic knowledge with a single or multiple sources of knowledge that need to be integrated for successful communication.

4.1.1 Static Knowledge Grounding

Static knowledge integration typically involves external knowledge bases, graphs, or document collections. For instance, Zhao et al. [215] propose a model (*KnowledGPT*) that uses a knowledge selection module and **jointly optimizes selection and response generation**. KnowledGPT consists of a context-aware encoder and a knowledge selector, trained with a policy-gradient method and a curriculum step that distinguishes between the "hard" and "easy" materials for grounding.

Feng et al. [47] propose the MultiDoc2Dial task and a dataset for modeling **goal-oriented dialogues that are grounded in multiple documents**. The task is to identify which parts of which documents are relevant at each dialogue turn. MultiDoc2Dial task focuses on (1) extracting the grounding span from the document collection and (2) generating the dialogue response given the history and the extracted spans.

Wu et al. [188] address knowledge-grounded dialogue generation with their *Section-Aware Commonsense Knowledge-Grounded Dialogue Generation with Pre-trained Language Model* (*SAKDP*). SAKDP utilizes a PriorRanking network with contrastive learning to estimate the **relevance of the retrieved knowledge facts**. All candidates are clustered into three groups according to their priority. SAKDP then uses section-aware strategies to encode knowledge in a linearized way and applies LMs to encode only the high-priority facts, thus making the encoding process more efficient. Another system called *PLUG* [101] **unifies different knowledge sources** for knowledge-grounded dialogue generation. The approach retrieves relevant information from various sources (e.g. wiki, dictionary, knowledge graph), converts the extracted knowledge into textual format and combines it with the dialogue history.

Chen et al. [26] focus on the task of **knowledge grounded dialogue generation with in-context learning**. Their goal is to produce faithful and informative responses that rely on the dialogue history as well as the knowledge base. To this end, they propose a retrieval-based framework, *IKA* (In-context Knowledge grounded dialogue Augmenter), combining in-context learning with retrieval techniques and adding the most relevant and diverse demonstrations to the LLM prompt for response generation.

As a part of static grounding, Xie et al. [192] consider **structured knowledge grounding** (*SKG*) and propose the *UnifiedSKG* framework that can standardize different task representations (e.g. semantic parsing, question answering, fact verification). The main idea behind UnifiedSKG is to unify different forms of structured knowledge through linearization. Xie et al. [192] also show that task-specific knowledge can be effectively shared via multi-task prefix tuning, improving the overall performance on the target task.

Another direction for static grounding is to ground the conversation in **social media interactions**. Choudhary and Kawahara [31] emphasize that most of the current work on knowledge-grounded dialogue focuses either on persona or fact-based structured knowledge. Thus, they propose a different approach and present a system that can mimic human responses through modeling social media interactions by training a joint retriever-generator on a mixture of open-domain dialogue data and a collection of Reddit comments.

Other recent work that incorporates static knowledge into dialogue processing uses knowledge graphs and performs entity-agnostic representation learning [220], generates dialogue acts to guide generation through tree-structured reasoning [114], focuses on document-grounded conversations, and uses graphs to capture the inter- and intra-document relations [194].

4.1.2 Dynamic Knowledge Grounding

Dynamic knowledge grounding happens when common ground is formed interactively. This is often achieved through negotiation and clarification [185, 134, 117]. Dynamic changes in common ground can be also modeled as knowledge updates. For instance, Tuan et al. [173] introduce the task of **dynamic knowledge-grounded conversation generation**. They pair every dialogue turn with a knowledge graph that includes a collection of triplets representing entities and relations between them (e.g. *"x IsEnemyOf y"*). The grounding task in this setting involves (1) text generation conditioned on the textual input plus the corresponding knowledge graph and (2) generation of relevant entities after each update of the graph.

Tuan et al. [173] propose a model (*Qadpt*) that predicts the knowledge graph entities and retrieves the relational paths in the graph by applying multi-hop reasoning. Qadpt proves to be beneficial even for zero-shot adaptation with dynamic knowledge graphs. Similarly, topic-grounded dialogues also require keeping track of **topic transitions throughout a conversation**. Wen et al. [184] present a model called Sequential Global Topic Attention (*SGTA*). It uses a latent space to integrate the global-level and sequence-level information and predicts the topic based on the distribution sampling. SGTA exploits topic co-occurrences and models post-to-response topic transitions as well as predicts the next likely topic in dialogue.

Udagawa and Aizawa [175] focus on creating and maintaining common ground in **dynamic environments**. Specifically, they collect a dataset of 5,617 dialogues (*OneCommon* Corpus) that represents entity attributes and their temporal dynamics based on continuous values that correspond to entity movements. Udagawa and Aizawa [175] consider a collaborative reference task as a multi-agent cooperative game. Each agent can observe several entities and exchange information about them with other agents. The task is accomplished successfully if all the agents select the same entity at the end of the game. The proposed model encodes dialogue utterances and utilizes spatial and temporal encoders to integrate the dynamic features.

4.2 Vision Grounding Tasks

Vision grounding is crucial for conversations that revolve around the content of images or videos and there are tasks such as visual dialogue generation and image grounded question answering.

Many multi-modal extensions of Transformer models (e.g., VL-BERT [163], VideoBERT [164], LXMERT [170], MTN [96], GTR [18], TransVG++ [41]) allow modeling both texts and images simultaneously and can be applied to such tasks. Below we exemplify several visual grounding tasks and showcase some models for image grounded conversations, visual and video-centered dialogue.

Mostafazadeh et al. [131] introduce the task of multi-modal **image grounded conversations** where natural-sounding conversations are generated about some shared image. This task has both elements of chit-chat and goal-oriented dialogue since the image constrains the topic of conversation.

Kang et al. [74] investigate the task of **reference resolution in visual dialogue**. The goal of this task is to answer a series of questions grounded in some image given the visual input together with the dialogue history. The authors propose Dual Attention Networks (*DANs*) to perform visual reference resolution. Their model consists of two attention modules: *REFER* and *FIND*. First, REFER applies multi-head attention mechanism to learn the relations between the question and the dialogue history. Next, FIND receives as input both image features and the outputs of the REFER module and combines them to perform visual grounding.

Kim et al. [79] address the **visual dialogue grounding task** in the context of question answering. They find that some questions can be answered by only looking at the image while others require both image and dialogue history. Therefore, they decide to maintain both models (image-only and image-history) and combine them in different ways. Specifically, they experiment with ensembling and consensus dropout fusion with shared parameters. The combined model demonstrates complementary gains for image-only and image-history models.

Video-grounded dialogue is also explored in [140, 95, 180]. The video grounding task involves modeling **video features across both spatial and temporal dimensions** as well as dialogue features that include dialogue history and interactions between the turns. Le and Hoi [95] extend GPT-2 models and formulate a video-grounded dialogue task as a sequence-to-sequence processing that combines both visual and textual representations. Their proposed model *VGD-GPT2* captures dependencies between different modalities at the spatio-temporal level (for videos) and token-sentence level (for dialogues).

Qin et al. [145] approach the task of **answering video-grounded questions in dialogue** using a Dual Temporal Grounding-enhanced Video Dialog model (*DTGVD*) that predicts turn-specific temporal regions while filtering out irrelevant video content and grounding free-text response in both video frame and dialog history. The proposed approach is based on the UniVL [112] visual-language model. DTGVD finds temporal segments in the dialog as well as contextually relevant video segments and grounds the response generation in both. This approach employs contrastive learning by utilizing grounded turn-clip pairs as positive samples and other turn-clip pairs as negative ones. The model is trained with answer generation and contrastive losses and achieves state-of-the-art results on several benchmark datasets for video-grounded dialogues.

Wang et al. [180] introduce a new Video-grounded Scene&Topic AwaRe dialogue (*VSTAR*) dataset and propose benchmarks for dialogue understanding based on **scene and topic segmentation as well as video-grounded dialogue generation**. Their experiments demonstrate that visual information is very important for the topic boundary detection and including such information can improve the performance by 7.1% F1. They also show that segment information is helpful for dialogue generation and the current encoder-decoder models still struggle to make full use of the visual input for the video-grounded dialogue generation.

4.3 *Persona Grounding Tasks*

Lim et al. [104] emphasize that it is important to ground **external knowledge and persona** simultaneously and propose a model called *INFO* that grounds persona information together with the external knowledge. They implement the knowledge and persona selector for the grounding task using poly-encoder and adopt retrieval-augmented generation to reduce the hallucinations and generate more coherent and engaging responses.

Majumder et al. [119] explore persona-grounded dialogue and focus on inferring simple implications of persona descriptions. For instance, if someone likes hiking, they probably also like nature. Majumder et al. [119] utilize commonsense knowledge bases to expand the set of persona descriptions. They also experiment with the **fine-grained persona grounding**, so that the model has to choose between different persona sentences when generating a dialogue response. To this end, the model uses variational learning to sample from various persona descriptions achieving good scores on the Persona-Chat dataset [209] with consistent and diverse responses.

Wang et al. [179] introduce a framework for **decoupling knowledge grounding** into different sources to aid response generation (e.g. persona, documents, memory). Their framework (*SAFARI*) can make a decision of whether to include a specific knowledge source and when to do so while generating a response. *SAFARI* has three modules that are responsible for planning, retrieval and assembling of information. Wang et al. [179] also construct a personalized knowledge-grounded dialogue dataset (*Knowledge Behind Persona*) where responses are conditioned on multiple knowledge sources for more informative and persona-consistent dialogues.

Gao et al. [53] focus on conversational agents that can infer listener's personas to generate appropriate responses and maintain consistent speaker profiles. They introduce *PeaCoK*, **a large-scale persona commonsense knowledge graph** with 100K human-validated facts, structured around five persona dimensions: characteristics, routines and habits, goals and plans, experiences, and relationships. *PeaCoK* is built by extracting and generating persona knowledge from existing knowledge graphs (ATOMIC [153], COMET [15]) and LMs.

Other related work includes modeling partner personas in dialogue [110], disentangling and recombining persona-related and persona-agnostic parts of the dialogue response [186] as well as personalized conversation generation based on journal entries [136].

5 Common Ground in LMs

Effective conversation requires building common ground that is ideally multimodal, dynamic, integrates commonsense and contextual knowledge. However, in the modern age of large LMs (LLMs) common ground is typically **defined by the user** or it is based on the **static training data**. For instance, prompts can be used to include persona, domain and task information that is supposed to represent shared knowledge, but this information is tailored in a way that reflects the user's point of view. Moreover, inputs to LLMs are usually based on text even when they reference other modalities (e.g. describing location, time or emotions).

LLM outputs may also contain hallucinations or incorrect assumptions. Jiang et al. [71] distinguish between two potential sources of factual hallucinations: **insufficient knowledge** within the model's parameters and knowledge memorization coupled with the **lack of generalization** capability. However, some hallucinations can also be a product of miscommunication caused by the lack of common ground: Shaikh et al. [157] show that LLMs tend to generate text with less conversational grounding (on average, 77.5% less likely compared to humans) and often presume common ground instead of building it incrementally over time.

Language models often exhibit reduced rates of grounding acts and show **poor grounding agreement with humans**. Existing supervised fine-tuning and preference optimization datasets are potential sources of this problem [157], because such datasets are meant for training models that simply "follow" instructions based on a limited number of interactions. Models trained on such data never learn how to build common ground throughout the conversation and adjust it depending on new inputs.

Jokinen [72] also emphasize the need for grounding in LM-based dialogue systems and argue that such systems need to build a shared understanding of dialogue context and intents, grounding generated utterances in real-world events and potentially bridging the gap between neural and symbolic computing. Furthermore, Schneider et al. [156] benchmark both open- and closed-source LMs on the grounding tasks and find that both are equally good at classifying grounding acts but **identifying grounded knowledge** proved to be very challenging and it is better handled by close-source LMs.

Another important issue is that LLM-based conversational agents may fail to generate safe and appropriate responses [40] and often go along with a problematic user input, generating offensive and toxic language. Kim et al. [84] aim to address this issue by proposing *GrounDial* that **grounds dialog responses in commonsense social rules** and does not require any additional fine-tuning. GrounDial combines in-context learning with guided decoding that follows human norms to generate more safe and appropriate responses.

Yu et al. [198] emphasize that commonly used techniques for dialogue response generation are based on **Chain-of-Thought (CoT)** [181] or **Retrieval Augmented Generation (RAG)** [214]. However, both methods have important drawbacks. On the one hand, CoT may overestimate the capabilities of LLMs by treating them as isolated knowledge sources, while the knowledge stored in LLMs can be outdated, and LLMs are prone to hallucinations [69]. On the other hand, approaches like RAG underplay the internalized knowledge of LLMs and mostly rely on external sources. Yu et al. [198] propose a different approach that considers **LLM as a collaborator** and includes several Thought-then-Generate stages to identify knowledge demands and then find relevant information via Demands-Guided Knowledge Retrieval.

Chiu et al. [28] draw attention to another limitation of LLMs in the context of grounded task-oriented dialogue: LLMs are difficult to steer towards **task objectives** and they have difficulties with handling **novel grounding**. To address these limitations Chiu et al. [28] propose an interpretable grounded dialogue system that combines **LLMs with a symbolic planner** to perform grounded code execution and response generation. The proposed system has a reader that uses LLM to convert utterances into executable code functions which represent the core meaning and map language to symbolic actions, and a symbolic planner that can plan over the symbolic actions and determine the next response. The task progress is tracked via Bayesian reasoning and information gain objective. This approach achieved promising results on the OneCommon task from [174] that involves collaborative reference resolution.

Factual inconsistency of LLMs [152] is another important concern for building a conversational agent. Previous work has shown that LLMs can generate factually incorrect responses even when provided with valid knowledge sources [69]. Xue et al. [195] tackle the inconsistency issue in knowledge-grounded conversations by enhancing the factual knowledge expression via extended Feed Forward Networks (FFN) in Transformers and apply reinforcement learning that implicitly aligns dialogue responses with gold knowledge using factual consistency preference. The parameters in extended FFNs are updated based on the knowledge-related tokens that appear in both grounding knowledge and response, e.g., if the term "Argentina" is part of the response it can be enhanced with the factual knowledge "Argentina won the 2022 World Cup champion."

Daheim et al. [34] also investigate **factuality in the context of document-grounded dialogue generation** and consider two components: a simple ungrounded response generation model that encourages fluency, and a component that encourages responses which can help reconstruct the response grounding document. The proposed method uses Bayes' Theorem to decompose the posterior distribution of a response given context and grounding into these components, and employs scaling factors to promote either greater correctness or fluency of the response. Although this approach results in improved factuality, the online decoding is **computationally demanding**.

Mohapatra et al. [127] provide a benchmark and design various tests to assess how well LLMs handle grounding as both speakers and listeners. As listeners, LLMs should integrate repaired or canceled information and identify ambiguous cases that need clarification. As speakers, they must generate accurate and unambiguous responses. Mohapatra et al. [127] analyze perplexity on the responses that are appropriate and grounded and responses that are fitting but contextually incorrect. They found **a strong correlation between conversational abilities and the size of the models and the pre-training data** with larger models and datasets leading to improved grounding capabilities and lower perplexity for correct responses.

Overall, building common ground with LMs is a very challenging task and the following issues that can hinder successful communication need to be taken into consideration:

1. **Lack of Shared Context** is an important factor as models typically do not have a persistent memory of previous interactions and do not have access to the same environment as the user (e.g., limited access to spacial and visual information). Integrating long-term memory and multimodal capabilities can help with alleviating this problem.
2. **Ambiguity and Underspecification** pose another challenge since humans often rely on implicit meanings, assumptions, and can use vague phrasing [178]. Clarifications are needed to resolve the ambiguities in conversation, the model needs to identify ambiguous knowledge and generate appropriate clarification requests [98].
3. **Trust and Interpretability** are essential for efficient communication and although LMs do not have own intentions, emotions, and self-awareness, users may anthropomorphize them [42, 155] and make incorrect assumptions based on that. Having more transparent and interpretable inner workings of LMs is a very desirable property as it can inform the user e.g. which knowledge was taken into account when generating the response.
4. **Knowledge and Response Misalignment** is a problem caused by the fact that LMs typically rely on static knowledge and have difficulties adapting to user intents if they change throughout the conversation. More research on dynamic grounding should be conducted to build flexible grounding models. Also, misalignment in tone or style can result in miscommunication and confusion. LMs struggle with emotional response, sarcasm and humor that are typical in human-to-human conversations [30] and LM responses need to be better adjusted to user persona and interaction style.

6 Current and Future Research Directions

In order to visualize the distribution of topics that are most prominent in the current research on dialogue grounding, we applied k-means clustering to the abstracts of 448 papers selected for the survey based on keyword match. We then categorized them according to common topics based on the top 25 words per cluster (after stopword removal). Figure 3 shows that almost 16% of all papers focus on grounded generation and selection, while 15.2% of the papers are concerned with cognitive and human-centered aspects of grounding. Knowledge grounding and LLM fine-tuning are two other relatively big topics (13% each). Interestingly, multimodal papers represent only 7.4% with a similar number of papers dedicated to learning-based approaches. Multimodal papers with emphasis on spatial grounding are even less prominent (only 5.37%), and, quite worryingly, only 2% of all publications focus on evaluation and benchmarking. These trends show that there are some considerable gaps in the current research on conversational grounding.

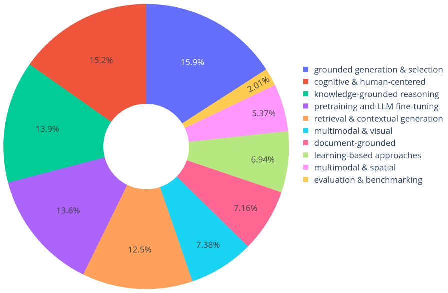


Figure 3. Distribution of topics relevant to conversational grounding based on the papers published in the ACL Anthology since 2015.

As demonstrated in this survey, well-grounded dialogue requires some commonsense knowledge as well as contextual knowledge that heavily depends on the dialogue history and previous interactions, and it may also include some domain-specific knowledge that needs to be derived based on the shared environment, experiences and expectations. Moreover, we believe that the **interdisciplinary collaboration** between different fields such as robotics, cognitive science and linguistics will greatly benefit this research area. There is also a need for more **diverse and realistic datasets** that are not purely text-based or combine text with just one modality e.g. images, the community needs a collection of various types of data that capture different dimensions of grounding with respect to modality, type and scope. More often than not research on grounding addresses the static setting which assumes that we have an access to some graph or a knowledge base, but grounding is a collaborative process and more research needs to be done on incorporating **dynamic features**.

Another important aspect that has only recently started gaining more attention is **evaluation**. For instance, Alghisi et al. [4] investigate the impact of incorporating external knowledge to ground dialogues with retrieval-augmented generation or gold knowledge and emphasize the importance of human evaluation. Chaudhary et al. [21] find that LLM-based evaluation does not align well with human judgments and show that evaluation results are not robust against perturbations. Ghaddar et al. [56] argue that knowledge-grounded dialogue needs to be more thoroughly evaluated with respect to hallucinations.

An important future avenue is research comparing common ground in **human-human and human-robot interactions**. For both, there is a need to collect more diverse data and combine LLM-based generation with neuro-symbolic approaches. For instance, Bonial et al. [13] use an Abstract Meaning Representation formalism to ground language concepts in the robot's world model, and Torres-Foncesca et al. [172] investigate an important dimension of knowledge grounding related to object permanence, i.e., the ability to maintain mental representations of objects even when they are not in view.

More research should be done towards **integrating common ground during pre-training** and fine-tuning stages of LLMs, e.g., by introducing additional loss functions and contrastive learning to distinguish between compatible and incompatible beliefs being formed in a conversation (or by measuring perplexity between correct and adversarial responses [127] and combining generation with grounding reconstruction [34]). Some recent works explore how one can build a knowledge-grounded dialog system that utilizes both dialog history and local knowledge base for response generation with a semi-supervised pre-training [202] and perform large-scale multi-party aware pre-training on conversational data [8] that shows promising results for knowledge grounded conversations.

7 Conclusion

In this survey we provided an overview of different dimensions of common ground and categorized them according to the modality, type and scope. We also discussed existing modeling approaches for knowledge-based, visual, and persona-based grounding, exemplifying promising research directions and attempts to integrate various aspects of common ground. We talked about leveraging common ground in LLMs, and summarized the issues related to the lack of conversational grounding in such models. We also described current and promising future research directions. We hope that this survey and our annotations[1] will serve as a guide for exploring the broad and dynamic landscape of conversational grounding.

Limitations

The current survey represents just a snapshot of the research on the topics of conversational grounding. This is an interdisciplinary field that ideally involves collaboration between the researchers who work on language, vision, robotics, and cognitive modeling. This work may not include all the relevant and very recent publications due to its scope and focus on the ACL Anthology.

Acknowledgements

This project was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005). We would also like to thank the anonymous reviewers and Brielen Madureira Lasota for their helpful and rigorous feedback.

References

[1] S. Agarwal, T. Bui, J.-Y. Lee, I. Konstas, and V. Rieser. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.728. URL https://aclanthology.org/2020.acl-main.728.

[1] Please refer to the Appendix A and our GitHub repository https://github.com/tanikina/common-ground-in-dialogue for additional statistics, annotated papers and datasets.

[2] J. Ahn, Y. Song, S. Yun, and G. Kim. MPCHAT: Towards multimodal persona-grounded conversation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.189. URL https://aclanthology.org/2023.acl-long.189/.

[3] J. Ahn, Y. Song, S. Yun, and G. Kim. MPCHAT: Towards Multimodal Persona-Grounded Conversation. *arXiv preprint arXiv:2305.17388*, 2023.

[4] S. Alghisi, M. Rizzoli, G. Roccabruna, S. M. Mousavi, and G. Riccardi. Should we fine-tune or RAG? evaluating different techniques to adapt LLMs for dialogue. In S. Mahamood, N. L. Minh, and D. Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 180–197, Tokyo, Japan, Sept. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.inlg-main.15/.

[5] M. Alikhani and M. Stone. Achieving common ground in multi-modal dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 10–15, 2020.

[6] T. Anikina. Towards efficient dialogue processing in the emergency response domain. In V. Padmakumar, G. Vallejo, and Y. Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 212–225, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.31. URL https://aclanthology.org/2023.acl-srw.31/.

[7] R. Artstein, J. Boberg, A. Gainer, J. Gratch, E. Johnson, A. Leuski, G. Lucas, and D. Traum. The Niki and Julie Corpus: Collaborative multimodal dialogues between humans, robots, and virtual agents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[8] S. Bao, H. He, F. Wang, H. Wu, H. Wang, W. Wu, Z. Wu, Z. Guo, H. Lu, X. Huang, X. Tian, X. Xu, Y. Lin, and Z.-Y. Niu. PLATO-XL: Exploring the large-scale pre-training of dialogue generation. In Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, editors, *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 107–118, Online only, Nov. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-aacl.10. URL https://aclanthology.org/2022.findings-aacl.10/.

[9] C.-P. Bara, S. CH-Wang, and J. Chai. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. *arXiv preprint arXiv:2109.06275*, 2021.

[10] L. Benotti and P. Blackburn. A recipe for annotating grounded clarifications. *arXiv preprint arXiv:2104.08964*, 2021.

[11] L. Benotti and P. Blackburn. Grounding as a Collaborative Process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.41.

[12] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, and B. He. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*, 2023.

[13] C. Bonial, J. Foresta, N. C. Fung, C. J. Hayes, P. Osteen, J. Arkin, B. Hedegaard, and T. Howard. Abstract Meaning Representation for grounded human-robot communication. In J. Bonn and N. Xue, editors, *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 34–44, Nancy, France, June 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.dmr-1.4/.

[14] J. Bonn, M. Palmer, Z. Cai, and K. Wright-Bettner. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.601.

[15] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1470. URL https://aclanthology.org/P19-1470.

[16] B. Byrne, K. Krishnamoorthi, S. Ganesh, and M. Kale. TicketTalk: Toward human-level performance with end-to-end, transaction-based dialog systems. 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 671–680, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.55. URL https://aclanthology.org/2021.acl-long.55.

[17] J. Cao, A. Suresh, J. Jacobs, C. Clevenger, A. Howard, C. Brown, B. Milne, T. Fischaber, T. Sumner, and J. H. Martin. Enhancing talk moves analysis in mathematics tutoring through classroom teaching discourse. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7671–7684, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.513/.

[18] M. Cao, L. Chen, M. Z. Shou, C. Zhang, and Y. Zou. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.773. URL https://aclanthology.org/2021.emnlp-main.773.

[19] K. R. Chandu, Y. Bisk, and A. W. Black. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.375. URL https://aclanthology.org/2021.findings-acl.375.

[20] K. R. Chandu, Y. Bisk, and A. W. Black. Grounding 'Grounding' in NLP, June 2021.

[21] M. Chaudhary, H. Gupta, S. Bhat, and V. Varma. Towards understanding the robustness of LLM-based evaluations under perturbations. In S. Lalitha Devi and K. Arora, editors, *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 197–205, AU-KBC Research Centre, Chennai, India, Dec. 2024. NLP Association of India (NLPAI). URL https://aclanthology.org/2024.icon-1.22/.

[22] C. Chen, S. Anjum, and D. Gurari. Grounding answers for visual questions asked by visually impaired people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19076–19085. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01851. URL https://doi.org/10.1109/CVPR52688.2022.01851.

[23] C.-Y. Chen, P.-H. Wang, S.-C. Chang, D.-C. Juan, W. Wei, and J.-Y. Pan. AirConcierge: Generating task-oriented dialogue via efficient large-scale knowledge retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 884–897, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.79. URL https://aclanthology.org/2020.findings-emnlp.79.

[24] K. Chen, Q. Huang, D. McDuff, X. Gao, H. Palangi, J. Wang, K. Forbus, and J. Gao. NICE: Neural image commenting with empathy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4456–4472, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.380. URL https://aclanthology.org/2021.findings-emnlp.380.

[25] N. Chen, Y. Wang, H. Jiang, D. Cai, Y. Li, Z. Chen, L. Wang, and J. Li. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.570. URL https://aclanthology.org/2023.findings-emnlp.570/.

[26] Q. Chen, W. Wu, and S. Li. Exploring in-context learning for knowledge grounded dialog generation. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10071–10081, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.675. URL https://aclanthology.org/2023.findings-emnlp.675.

[27] Z. Chen, B. Liu, S. Moon, C. Sankar, P. Crook, and W. Y. Wang. KETOD: Knowledge-enriched task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2581–2593, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.197. URL https://aclanthology.org/2022.findings-naacl.197.

[28] J. Chiu, W. Zhao, D. Chen, S. Vaduguru, A. Rush, and D. Fried. Symbolic planning and code generation for grounded dialogue. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*,

pages 7426–7436, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.460. URL https://aclanthology.org/2023.emnlp-main.460/.

[29] H. Cho and J. May. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.218. URL https://aclanthology.org/2020.acl-main.218.

[30] M. Choi, J. Pei, S. Kumar, C. Shu, and D. Jurgens. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.699. URL https://aclanthology.org/2023.emnlp-main.699/.

[31] R. Choudhary and D. Kawahara. Grounding in social media: An approach to building a chit-chat dialogue model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 9–15, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-srw.2. URL https://aclanthology.org/2022.naacl-srw.2.

[32] H. H. Clark. Context and Common Ground. In *Encyclopedia of Language & Linguistics*, pages 105–108. Elsevier, 2006. ISBN 978-0-08-044854-1. doi: 10.1016/B0-08-044854-2/01088-9.

[33] H. H. Clark and S. E. Brennan. Grounding in communication. 1991.

[34] N. Daheim, D. Thulke, C. Dugast, and H. Ney. Controllable factuality in document-grounded dialog systems using a noisy channel model. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1365–1381, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.98. URL https://aclanthology.org/2022.findings-emnlp.98/.

[35] N. Daheim, J. Macina, M. Kapur, I. Gurevych, and M. Sachan. Stepwise verification and remediation of student reasoning errors with large language model tutors. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.478. URL https://aclanthology.org/2024.emnlp-main.478/.

[36] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[37] J. Davison, J. Feldman, and A. M. Rush. Commonsense knowledge mining from pretrained models. 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 1173–1178, 2019.

[38] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. GuessWhat?! Visual object discovery through multimodal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

[39] H. De Vries, K. Shuster, D. Batra, D. Parikh, J. Weston, and D. Kiela. Talk the walk: Navigating New York City through grounded dialogue. *arXiv preprint arXiv:1807.03367*, 2018.

[40] J. Deng, J. Cheng, H. Sun, Z. Zhang, and M. Huang. Towards safer generative language models: A survey on safety risks, evaluations, and improvements, 2023. URL https://arxiv.org/abs/2302.09270.

[41] J. Deng, Z. Yang, D. Liu, T. Chen, W. Zhou, Y. Zhang, H. Li, and W. Ouyang. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13636–13652, 2023. doi: 10.1109/TPAMI.2023.3296823. URL https://doi.org/10.1109/TPAMI.2023.3296823.

[42] A. Deshpande, T. Rajpurohit, K. Narasimhan, and A. Kalyan. Anthropomorphization of AI: Opportunities and risks. In D. Preoţiuc-Pietro, C. Goanta, I. Chalkidis, L. Barrett, G. Spanakis, and N. Aletras, editors, *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 1–7, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nllp-1.1. URL https://aclanthology.org/2023.nllp-1.1/.

[43] P. Dillenbourg, D. Traum, and D. Schneider. Grounding in multimodal task-oriented collaboration. In *Proceedings of the European Conference on AI in Education*, pages 401–407, 1996.

[44] T. Dong, A. Testoni, L. Benotti, and R. Bernardi. Visually grounded follow-up questions: A dataset of spatial questions which require dialogue history. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 22–31, 2021.

[45] M. Eric, L. Krishnan, F. Charette, and C. D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5506. URL https://aclanthology.org/W17-5506.

[46] S. Feng, H. Wan, C. Gunasekara, S. Patel, S. Joshi, and L. Lastras. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.652. URL https://aclanthology.org/2020.emnlp-main.652.

[47] S. Feng, S. S. Patel, H. Wan, and S. Joshi. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.498. URL https://aclanthology.org/2021.emnlp-main.498.

[48] J. Françoise, Y. Candau, S. Fdili Alaoui, and T. Schiphorst. Designing for kinesthetic awareness: Revealing user experiences through second-person inquiry. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5171–5183, Denver Colorado USA, May 2017. ACM. ISBN 978-1-4503-4655-9. doi: 10.1145/3025453.3025714.

[49] H. Fu, Y. Zhang, H. Yu, J. Sun, F. Huang, L. Si, Y. Li, and C. T. Nguyen. Doc2Bot: Accessing heterogeneous documents via conversational bots. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1820–1836, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.131. URL https://aclanthology.org/2022.findings-emnlp.131/.

[50] F. Galetzka, C. U. Eneh, and D. Schlangen. A corpus of controlled opinionated and knowledgeable movie discussions for training neural conversation models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 565–573, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.71.

[51] J. Gao, M. Galley, and L. Li. *Neural Approaches to Conversational AI: Question Answering, Task-Oriented Dialogues and Social Chatbots*. Now Foundations and Trends, 2019. ISBN 1-68083-552-1.

[52] J. Gao, Y. Lian, Z. Zhou, Y. Fu, and B. Wang. LiveChat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.858. URL https://aclanthology.org/2023.acl-long.858/.

[53] S. Gao, B. Borges, S. Oh, D. Bayazit, S. Kanno, H. Wakaki, Y. Mitsufuji, and A. Bosselut. PeaCoK: Persona commonsense knowledge for consistent and engaging narratives. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6569–6591, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.362. URL https://aclanthology.org/2023.acl-long.362/.

[54] F. Gervits, K. Eberhard, and M. Scheutz. Disfluent but effective? a quantitative study of disfluencies and conversational moves in team discourse. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3359–3369, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1317.

[55] F. Gervits, A. Roque, G. Briggs, M. Scheutz, and M. Marge. How should agents ask questions for situated learning? an annotated dialogue corpus. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 353–359, Singapore and Online, July 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.sigdial-1.37.

[56] A. Ghaddar, D. Alfonso-Hermelo, P. Langlais, M. Rezagholizadeh, B. Chen, and P. Parthasarathi. CHARP: Conversation history AwaReness probing for knowledge-grounded dialogue systems. In L.-W. Ku,

A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1534–1551, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.90. URL https://aclanthology.org/2024.findings-acl.90/.

[57] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. ISBN 2374-3468.

[58] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*, 2020.

[59] L. Golany, F. Galgani, M. Mamo, N. Parasol, O. Vandsburger, N. Bar, and I. Dagan. Efficient data generation for source-grounded information-seeking dialogs: A use case for meeting transcripts. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1908–1925, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.106. URL https://aclanthology.org/2024.findings-emnlp.106/.

[60] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00380. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html.

[61] J. Haber, T. Baumgärtner, E. Takmaz, L. Gelderloos, E. Bruni, and R. Fernández. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1184. URL https://aclanthology.org/P19-1184.

[62] T. K. Harris and A. I. Rudnicky. Teamtalk: A platform for multi-human-robot dialog research in coherent real and virtual spaces. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, pages 1864–1865. AAAI Press, 2007. URL http://www.aaai.org/Library/AAAI/2007/aaai07-307.php.

[63] B. Hedayatnia, D. Jin, Y. Liu, and D. Hakkani-Tur. A systematic evaluation of response selection for open domain dialogue. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 298–311, Edinburgh, UK, Sept. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sigdial-1.30.

[64] D. Hofs, M. Theune, and R. op den Akker. Natural interaction with a virtual guide in a virtual environment: A multimodal dialogue system. *Journal on Multimodal User Interfaces*, 3:141–153, 2010. ISSN 1783-7677.

[65] R. Hu, D. Fried, A. Rohrbach, D. Klein, T. Darrell, and K. Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation. *arXiv preprint arXiv:1906.00347*, 2019.

[66] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55, 2025. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

[67] T. Iki and A. Aizawa. Language-Conditioned Feature Pyramids for Visual Selection Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4687–4697, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.420. URL https://aclanthology.org/2020.findings-emnlp.420.

[68] N. Ilinykh, S. Zarrieß, and D. Schlangen. MeetUp! A corpus of joint activity dialogues in a visual environment. *arXiv preprint arXiv:1907.05084*, 2019.

[69] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

[70] Z. Ji, Z. Liu, N. Lee, T. Yu, B. Wilie, M. Zeng, and P. Fung. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, 2023.

[71] C. Jiang, B. Qi, X. Hong, D. Fu, Y. Cheng, F. Meng, M. Yu, B. Zhou, and J. Zhou. On large language models' hallucination with regard to known facts. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1041–1053, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.60. URL https://aclanthology.org/2024.naacl-long.60.

[72] K. Jokinen. The need for grounding in LLM-based dialogue systems. In T. Dong, E. Hinrichs, Z. Han, K. Liu, Y. Song, Y. Cao, C. F. Hempelmann, and R. Sifa, editors, *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 45–52, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.neusymbridge-1.5/.

[73] H. Kamezawa, N. Nishida, N. Shimizu, T. Miyazaki, and H. Nakayama. A visually-grounded first-person dialogue dataset with verbal and non-verbal responses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3299–3310, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.267. URL https://aclanthology.org/2020.emnlp-main.267.

[74] G.-C. Kang, J. Lim, and B.-T. Zhang. Dual attention networks for visual reference resolution in visual dialog. 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 2024–2033, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1209. URL https://aclanthology.org/D19-1209.

[75] S. Katayama, A. Mathur, M. Van den Broeck, T. Okoshi, J. Nakazawa, and F. Kawsar. Situation-aware emotion regulation of conversational agents with kinetic earables. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 725–731. IEEE, 2019. ISBN 1-72813-888-4.

[76] C. Kennington, S. Kousidis, and D. Schlangen. Interpreting situated dialogue utterances: An update model that uses speech, gaze, and gesture information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 173–182, 2013.

[77] I. K. Khebour, K. Lai, M. Bradford, Y. Zhu, R. A. Brutti, C. Tam, J. Tu, B. A. Ibarra, N. Blanchard, N. Krishnaswamy, and J. Pustejovsky. Common ground tracking in multimodal dialogue. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.318/.

[78] S. Khosla. Information extraction and program synthesis from goal-oriented dialogue. In V. Hudecek, P. Schmidtova, T. Dinkar, J. Chiyah-Garcia, and W. Sieinska, editors, *Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems*, pages 51–53, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.yrrsds-1.19/.

[79] H. Kim, H. Tan, and M. Bansal. Modality-balanced models for visual dialogue. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8091–8098. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6320. URL https://doi.org/10.1609/aaai.v34i05.6320.

[80] H. Kim, Y. Yu, L. Jiang, X. Lu, D. Khashabi, G. Kim, Y. Choi, and M. Sap. ProsocialDialog: A prosocial backbone for conversational agents. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.267. URL https://aclanthology.org/2022.emnlp-main.267/.

[81] J.-H. Kim, N. Kitaev, X. Chen, M. Rohrbach, B.-T. Zhang, Y. Tian, D. Batra, and D. Parikh. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1651. URL https://aclanthology.org/P19-1651.

[82] M. Kim, C. Kim, Y. H. Song, S.-w. Hwang, and J. Yeo. BotsTalk: Machine-sourced framework for automatic curation of large-scale multi-skill dialogue datasets. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empiri-*

cal Methods in Natural Language Processing, pages 5149–5170, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.344. URL https://aclanthology.org/2022.emnlp-main.344/.

[83] S. Kim, M. Eric, K. Gopalakrishnan, B. Hedayatnia, Y. Liu, and D. Hakkani-Tur. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.sigdial-1.35.

[84] S. Kim, S. Dai, M. Kachuee, S. Ray, T. Taghavi, and S. Yoon. GrounDial: Human-norm grounded safe dialog response generation. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1582–1588, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.109/.

[85] T. Kodama, H. Kiyomaru, Y. J. Huang, T. Okahisa, and S. Kurohashi. Is a knowledge-based response engaging?: An analysis on knowledge-grounded dialogue with information source annotation. In V. Padmakumar, G. Vallejo, and Y. Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 237–243, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.34. URL https://aclanthology.org/2023.acl-srw.34/.

[86] I. Kola, P. K. Murukannaiah, C. M. Jonker, and M. B. Van Riemsdijk. Toward social situation awareness in support agents. *IEEE intelligent systems*, 37(5):50–58, 2022. ISSN 1541-1672.

[87] D. Kontogiorgos, E. Sibirtseva, and J. Gustafson. Chinese whispers: A multimodal dataset for embodied language grounding. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 743–749, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.93.

[88] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, and M. Rohrbach. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 582–595, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1058. URL https://aclanthology.org/N19-1058.

[89] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.401. URL https://aclanthology.org/2021.emnlp-main.401.

[90] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi. Navigating connected memories with a task-oriented dialog system. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2495–2507, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.160. URL https://aclanthology.org/2022.emnlp-main.160/.

[91] J. Kruijt and P. Vossen. The role of common ground for referential expressions in social dialogues. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 99–110, Gyeongju, Republic of Korea, Oct. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.crac-1.10.

[92] S. Larsson. Grounding as a side-effect of grounding. *Top. Cogn. Sci.*, 10(2):389–408, 2018. doi: 10.1111/TOPS.12317. URL https://doi.org/10.1111/tops.12317.

[93] S. Larsson. Grounding as a side-effect of grounding. *Topics in cognitive science*, 10(2):389–408, 2018.

[94] D. Le, R. Guo, W. Xu, and A. Ritter. Improved instruction ordering in recipe-grounded conversation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10086–10104, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.561. URL https://aclanthology.org/2023.acl-long.561/.

[95] H. Le and S. C. Hoi. Video-grounded dialogues with pretrained generation language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5842–5848, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.518. URL https://aclanthology.org/2020.acl-main.518.

[96] H. Le, D. Sahoo, N. Chen, and S. Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1564. URL https://aclanthology.org/P19-1564.

[97] H. Le, C. Sankar, S. Moon, A. Beirami, A. Geramifard, and S. Kottur. DVD: A diagnostic dataset for multi-step reasoning in video grounded dialogue. 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 5651–5665, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.439. URL https://aclanthology.org/2021.acl-long.439.

[98] A. Leippert, T. Anikina, B. Kiefer, and J. Genabith. To clarify or not to clarify: A comparative analysis of clarification classification with fine-tuning, prompt tuning, and prompt engineering. In Y. T. Cao, I. Papadimitriou, A. Ovalle, M. Zampieri, F. Ferraro, and S. Swayamdipta, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 105–115, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-srw.12. URL https://aclanthology.org/2024.naacl-srw.12/.

[99] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[100] S. Li, Y. Yin, C. Yang, W. Jiang, Y. Li, Z. Cheng, L. Shang, X. Jiang, Q. Liu, and Y. Yang. NewsDialogues: Towards proactive news grounded conversation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3634–3649, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.224. URL https://aclanthology.org/2023.findings-acl.224/.

[101] Y. Li, B. Peng, Y. Shen, Y. Mao, L. Liden, Z. Yu, and J. Gao. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.15. URL https://aclanthology.org/2022.naacl-main.15.

[102] Y. Li, J. Zhao, M. Lyu, and L. Wang. Eliciting knowledge from large pre-trained models for unsupervised knowledge-grounded conversation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10551–10564, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.721. URL https://aclanthology.org/2022.emnlp-main.721/.

[103] Y. Li, D. Hazarika, D. Jin, J. Hirschberg, and Y. Liu. From pixels to personas: Investigating and modeling self-anthropomorphism in human-robot dialogues. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9695–9713, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.567. URL https://aclanthology.org/2024.findings-emnlp.567/.

[104] J. Lim, M. Kang, Y. Hur, S. W. Jeong, J. Kim, Y. Jang, D. Lee, H. Ji, D. Shin, S. Kim, and H. Lim. You truly understand what I need : Intellectual and friendly dialog agents grounding persona and knowledge. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1053–1066, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.75. URL https://aclanthology.org/2022.findings-emnlp.75/.

[105] B. Liu and S. Mazumder. Lifelong and continual learning dialogue systems: Learning during conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15058–15063, 2021. ISBN 2374-3468.

[106] S. Liu, X. Zhao, B. Li, F. Ren, L. Zhang, and S. Yin. A Three-Stage Learning Framework for Low-Resource Knowledge-Grounded Dialogue Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2272, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for

Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.173. URL https://aclanthology.org/2021.emnlp-main.173.

[107] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang. Towards emotional support dialog systems. 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 3469–3483, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269.

[108] S. Liu, A. Hasan, K. Hong, R. Wang, P. Chang, Z. Mizrachi, J. Lin, D. L. McPherson, W. A. Rogers, and K. D. Campbell. DRAGON: A dialogue-based robot for assistive navigation with visual language grounding. *IEEE Robotics Autom. Lett.*, 9(4):3712–3719, 2024. doi: 10.1109/LRA.2024.3362591. URL https://doi.org/10.1109/lra.2024.3362591.

[109] Z. Liu, M. Patwary, R. Prenger, S. Prabhumoye, W. Ping, M. Shoeybi, and B. Catanzaro. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1317–1337, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.104. URL https://aclanthology.org/2022.findings-acl.104.

[110] H. Lu, W. Lam, H. Cheng, and H. Meng. Partner personas generation for dialogue response generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5200–5212, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.382. URL https://aclanthology.org/2022.naacl-main.382.

[111] S. M. Lukin, F. Gervits, C. J. Hayes, P. Moolchandani, A. Leuski, J. G. Rogers III, C. Sanchez Amaro, M. Marge, C. R. Voss, and D. Traum. ScoutBot: A dialogue system for collaborative navigation. In *Proceedings of ACL 2018, System Demonstrations*, pages 93–98, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4016. URL https://aclanthology.org/P18-4016.

[112] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, X. Chen, and M. Zhou. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *ArXiv*, abs/2002.06353, 2020. URL https://api.semanticscholar.org/CorpusID:211132410.

[113] S.-B. Luo, C.-C. Fan, K.-Y. Chen, Y. Tsao, H.-M. Wang, and K.-Y. Su. Chinese movie dialogue question answering dataset. In Y.-C. Chang and Y.-C. Huang, editors, *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 7–14, Taipei, Taiwan, Nov. 2022. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). URL https://aclanthology.org/2022.rocling-1.2/.

[114] W. Ma, R. Takanobu, and M. Huang. CR-walker: Tree-structured graph reasoning and dialog acts for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1851, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.139. URL https://aclanthology.org/2021.emnlp-main.139.

[115] J. Macina, N. Daheim, S. Chowdhury, T. Sinha, M. Kapur, I. Gurevych, and M. Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.372. URL https://aclanthology.org/2023.findings-emnlp.372/.

[116] B. Madureira. Incrementally enriching the common ground: A research path. In V. Hudecek, P. Schmidtova, T. Dinkar, J. Chiyah-Garcia, and W. Sieinska, editors, *Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems*, pages 57–58, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.yrrsds-1.21/.

[117] B. Madureira and D. Schlangen. It couldn't help but overhear: On the limits of modelling meta-communicative grounding acts with supervised learning. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, and K. Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 149–158, Kyoto, Japan, Sept. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.13. URL https://aclanthology.org/2024.sigdial-1.13/.

[118] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang. Evaluating very long-term conversational memory of LLM agents. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

[119] B. P. Majumder, H. Jhamtani, T. Berg-Kirkpatrick, and J. McAuley. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.739. URL https://aclanthology.org/2020.emnlp-main.739.

[120] M. Markowska, M. Taghizadeh, A. Soubki, S. Mirroshandel, and O. Rambow. Finding common ground: Annotating and predicting common ground in spoken conversations. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.551. URL https://aclanthology.org/2023.findings-emnlp.551/.

[121] M. Mazuecos, F. M. Luque, J. Sánchez, H. Maina, T. Vadora, and L. Benotti. Region under Discussion for visual dialog. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4745–4759, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.390. URL https://aclanthology.org/2021.emnlp-main.390.

[122] J. Miehle, K. Yoshino, L. Pragst, S. Ultes, S. Nakamura, and W. Minker. Cultural communication idiosyncrasies in human-computer interaction. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 74–79, Los Angeles, Sept. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3610. URL https://aclanthology.org/W16-3610.

[123] K. Mitsuda, R. Higashinaka, Y. Oga, and S. Yoshida. Dialogue collection for recording the process of building common ground in a collaborative task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5749–5758, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.618.

[124] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1255. URL https://aclanthology.org/D18-1255.

[125] B. Mohapatra. Conversational grounding in multimodal dialog systems. In V. Hudecek, P. Schmidtova, T. Dinkar, J. Chiyah-Garcia, and W. Sieinska, editors, *Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems*, pages 15–17, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.yrrsds-1.5/.

[126] B. Mohapatra, S. Hassan, L. Romary, and J. Cassell. Conversational grounding: Annotation and analysis of grounding acts and grounding units. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3967–3977, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.352/.

[127] B. Mohapatra, M. N. Kapadnis, L. Romary, and J. Cassell. Evaluating the effectiveness of large language models in establishing conversational grounding. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.545. URL https://aclanthology.org/2024.emnlp-main.545/.

[128] S. Moon, P. Shah, A. Kumar, and R. Subba. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1081. URL https://aclanthology.org/P19-1081.

[129] S. Moon, P. Shah, R. Subba, and A. Kumar. Memory grounded conversational reasoning. 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): System Demonstrations, pages 145–150, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3025. URL https://aclanthology.org/D19-3025.

[130] S. Moon, S. Kottur, P. Crook, A. De, S. Poddar, T. Levin, D. Whitney, D. Difranco, A. Beirami, E. Cho, R. Subba, and A. Geramifard. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.96. URL https://aclanthology.org/2020.coling-main.96.

[131] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1047.

[132] M. Moutti, S. Eleftheriou, P. Koromilas, and T. Giannakopoulos. A dataset for speech emotion recognition in Greek theatrical plays. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1040–1046, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.111.

[133] K. Nakamura, S. Levy, Y.-L. Tuan, W. Chen, and W. Y. Wang. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.41. URL https://aclanthology.org/2022.findings-acl.41.

[134] K. Naszadi, P. Manggala, and C. Monz. Aligning predictive uncertainty with clarification questions in grounded dialog. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14988–14998, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.999. URL https://aclanthology.org/2023.findings-emnlp.999/.

[135] S. Ostermann, S. Zhang, M. Roth, and P. Clark. Commonsense inference in natural language processing (COIN) - shared task report. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 66–74, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6007. URL https://aclanthology.org/D19-6007.

[136] S. Pal, S. Das, and R. K. Srihari. Beyond discrete personas: Personality modeling through journal intensive conversations. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7055–7074, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.470/.

[137] L. Parcalabescu, N. Trost, and A. Frank. What is multimodality? *arXiv preprint arXiv:2103.06304*, 2021.

[138] G. Paré and S. Kitsiou. Chapter 9 methods for literature reviews. handbook of ehealth evaluation: An evidence-based approach [internet]. university of victoria, 2017.

[139] J. Park, M. Joo, J.-K. Kim, and H. J. Kim. Generative subgraph retrieval for knowledge graph–grounded dialog generation. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21167–21182, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1179. URL https://aclanthology.org/2024.emnlp-main.1179/.

[140] R. Pasunuru and M. Bansal. Game-based video-context dialogue. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 125–136, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1012. URL https://aclanthology.org/D18-1012.

[141] D. Petrak, T. T. Tran, and I. Gurevych. Learning from implicit user feedback, emotions and demographic information in task-oriented and document-grounded dialogues. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4573–4603, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.264. URL https://aclanthology.org/2024.findings-emnlp.264/.

[142] C. Pryor, Q. Yuan, J. Liu, M. Kazemi, D. Ramachandran, T. Bedrax-Weiss, and L. Getoor. Using domain knowledge to guide dialog structure induction via neural probabilistic soft logic. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7631–7652, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.422. URL https://aclanthology.org/2023.acl-long.422/.

[143] J. Pustejovsky and N. Krishnaswamy. Situated meaning in multimodal dialogue: Human-robot and human-computer interactions. *Traitement Automatique des Langues*, 61(3):17–41, 2020.

[144] K. Qian and Z. Yu. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1253. URL https://aclanthology.org/P19-1253.

[145] Y. Qin, W. Ji, X. Lan, H. Fei, X. Yang, D. Guo, R. Zimmermann, and L. Liao. Grounding is all you need? dual temporal grounding for video dialog. *ArXiv*, abs/2410.05767, 2024. URL https://api.semanticscholar.org/CorpusID:273228747.

[146] D. Raghu, S. Agarwal, S. Joshi, and Mausam. End-to-end learning of flowchart grounded task-oriented dialogs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4348–4366, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.357. URL https://aclanthology.org/2021.emnlp-main.357.

[147] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534.

[148] A. Razzhigaev, M. Kurkin, E. Goncharova, I. Abdullaeva, A. Lysenko, A. Panchenko, A. Kuznetsov, and D. Dimitrov. OmniDialog: A multimodal benchmark for generalization across text, visual, and audio modalities. In D. Hupkes, V. Dankers, K. Batsuren, A. Kazemnejad, C. Christodoulopoulos, M. Giulianelli, and R. Cotterell, editors, *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 183–195, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.genbench-1.12. URL https://aclanthology.org/2024.genbench-1.12/.

[149] C. Richardson and L. Heck. Commonsense reasoning for conversational AI: A survey of the state of the art. *arXiv preprint arXiv:2302.07926*, 2023.

[150] P. Rodriguez, P. Crook, S. Moon, and Z. Wang. Information seeking in the spirit of learning: A dataset for conversational curiosity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8153–8172, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.655. URL https://aclanthology.org/2020.emnlp-main.655.

[151] F. Ruggeri, M. Mesgar, and I. Gurevych. A dataset of argumentative dialogues on scientific papers. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.425. URL https://aclanthology.org/2023.acl-long.425/.

[152] S. Santhanam, B. Hedayatnia, S. Gella, A. Padmakumar, S. Kim, Y. Liu, and D. Hakkani-Tur. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *CoRR*, abs/2110.05456, 2021. URL https://arxiv.org/abs/2110.05456.

[153] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press, 2019. doi: 10.1609/AAAI.V33I01.33013027. URL https://doi.org/10.1609/aaai.v33i01.33013027.

[154] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035, 2019. ISBN 2374-3468.

[155] K. Schaaff and M.-A. Heidelmann. Impacts of anthropomorphizing large language models in learning environments, 2024. URL https://arxiv.org/abs/2408.03945.

[156] P. Schneider, N. Machner, K. Jokinen, and F. Matthes. Bridging information gaps in dialogues with grounded exchanges using knowledge graphs. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft, and K. Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 110–120, Kyoto, Japan, Sept. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.10. URL https://aclanthology.org/2024.sigdial-1.10/.

[157] O. Shaikh, K. Gligoric, A. Khetan, M. Gerstgrasser, D. Yang, and D. Jurafsky. Grounding gaps in language model generations. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.348. URL https://aclanthology.org/2024.naacl-long.348.

[158] K. Shuster, S. Humeau, A. Bordes, and J. Weston. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.219. URL https://aclanthology.org/2020.acl-main.219.

[159] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. ISBN 2374-3468.

[160] R. Stalnaker. Common ground. *Linguistics and Philosophy*, 25:701–721, 2002. URL https://api.semanticscholar.org/CorpusID:265871097.

[161] C. Strathearn and D. Gkatzia. Task2Dial: A novel task and dataset for commonsense-enhanced task-based dialogue grounded in documents. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 187–196, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.dialdoc-1.21. URL https://aclanthology.org/2022.dialdoc-1.21.

[162] F. Strub, H. De Vries, J. Mary, B. Piot, A. Courville, and O. Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*, 2017.

[163] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SygXPaEYvH.

[164] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472, 2019. doi: 10.1109/ICCV.2019.00756.

[165] H. Sun, Z. Cao, and D. Yang. SPORTSINTERVIEW: A large-scale sports interview benchmark for entity-centric dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5821–5828, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.626.

[166] M. Sung, S. Feng, J. Gung, R. Shu, Y. Zhang, and S. Mansour. Structured list-grounded question answering. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8347–8359, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.558/.

[167] S. K. Suresh, M. Wu, T. Pranav, and E. Chng. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 673–690. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-naacl.40/.

[168] T. Takenobu, I. Ryu, T. Asuka, and K. Naoko. The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of LREC*, pages 422–429, 2012.

[169] C. Tam, R. Brutti, K. Lai, and J. Pustejovsky. Annotating situated actions in dialogue. In J. Bonn and N. Xue, editors, *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 45–51, Nancy, France, June 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.dmr-1.5/.

[170] H. Tan and M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. 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 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

[171] R. Titung and C. O. Alm. FUSE - FrUstration and surprise expressions: A subtle emotional multimodal language corpus. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7544–7555, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.666/.

[172] J. Torres-Foncesca, C. Henry, and C. Kennington. Symbol and communicative grounding through object permanence with a mobile robot. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 124–134, Edinburgh, UK, Sept. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sigdial-1.14.

[173] Y.-L. Tuan, Y.-N. Chen, and H.-y. Lee. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. 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 1855–1865, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1194. URL https://aclanthology.org/D19-1194.

[174] T. Udagawa and A. Aizawa. A natural language corpus of common grounding under continuous and partially-observable context. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7120–7127. AAAI Press, 2019. doi: 10.1609/AAAI.V33I01.33017120. URL https://doi.org/10.1609/aaai.v33i01.33017120.

[175] T. Udagawa and A. Aizawa. Maintaining common ground in dynamic environments. *Transactions of the Association for Computational Linguistics*, 9:995–1011, 2021. doi: 10.1162/tacl_a_00409. URL https://aclanthology.org/2021.tacl-1.59.

[176] T. Udagawa, T. Yamazaki, and A. Aizawa. A linguistic analysis of visually grounded dialogues based on spatial expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 750–765, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.67. URL https://aclanthology.org/2020.findings-emnlp.67.

[177] N. Ueda, H. Habe, A. Yuguchi, S. Kawano, Y. Kawanishi, S. Kurohashi, and K. Yoshino. J-CRe3: A Japanese conversation dataset for real-world reference resolution. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9489–9502, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.829/.

[178] N. Waights Hickman. (implicit) knowledge, reasons, and semantic understanding. *Mind & Language*, 36(5):707–728, 2021. doi: https://doi.org/10.1111/mila.12286. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mila.12286.

[179] H. Wang, M. Hu, Y. Deng, R. Wang, F. Mi, W. Wang, Y. Wang, W.-C. Kwan, I. King, and K.-F. Wong. Large language models as source planner for personalized knowledge-grounded dialogues. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9556–9569, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.641. URL https://aclanthology.org/2023.findings-emnlp.641.

[180] Y. Wang, Z. Zheng, X. Zhao, J. Li, Y. Wang, and D. Zhao. VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5036–5048, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.276. URL https://aclanthology.org/2023.acl-long.276/.

[181] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: An-*

nual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[182] N. Weir, R. Thomas, R. d'Amore, K. Hill, B. Van Durme, and H. Jhamtani. Ontologically faithful generation of non-player character dialogues. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9212–9242, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.520. URL https://aclanthology.org/2024.emnlp-main.520/.

[183] A. Welivita, C.-H. Yeh, and P. Pu. Empathetic response generation for distress support. In S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, and M. Alikhani, editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 632–644, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sigdial-1.59. URL https://aclanthology.org/2023.sigdial-1.59/.

[184] X.-F. Wen, W. Wei, and X.-L. Mao. Sequential topic selection model with latent variable for topic-grounded dialogue. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1209–1219, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.87. URL https://aclanthology.org/2022.findings-emnlp.87.

[185] J. White, G. Poesia, R. Hawkins, D. Sadigh, and N. Goodman. Open-domain clarification question generation without question examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 563–570, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.44. URL https://aclanthology.org/2021.emnlp-main.44.

[186] C. H. Wu, Y. Zheng, X. Mao, and M. Huang. Transferable persona-grounded dialogues via grounded minimal edits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2368–2382, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.183. URL https://aclanthology.org/2021.emnlp-main.183.

[187] C.-S. Wu, A. Madotto, W. Liu, P. Fung, and C. Xiong. QAConv: Question answering on informative conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5389–5411, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.370. URL https://aclanthology.org/2022.acl-long.370.

[188] S. Wu, Y. Li, P. Xue, D. Zhang, and Z. Wu. Section-aware commonsense knowledge-grounded dialogue generation with pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 521–531, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.43.

[189] S. Wu, W. Hsu, and M. L. Lee. EHDChat: A knowledge-grounded, empathy-enhanced language model for healthcare interactions. In J. Hale, K. Chawla, and M. Garg, editors, *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 141–151, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sicon-1.10. URL https://aclanthology.org/2024.sicon-1.10/.

[190] T.-L. Wu, S. Kottur, A. Madotto, M. Azab, P. Rodriguez, B. Damavandi, N. Peng, and S. Moon. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6273–6291, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.345. URL https://aclanthology.org/2023.acl-long.345/.

[191] F. Xia, B. Li, Y. Weng, S. He, K. Liu, B. Sun, S. Li, and J. Zhao. MedConQA: Medical conversational question answering system based on knowledge graphs. In W. Che and E. Shutova, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 148–158, Abu Dhabi, UAE, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.15. URL https://aclanthology.org/2022.emnlp-demos.15/.

[192] T. Xie, C. H. Wu, P. Shi, R. Zhong, T. Scholak, M. Yasunaga, C.-S. Wu, M. Zhong, P. Yin, S. I. Wang, V. Zhong, B. Wang, C. Li, C. Boyle, A. Ni, Z. Yao, D. Radev, C. Xiong, L. Kong, R. Zhang, N. A. Smith, L. Zettlemoyer, and T. Yu. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 602–631, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.39. URL https://aclanthology.org/2022.emnlp-main.39.

[193] H. Xu, S. Moon, H. Liu, B. Liu, P. Shah, B. Liu, and P. Yu. User memory reasoning for conversational recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5288–5308, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.463. URL https://aclanthology.org/2020.coling-main.463.

[194] L. Xu, Q. Zhou, J. Fu, M.-Y. Kan, and S.-K. Ng. CorefDiffs: Co-referential and differential knowledge flow in document grounded conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 471–484, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.38.

[195] B. Xue, W. Wang, H. Wang, F. Mi, R. Wang, Y. Wang, L. Shang, X. Jiang, Q. Liu, and K.-F. Wong. Improving factual consistency for knowledge-grounded dialogue systems via knowledge enhancement and alignment. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7829–7844, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.525. URL https://aclanthology.org/2023.findings-emnlp.525/.

[196] S. Yamashita, K. Inoue, A. Guo, S. Mochizuki, T. Kawahara, and R. Higashinaka. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li, and J. Li, editors, *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861, Hong Kong, China, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.paclic-1.85/.

[197] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. ISBN 2374-3468.

[198] J. Yu, S. Wu, J. Chen, and W. Zhou. LLMs as collaborator: Demands-guided collaborative retrieval-augmented generation for commonsense knowledge-grounded open-domain dialogue systems. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13586–13612, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.794. URL https://aclanthology.org/2024.findings-emnlp.794/.

[199] Y. Yu, A. Eshghi, G. Mills, and O. Lemon. The BURCHAK corpus: a challenge data set for interactive learning of visually grounded word meanings. In *Proceedings of the Sixth Workshop on Vision and Language*, pages 1–10, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2001. URL https://aclanthology.org/W17-2001.

[200] M. Zare, A. Wagner, and R. Passonneau. A POMDP dialogue policy with 3-way grounding and adaptive Sensing for learning through communication. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6767–6780, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.504. URL https://aclanthology.org/2022.findings-emnlp.504/.

[201] S. Zarrieß, J. Hough, C. Kennington, R. Manuvinakurike, D. DeVault, R. Fernández, and D. Schlangen. Pentoref: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131, 2016.

[202] W. Zeng, K. He, Z. Wang, D. Fu, G. Dong, R. Geng, P. Wang, J. Wang, C. Sun, W. Wu, and W. Xu. Semi-supervised knowledge-grounded pre-training for task-oriented dialog systems. In Z. Ou, J. Feng, and J. Li, editors, *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 39–47, Abu Dhabi, Beijing (Hybrid), Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.seretod-1.6. URL https://aclanthology.org/2022.seretod-1.6/.

[203] H. Zhan, S. Maruf, I. Zukerman, and G. Haffari. Going beyond imagination! enhancing multi-modal dialogue agents with synthetic visual descriptions. In T. Kawahara, V. Demberg, S. Ultes, K. Inoue,

S. Mehri, D. Howcroft, and K. Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 420–427, Kyoto, Japan, Sept. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.36. URL https://aclanthology.org/2024.sigdial-1.36/.

[204] H. Zhang, Z. Liu, C. Xiong, and Z. Liu. Grounded conversation generation as guided traverses in commonsense knowledge graphs. *arXiv preprint arXiv:1911.02707*, 2019.

[205] J. Zhang, K. Qian, Z. Liu, S. Heinecke, R. Meng, Y. Liu, Z. Yu, H. Wang, S. Savarese, and C. Xiong. DialogStudio: Towards richest and most diverse unified dataset collection for conversational AI. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2299–2315, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.152/.

[206] K. Zhang, Y. Kang, F. Zhao, and X. Liu. LLM-based medical assistant personalization with short- and long-term memory coordination. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2386–2398, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.132. URL https://aclanthology.org/2024.naacl-long.132/.

[207] R. Zhang and C. Eickhoff. SOCCER: An information-sparse discourse state tracking collection in the sports commentary domain. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4325–4333, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.342. URL https://aclanthology.org/2021.naacl-main.342.

[208] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1205. URL https://aclanthology.org/P18-1205/.

[209] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205.

[210] X. Zhang, R. Divekar, R. Ubale, and Z. Yu. GrounDialog: A dataset for repair and grounding in task-oriented spoken dialogues for language learning. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 300–314, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.26. URL https://aclanthology.org/2023.bea-1.26/.

[211] X. Zhang, R. Divekar, R. Ubale, and Z. Yu. GrounDialog: A Dataset for Repair and Grounding in Task-oriented Spoken Dialogues for Language Learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 300–314, 2023.

[212] Y. Zhang, P. Ren, W. Deng, Z. Chen, and M. Rijke. Improving multi-label malevolence detection in dialogues through multi-faceted label correlation enhancement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3543–3555, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.248. URL https://aclanthology.org/2022.acl-long.248.

[213] C. Zhao, S. Gella, S. Kim, D. Jin, D. Hazarika, A. Papangelis, B. Hedayatnia, M. Namazifar, Y. Liu, and D. Hakkani-Tur. "what do others think?": Task-oriented conversational modeling with subjective knowledge. In S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, and M. Alikhani, editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 309–323, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sigdial-1.28. URL https://aclanthology.org/2023.sigdial-1.28/.

[214] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui. Retrieval-augmented generation for ai-generated content: A survey, 2024. URL https://arxiv.org/abs/2402.19473.

[215] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.272. URL https://aclanthology.org/2020.emnlp-main.272.

[216] X. Zhao, T. Fu, C. Tao, and R. Yan. There is no standard answer: Knowledge-grounded dialogue generation with adversarial activated multi-reference learning. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1878–1891, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.123. URL https://aclanthology.org/2022.emnlp-main.123/.

[217] Y. Zheng, G. Chen, X. Liu, and J. Sun. MMChat: Multi-modal chat dataset on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5778–5786, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.621.

[218] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4623–4629, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-2-7. doi: 10.24963/ijcai.2018/643.

[219] H. Zhou, C. Zheng, K. Huang, M. Huang, and X. Zhu. KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7098–7108, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.635. URL https://aclanthology.org/2020.acl-main.635.

[220] H. Zhou, M. Huang, Y. Liu, W. Chen, and X. Zhu. EARL: Informative knowledge-grounded conversation generation with entity-agnostic representation learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2395, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.184. URL https://aclanthology.org/2021.emnlp-main.184.

[221] P. Zhou, H. Cho, P. Jandaghi, D.-H. Lee, B. Y. Lin, J. Pujara, and X. Ren. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10468, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.714. URL https://aclanthology.org/2022.emnlp-main.714/.

[222] W. Zhu, K. Mo, Y. Zhang, Z. Zhu, X. Peng, and Q. Yang. Flexible end-to-end dialogue system for knowledge grounded conversation. *CoRR*, abs/1709.04264, 2017. URL http://arxiv.org/abs/1709.04264.

A Appendix

A.1 Paper Selection Process

The initial selection of papers is based on the ACL Anthology, the selection was done using a keyword match, i.e. we considered all papers published between 2015 and 2025 whose title or abstract contain both terms “ground(ing)” and “dialogue”. In total, we retrieved and annotated 448 papers with respect to topic (models, datasets, evaluation, theory, supplementary), modality (knowledge, visual, multimodal, persona, other), scope (contextual, domain, commonsense, mixed), and type (static, dynamic, mixed). Note that among the keyword-selected papers there were also some that do not directly relate to the topic of grounding in dialogue and the match could happen e.g. if the paper talks about dialogue processing using ground-truth labels (i.e. term “ground” matches but has a different meaning). Such papers received an additional annotation “irrelevant” and were excluded from the final statistics. We also excluded supplementary papers that are relevant for the grounding in dialogue but focus on very specialized topics (e.g. self-anthropomorphism in robots [103]) or papers introducing dialogue researchers participating in YRRSDS round table [125, 78, 116]. The number of grounding papers decreased from 448 to 384 after filtering out spurious matches and supplementary papers. The initial selection was extended with the papers published in the venues other than the ACL Anthology based on the citations and background knowledge of the authors.

A.2 Topic, Modality, Scope and Type Distribution

Figure 4 shows the topic distribution of the selected papers and indicates a substantial imbalance with the majority of papers (61%) dedicated to the modeling approaches, while 23% of the papers introduce new datasets, and only 10% focus on evaluation. It is important to note that new modeling approaches are typically accompanied with the evaluation results, but the evaluation is usually quite limited and includes only 1-2 baselines and the proposed approach on a few selected datasets. More rigorous benchmarking and model comparison is still missing in the current research on grounding in dialogue.

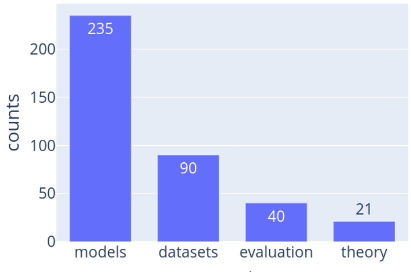


Figure 4. Topic distribution of the papers on grounding in dialogue.

This survey annotates the papers according to the following core modalities: knowledge, visual, multimodal, persona and other. Figure 5 shows that the majority of the papers (60%) is about knowledge grounding, visual grounding is represented at 16%, multimodality is discussed in 12% of the papers, and the rest is almost equally spread among the mixed topics and persona-grounding.

Grounding scope an be characterized as contextual, domain-specific, commonsense or mixed (see Figure 9). It is interesting to see that contextual common ground is the most commonly researched topic, while domain-specific knowledge is also often considered. A significant proportion of papers (27%) has mixed scope and only 5 papers (2%) focus exclusively on the commonsense grounding.

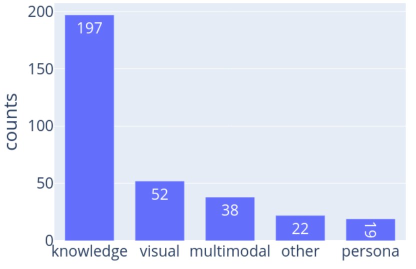


Figure 5. Modality distribution of the papers on grounding in dialogue.

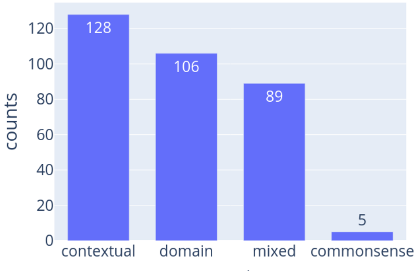


Figure 6. Scope distribution of the papers on grounding in dialogue.

The distribution of grounding types: static vs. dynamic (see Figure 7) makes it clear that static grounding is much better researched than dynamic or mixed cases (46% vs. 30 and 24% correspondingly). This can be likely attributed to the fact that static grounding is easier to model because it often requires an access to a knowledge base or a document collection and the grounding knowledge does not change throughout the conversation. Dynamic grounding requires the dialogue agent to be pro-active, being able to identify ambiguous cases and resolve misunderstanding, e.g. by asking clarification questions.

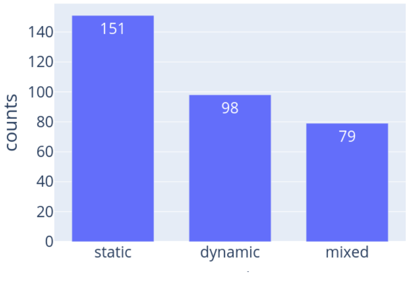


Figure 7. Type distribution of the papers on grounding in dialogue.

A.3 Modality, Scope and Type of Datasets

In this section we summarize the annotation results based on the papers describing the datasets. Figure 8 shows that knowledge modality is much more prevalent than others (51%). Only 16% of the datasets are visual and 17% are multimodal. Also, the grounding scope has unbalanced distribution with 49% datasets related to contextual grounding and 29% domain-specific ones. Interestingly, static and dynamic grounding are almost equally represented in the existing datasets (which is different from the type distribution for modeling approaches as shown in Section A.4)

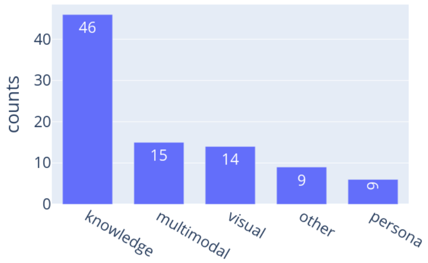


Figure 8. Modality distribution of the *datasets* for grounding in dialogue.

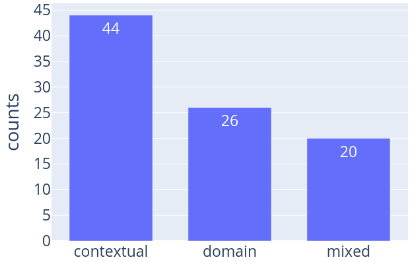


Figure 9. Scope distribution of the *datasets* for grounding in dialogue.

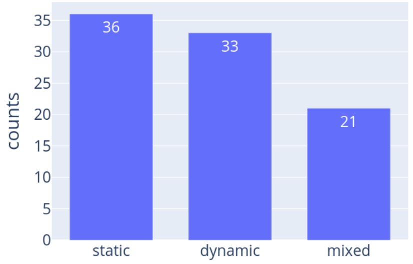


Figure 10. Type distribution of the *datasets* for grounding in dialogue.

A.4 Modality, Scope and Type of Models

If we consider only those papers that describe modeling approaches and plot their distribution per modality, we can see in Figure 11 that knowledge-based approaches are the most common ones. Multimodal and visual grounding receive less attention (11 and 17% each) and the least researched modality is persona-based grounding with only 6% of all papers addressing this topic. These statistics emphasize that there is a lack of research on grounding in the modalities that go beyond knowledge, especially when multiple modalities and persona-related features should be taken into consideration.

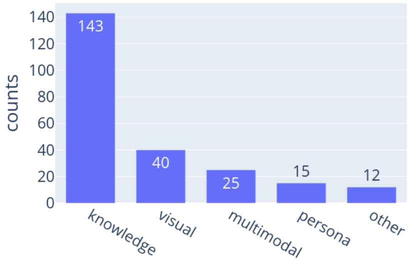


Figure 11. Modality distribution of the *models* for grounding in dialogue.

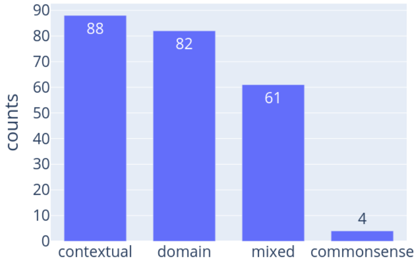


Figure 12. Scope distribution of the *models* for grounding in dialogue.

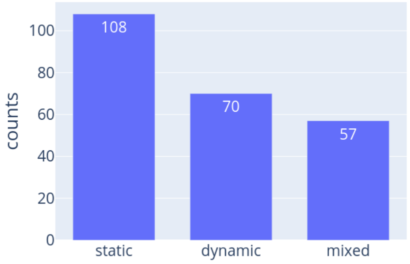


Figure 13. Type distribution of the *models* for grounding in dialogue.

A.5 Grounding Datasets

In the scope of this survey, we compiled a list of the datasets for grounding in dialogue categorized according to the modality, type, and scope (see Table 2-7). The datasets include several additional resources that were not published through the ACL Anthology. This information along with our paper annotations is available to the research community in the following GitHub repository: https://github.com/tanikina/common-ground-in-dialogue.

Table 2. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
SAGA22 [17]	The SAGA22 dataset is based on 148 transcribed videos, it is a dataset of teacher and student talk moves and annotated math tutoring sessions. Talk moves use dialogue acts grounded in Accountable Talk theory.	Knowl.	Dyn.	Cont.	upon request/unknown
Reannotated Spot the Difference and Meetup Datasets [126]	The Meetup and Spot the Difference datasets were (re-)annotated with Grounding Acts, Common Grounding units, and degrees of grounding.	Knowl.	Dyn.	Cont.	https://osf.io/qfcnm/?view_only=34e7259fe8fc4ade82d55ba7d5105ffe
The Common Ground Corpus [120]	The Common Ground Corpus is annotated on the top of the LDC CALLHOME American Speech corpus, which consists of collections of 120 unscripted dialogs between close friends or family members. The dialogs are available in both written form and audio. The Common Ground corpus is the first attempt at annotating common ground in a discourse, providing the annotations for beliefs and common ground updates.	Knowl.	Dyn.	Cont.	https://github.com/cogstates/2023-emnlp-common-ground
GrounDialog [210]	An annotated dataset of spoken conversations with repair and grounding patterns. The dataset contains 42 dialogues with 1569 turns.	Knowl.	Dyn.	Cont.	upon request/unknown
CoomonLayout [123]	The dataset is built for the CommonLayout task in which two workers lay out the same figure set into a common design through text chat. To perform the task, they discuss the idea of a final layout and move figures into the same position one by one. The dataset contains 984 dialogues and each dialogue has 28.8 utterances on average.	Knowl.	Dyn.	Cont.	upon request/unknown
Reflect [221]	Reflect is a dataset that annotates dialogues with explicit common ground (represented as inferences approximating shared knowledge and beliefs) and contains 9K diverse human-generated responses each following one common ground.	Knowl.	Dyn.	Cont.	https://inklab.usc.edu/Reflect/
SPOLIN [29]	Selected Pairs Of Learnable ImprovisatioN (SPOLIN) corpus is a collection of more than 26K English dialogue turn pairs, each consisting of a prompt and subsequent grounded response, where responses are not only coherent with dialogue context but also initiate the next relevant contribution.	Knowl.	Dyn.	Cont.	https://justin-cho.com/spolin
KNUDGE [182]	KNUDGE (KNowledge Constrained User-NPC Dialogue GEneration) is constructed from side quest dialogues drawn directly from game data of Obsidian Entertainment's The Outer Worlds, leading to real-world complexities in generation: (1) utterances must remain faithful to the game lore, including character personas and backstories; (2) a dialogue must accurately reveal new quest details to the human player; and (3) dialogues are large trees as opposed to linear chains of utterances. KNUDGE contains 159 dialogue trees.	Knowl.	Mix	Cont.	https://github.com/nweir127/KNUDGE
KETOD [27]	KETOD (Knowledge-Enriched Task-Oriented Dialogue) enriches task-oriented dialogues with chit-chat based on relevant entity knowledge. It contains >5K dialogues.	Knowl.	Mix	Cont.	https://github.com/facebookresearch/ketod
ChattyChef [94]	ChattyChef is a dataset of cooking dialogues, designed to support research on instruction-grounded conversational agents. ChattyChef contains 267 dialogues with 26 utterances per dialogue.	Knowl.	Dyn.	Dom.	https://github.com/octaviaguo/ChattyChef
EHD [189]	Empathetic Healthcare Dialogue (EHD) dataset can help with generating human-like empathetic responses within the healthcare domain. It contains a wide range of synthetic, multi-turn dialogues between doctors and patients that are not only emotionally supportive, but also clinically informative. EHD contains 33K dialogues, with an average of 12 utterances per dialogue.	Knowl.	Mix	Dom.	https://huggingface.co/datasets/ericw955/EHD
MathDial [115]	MathDial is a dataset of 3K one-to-one teacher-student tutoring dialogues grounded in multi-step math reasoning problems.	Knowl.	Mix	Dom.	https://github.com/eth-nlped/mathdial
ArgSciChat [151]	ArgSciChat is a dataset of 41 argumentative dialogues between scientists on 20 NLP papers. The dataset includes both exploratory and argumentative questions and answers in a dialogue discourse on a scientific paper.	Knowl.	Mix	Dom.	https://github.com/UKPLab/acl2023-argscichat
KdConv [219]	KdConv, a Chinese multi-domain dataset towards multi-turn Knowledge-driven Conversation with 86K utterances and 4.5K dialogues in three domains.	Knowl.	Mix	Dom.	https://github.com/thu-coai/KdConv
List2QA [166]	List2QA dataset is designed to evaluate the ability of QA systems to respond effectively using list information. The dataset is created from unlabeled customer service documents with language models and model-based filtering, it has >2K utterances.	Knowl.	Stat.	Dom.	upon request/unknown
MISeD – Meeting Information Seeking Dialogs dataset [59]	MISeD – Meeting Information Seeking Dialogs dataset is a dataset of information-seeking dialogues focusing on meeting transcripts for 225 meetings, comprising 432 dialogues, and 4161 query-response pairs.	Knowl.	Stat.	Dom.	https://github.com/google-research-datasets/MISeD
Verify-then-Generate [35]	1K student solutions and their stepwise reasoning chains in the domain of multi-step math problem-solving.	Knowl.	Stat.	Dom.	https://github.com/eth-lre/verify-then-generate
NewsDialogues [100]	A human-to-human Chinese dialogue dataset with 1K conversations with a total of 14.6K utterances and detailed annotations for target topics and knowledge spans.	Knowl.	Stat.	Dom.	https://github.com/SihengLi99/NewsDialogues

Table 3. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
CMDQA [113]	Chinese dialogue-based information-seeking question answering dataset CMDQA, which is mainly applied to the scenario of getting Chinese movie related information. It contains 10K QA dialogs (40K turns in total).	Knowl.	Stat.	Dom.	upon request/unknown
SPORTSINTERVIEW [165]	Dataset in the domain of sports interview, it contains two types of external knowledge sources as knowledge grounding, 150K interview sessions and 34K distinct interviewees.	Knowl.	Stat.	Dom.	upon request/unknown
Doc2Bot [49]	Dataset with over 100K turns based on Chinese documents from five domains.	Knowl.	Stat.	Dom.	https://github.com/Doc2Bot/Doc2Bot
MultiRefKGC [216]	A multi-reference Knowledge-Grounded Conversation (KGC) dataset based on conversations from Reddit with 130K dialogues.	Knowl.	Stat.	Dom.	https://github.com/TingchenFu/MultiRefKGC
CM-CQA [191]	A large-scale Chinese Medical CQA (CM-CQA) dataset based on 45 medical subdomains, 33615 entities, 8808 symptoms, 1294753 dialogues.	Knowl.	Stat.	Dom.	https://github.com/WENGSYX/LingYi
Social-Dialogues-Coreference [91]	Dataset for resolving third-person references in social dialogues (inner and outer-circle references), based on the episodes of the Friends series. It contains social dialogue and long-term connections between mentions that go beyound a single document.	Knowl.	Stat.	Dom.	https://github.com/cltl/inner-outer-coreference
MultiDoc2Dial [47]	Conversations grounded in 488 documents, 4796 dialogues in total.	Knowl.	Stat.	Dom.	https://doc2dial.github.io/multidoc2dial/
TicketTalk [16]	A movie ticketing dialog dataset with 23,789 annotated conversations that range from completely open-ended and unrestricted to more structured in terms of the knowledge base, discourse features, and number of turns.	Knowl.	Stat.	Dom.	https://git.io/JL8an
Doc2Dial [46]	The dataset of goal-oriented dialogues that are grounded in the documents. 4500 annotated conversations grounded in over 450 documents from four domains.	Knowl.	Stat.	Dom.	http://doc2dial.github.io/
Background-aware movie dataset [124]	Background-aware conversation dataset about movies with 90K utterances from 9K conversations grounded in plots, reviewes, comments.	Knowl.	Stat.	Dom.	https://github.com/nikitacs16/Holl-E
Multi-turn and multi-domain dataset [45]	The dataset of 3031 dialogues that are grounded through knowledge bases and span three distinct tasks in the in-car personal assistant space: calendar scheduling, weather information retrieval, and point-of-interest navigation.	Knowl.	Stat.	Dom.	http://nlp.stanford.edu/projects/kvret/kvret_dataset_public.zip
SOCCER [207]	2263 soccer matches including with time-stamped natural language commentary accompanied by discrete events such as a team scoring goals, switching players or being penalized with cards.	Knowl.	Dyn.	Mix	https://github.com/bcbi-edu/p_eickhoff_SOCCER
FloDial [146]	FloDial has 2738 dialogs grounded on 12 different troubleshooting flowcharts.	Knowl.	Dyn.	Mix	https://dair-iitd.github.io/FloDial
OpenDialKG [128]	Open-ended Dialog and KG parallel corpus called OpenDialKG, where each utterance from 15K human-to-human role-playing dialogs is manually annotated with ground-truth reference to corresponding entities and paths from a large-scale KG with 1M+ facts.	Knowl.	Dyn.	Mix	https://github.com/facebookresearch/opendialkg
FEDI [141]	FEDI, the first English task-oriented and document-grounded dialogue dataset annotated with implicit user feedback, emotions and demographic information.	Knowl.	Mix	Mix	https://github.com/UKPLab/FEDI
Situated Actions in Dialogue [169]	Action and Abstract Meaning Representation annotations for first-person point-of-view videos (based on the Fibonacci Weights Task dataset and Epic Kitchens dataset).	Knowl.	Mix	Mix	upon request/unknown
Japanese Move Recommendations with external and speaker-derived grounding [85]	Annotated knowledge-grounded dialogue corpus Japanese Movie Recommendation Dialogue that contains >5K dialogues. Each entity is annotated with its information source, either derived from external knowledge (database-derived) or the speaker's own knowledge, experiences, and opinions (speaker-derived).	Knowl.	Mix	Mix	upon request/unknown
Task2Dial [161]	A dataset of document-grounded task-based dialogues, where an Information Giver (IG) provides instructions (by consulting a document) to an Information Follower (IF). The dataset contains dialogues with an average 18.15 number of turns grounded in 353 documents.	Knowl.	Mix	Mix	http://www.huggingface.co/datasets/cstrathe435/Task2Dial
QAConv [187]	A question-answering (QA) dataset that uses conversations as a knowledge source and offers 34608 QA pairs with both human-written and machine-generated questions.	Knowl.	Mix	Mix	https://github.com/salesforce/QAConv
A Dataset for Conversational Curiosity [150]	14K dialogues (181K utterances) where users and assistants converse about geographic topics like geopolitical entities and locations. This dataset is annotated with pre-existing user knowledge, message-level dialog acts, grounding to Wikipedia, and user reactions to messages.	Knowl.	Mix	Mix	http://curiosity.pedro.ai/
BridgeKG [156]	Annotated human conversations across five knowledge domains, 26 information-seeking conversations and 669 dialogue turns.	Knowl.	Stat.	Mix	https://github.com/philotron/Bridge-KG
DialogStudio [205]	Collection with diverse data from open-domain dialogues, task-oriented dialogues, natural language understanding, conversational recommendation, dialogue summarization, and knowledge-grounded dialogues.	Knowl.	Stat.	Mix	https://github.com/salesforce/DialogStudio

Table 4. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
SK-TOD [213]	Subjective-Knowledge Task-Oriented Dialogue (SK-TOD) dataset contains subjective knowledge-seeking dialogue contexts and manually annotated responses grounded in subjective knowledge sources. SK-TOD has >9K instances consisting of subjective user requests and subjective knowledge-grounded responses.	Knowl.	Stat.	Mix	https://github.com/alexa/dstc11-track5
HPD [25]	Harry Potter Dialogue (HPD) dataset in English and Chinese is annotated with vital background information, including dialogue scenes, speakers, character relationships, and attributes. It has >1K dialogues.	Knowl.	Stat.	Mix	https://nuochenpku.github.io/HPD.github.io
RSD [63]	Response Selection Data (RSD) dataset where responses from multiple response generators produced for the same dialog context are manually annotated as appropriate (positive) and inappropriate (negative). The data has 100K interactiona and 2.5 million turns.	Knowl.	Stat.	Mix	upon request/unknown
COMET [90]	A new task-oriented dialog dataset COMET, which contains 11.5K user-assistant dialogs (totalling 103K utterances), grounded in simulated personal memory graphs.	Knowl.	Stat.	Mix	https://github.com/facebookresearch/comet_memory_dialog
Augmented MultiWOZ 2.1 [83]	An augmented version of MultiWOZ 2.1, which includes new out-of-API-coverage turns and responses grounded on external knowledge sources. The dataset contains >10K dialogues with >9K augmented turns.	Knowl.	Stat.	Mix	upon request/unknown
MGConvRex [193]	A new Memory Graph (MG) - Conversational Recommendation parallel corpus called MGConvRex with 7K+ human-to-human role-playing dialogs, grounded on a large-scale user memory bootstrapped from real-world user scenarios.	Knowl.	Stat.	Mix	upon request/unknown
Annotated Weights Task Dataset [77]	A dataset of multimodal interactions in a shared physical space with speech transcriptions, prosodic features, gestures, actions, and facets of collaboration (based on the Weights Task).	Multi	Dyn.	Cont.	https://github.com/csu-signal/Common-Ground-detection
J-CRe3 [177]	A Japanese Conversation dataset for Real-world Reference Resolution (J-CRe3) that contains video and dialogue audio of real-world conversations between two people acting as a master and an assistant robot at home. The dataset is annotated with crossmodal tags between phrases in the utterances and the object bounding boxes in the video frames. These tags include indirect reference relations, such as predicate-argument structures and bridging references as well as direct reference relations.	Multi	Dyn.	Cont.	https://github.com/riken-grp/J-CRe3
LoCoMo [118]	LoCoMo, a dataset of very long-term conversations, each encompassing 600 turns and 16K tokens on avg., over up to 32 sessions. The dialogues are grounded on personas and temporal event graphs.	Multi	Dyn.	Cont.	https://snap-research.github.io/locomo
Chinese Whispers [87]	The corpus with 34 interactions, where each subject first assembles and then instructs how to assemble IKEA furniture. The dataset has speech, eye-gaze, pointing gestures, and object movements, as well as subjective interpretations of mutual understanding, collaboration and task recall.	Multi	Dyn.	Cont.	https://www.kth.se/profile/diko/page/material
Spatial AMR and Grounded Minecraft Dataset [14]	A multimodal corpus consisting of 170 3D structure-building dialogues between a human architect and human builder in Minecraft. The data contain sentence-level and document-level annotations designed to capture implicit information, the coordinates and the spatial framework annotation ground the spatial language in the dialogues.	Multi	Dyn.	Cont.	https://github.com/cu-clear/Spatial-AMR/
OneCommon [176]	OneCommon Corpus for visual conversational grounding with 600 dialogues annotated with spatial expressions that capture predicate-argument structure, modification and ellipsis.	Multi	Dyn.	Cont.	https://github.com/Alab-NII/onecommon
The Niki and Julie Corpus [7]	The Niki and Julie corpus contains more than 600 dialogues between human participants and a human-controlled robot or virtual agent, engaged in a series of collaborative item-ranking tasks designed to measure influence. Some of the dialogues contain deliberate conversational errors by the robot, designed to simulate the kinds of conversational breakdown that are typical of present-day automated agents. Data collected include audio and video recordings, the results of the ranking tasks, and questionnaire responses; some of the recordings have been transcribed and annotated for verbal and nonverbal feedback.	Multi	Dyn.	Cont.	upon request/unknown
REX Corpora [168]	A collection of multimodal corpora of referring expressions, the REX corpora. The corpora include time-aligned extra-linguistic information such as participant actions and eye-gaze on top of linguistic information, also the dialogues were collected with various configurations in terms of the puzzle type, hinting and language. The REX corpora contain 226 dialogues.	Multi	Dyn.	Cont.	upon request/unknown
GreThE [132]	GreThE, the Greek Theatrical Emotion dataset, a publicly available data collection for speech emotion recognition in Greek theatrical plays. The dataset contains 500 utterances that have been annotated in terms of their emotional content (valence and arousal).	Multi	Mix	Cont.	https://github.com/magcil/GreThE
Memory Dialog [129]	A corpus of memory grounded conversations, which comprises human-to-human role-playing dialogues given synthetic memory graphs with simulated attributes and connections to real entities (e.g. locations, events, public entities).	Multi	Mix	Cont.	upon request/unknown

Table 5. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
FUSE [171]	FrUstration and Surprise Expressions (FUSE) is a multimodal corpus for expressive task-based spoken language and dialogue, focusing on language use under frustration and surprise.	Multi	Stat.	Cont.	https://fusecorpus.github.io/FUSE/
MPCHAT [2]	MPCHAT is the first multimodal persona-based dialogue dataset which extends persona with both text and images to contain episodic memories. It contains 15K dialogues sourced from Reddit.	Multi	Stat.	Cont.	https://github.com/ahnjaewoo/MPCHAT
NICE [24]	Neural Image Commenting with Empathy (NICE) dataset consists of almost two million images and the corresponding human-generated comments with a set of human annotations. The dataset can be used to generate dialogues grounded in a user-shared image with increased emotion and empathy while minimizing offensive outputs.	Multi	Stat.	Cont.	https://nicedataset.github.io/
SIMMC 2.0 [89]	A dataset for Situated and Interactive Multimodal Conversations, SIMMC 2.0, which includes 11K task-oriented user-assistant dialogs (117K utterances) in the shopping domain, grounded in immersive and photo-realistic scenes.	Multi	Mix	Dom.	https://github.com/facebookresearch/simmc2
HybriDialogue [133]	A dialogue dataset, HybriDialogue, which consists of crowdsourced natural conversations grounded on both Wikipedia text and tables. The conversations are created through the decomposition of complex multihop questions into simple, realistic multiturn dialogue interactions.	Multi	Stat.	Dom.	https://github.com/entitize/HybridDialogue
KOMODIS [50]	Knowledgable and Opinionated MOvie DIScussions (KOMODIS) is a labeled dialogue dataset in the domain of movie discussions, where every dialogue is based on pre-specified facts and opinions. It contains >7K dialogues and >103K utterances.	Multi	Stat.	Dom.	https://github.com/fabiangal/komodis-dataset
SIMMC [130]	Situated Interactive MultiModal Conversations (SIMMC) is a dataset with 13K human-human dialogs (169K utterances) collected using a multimodal Wizard-of-Oz (WoZ) setup, on two shopping domains: (a) furniture – grounded in a shared virtual environment; and (b) fashion – grounded in an evolving set of images. Data include multimodal context of the items appearing in each scene, and contextual NLU, NLG and coreference annotations.	Multi	Dyn.	Mix	https://github.com/facebookresearch/simmc
RED [183]	Reddit Emotional Distress (RED) is a large-scale dialogue dataset that contains 1.3M peer support dialogues spanning across more than 4K distress-related topics.	Other	Dyn.	Cont.	https://github.com/yehchunhung/EPIMEED
MDMD [212]	A multi-label dialogue malevolence detection (MDMD) dataset where a dialogue response is considered malevolent if it is grounded in negative emotions, inappropriate behavior, or an unethical value basis in terms of content and dialogue acts. MDMD contains >8K utterances.	Other	Dyn.	Cont.	https://github.com/repozhang/malevolent_dialogue
Dynamic OneCommon [175]	A large-scale dataset of 5617 dialogues to enable fine-grained evaluation, using complex spatio-temporal expressions to create and maintain common ground over time in dynamic environments.	Other	Dyn.	Cont.	https://github.com/Alab-NII/dynamic-onecommon
HuRDL [55]	The Human-Robot Dialogue Learning (HuRDL) corpus is a dialogue corpus with 22 dialogues and 1122 turns collected in an online interactive virtual environment in which human participants play the role of a robot performing a collaborative tool-organization task. The data can be used to improve question generation in situated intelligent agents.	Other	Dyn.	Cont.	https://github.com/USArmyResearchLab/ARL-HuRDL
ESConv [107]	Emotion Support Conversation dataset (ESConv) with rich annotation (especially support strategy) in a help-seeker and supporter mode for 1K dialogues.	Other	Dyn.	Cont.	https://github.com/thu-coai/Emotional-Support-Conversation
CreST [54]	A corpus of spontaneous, task-oriented dialogue (CReST corpus), which was annotated for disfluencies and conversational moves that can facilitate grounding and coordination.	Other	Dyn.	Cont.	upon request/unknown
EmpatheticDialogues [147]	EmpatheticDialogues is a dataset of 25K conversations grounded in emotional situations, the data were gathered from 810 different participants.	Other	Mix	Cont.	https://parl.ai/
ProsocialDialog [80]	The ProsocialDialog dataset consists of 58K dialogues, with 331K utterances, and 497K dialogue safety labels accompanied by free-form rationales. It can be used for generating more socially acceptable dialogues grounded in social norms.	Other	Stat.	Dom.	https://hyunw.kim/prosocial-dialog
BSBT [82]	Blended Skill BotsTalk (BSBT), a large-scale multi-skill dialogue dataset comprising 300K conversations where agents are grounded to the specific target skills.	Other	Stat.	Dom.	https://github.com/convei-lab/BotsTalk
JIC [136]	Journal Intensive Conversations (JIC) is a journal-based conversational dataset with around 400,000 dialogues and a framework for generating personalized conversations using long-form journal entries from Reddit. The data capture common personality traits — openness, conscientiousness, extraversion, agreeableness, and neuroticism — ensuring that dialogues authentically reflect an individual's personality.	Persona	Mix	Cont.	https://github.com/Sayantan-world/Beyond-Discrete-Personas
KBP [179]	A personalized knowledge-grounded dialogue dataset Knowledge Behind Persona (KBP) is the first to consider the dependency between persona and implicit knowledge. It comes with >2K dialogues grounded in persona and knowledge.	Persona	Stat.	Cont.	https://github.com/ruleGreen/SAFARI

Table 6. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
LiveChat [52]	The LiveChat dataset is composed of 1.33 million real-life Chinese dialogues with almost 3800 average sessions across 351 personas and fine-grained profiles for each persona representing multi-party conversations.	Persona	Stat.	Cont.	https://github.com/gaojingsheng/LiveChat
PersonaMinEdit [186]	The PersonaMinEdit dataset is derived from PersonaChat with multiple human references for the edited response, it can be used to evaluate persona-grounded minimal editing.	Persona	Stat.	Cont.	https://github.com/thu-coai/grounded-minimal-edit
MaLP [206]	The dataset contains 11K dialogues, it is based on an open-source medical corpus and can help with building personalized medical assistants. The dataset is focusing on medical scenarios, including domain and commonsense information as well as personal details (e.g., chronic diseases, dialogue preferences).	Persona	Mix	Dom.	https://github.com/MatthewKKai/MaLP
PeaCoK [53]	A large-scale persona commonsense knowledge graph, PeaCoK, contains 100K human-validated persona facts. It formalizes five common aspects of persona knowledge: characteristics, routines and habits, goals and plans, experiences, and relationships.	Persona	Stat.	Mix	https://github.com/Silin159/PeaCoK
Persona-Chat [209]	Persona-Chat is a crowd-sourced dataset, collected via Amazon Mechanical Turk, where each of the pair of speakers condition their dialogue on a given profile, which is provided. The dataset is based on 1155 possible personas and provides 11K dialogues.	Persona	Stat.	Mix	https://github.com/facebookresearch/ParlAI/tree/master/projects/personachat
RealPersonaChat [196]	RealPersonaChat (RPC) corpus is based on collecting the actual personality traits and personas of interlocutors and having them freely engage in dialogue. This corpus contains 14K dialogues in Japanese and represents one of the largest corpora of dialogue data annotated with personas and personality traits.	Persona	Stat.	Mix	https://github.com/nu-dialogue/real-persona-chat
VSTAR [180]	Video-grounded Scene &Topic AwaRe dialogue (VSTAR) dataset is a large scale video-grounded dialogue understanding dataset based on 395 TV series. It contains annotations for scene and topic transitions. VSTAR contains 185K dialogues.	Visual	Dyn.	Cont.	https://vstar-benchmark.github.io/
SIMMC-2.0 [190]	SIMMC-2.0 is a video-grounded task-oriented dialog dataset that captures real-world AI-assisted user scenarios in virtual reality. It contains fine-grained and scene-grounded annotations for 4K dialogues.	Visual	Dyn.	Cont.	https://github.com/patrick-tssn/VSTAR
DVD [97]	A Diagnostic Dataset for Video-grounded Dialogue (DVD) was designed to contain minimal biases and has detailed annotations for the different types of reasoning over the spatio-temporal space of video. Dialogues were synthesized over multiple question turns, each of which was injected with a set of cross-turn semantic relationships. DVD was built from 11K CATER synthetic videos and contains 10 instances of 10-round dialogues for each video, resulting in more than 100K dialogues and 1M question-answer pairs.	Visual	Dyn.	Cont.	https://github.com/facebookresearch/DVDialogues
VFD [73]	A visually-grounded first-person dialogue (VFD) dataset with verbal and non-verbal responses. The VFD dataset provides manually annotated (1) first-person images of agents, (2) utterances of human speakers, (3) eye-gaze locations of the speakers, and (4) the agents' verbal and non-verbal responses. For the verbal response selection task, VFD dataset has almost 600K dialogues. For the non-verbal response selection task it contains around 160K dialogues.	Visual	Dyn.	Cont.	https://randd.yahoo.co.jp/en/softwaredata
PhotoBook [61]	The dataset was collected through a collaborative game prompting two online participants to refer to images utilising both their visual context as well as previously established referring expressions. This resulted in 2500 annotated dialogues.	Visual	Dyn.	Cont.	https://dmg-photobook.github.io/
CoDraw [81]	This dataset is based on a Collaborative image-Drawing game between two agents, called CoDraw. The game is grounded in a virtual world that contains movable clip art objects and involves two players: a Teller and a Drawer. The Teller sees an abstract scene containing multiple clip art pieces in a semantically meaningful configuration, while the Drawer tries to reconstruct the scene on an empty canvas using available clip art pieces. The two players communicate with each other using natural language. The CoDraw dataset contains 10K dialogs with 138K messages exchanged between human players.	Visual	Dyn.	Cont.	https://github.com/facebookresearch/CoDraw
CLEVR-Dialog [88]	CLEVR-Dialog is a large diagnostic dataset for studying multi-round reasoning in visual dialog. The dialog grammar is grounded in the scene graphs of the images from the CLEVR dataset. This combination results in a dataset where all aspects of the visual dialog are fully annotated. In total, CLEVR-Dialog contains 5 instances of 10-round dialogs for about 85K CLEVR images, totaling to 4.25M question-answer pairs.	Visual	Dyn.	Cont.	https://github.com/satwikkottur/clevr-dialog
Twitch-FIFA [140]	The Twitch-FIFA dataset is a video-context, many-speaker dialogue dataset based on live-broadcast soccer game videos and chats from Twitch.tv. It is based on 49 FIFA-18 game videos along with their users' chat. The dataset provides the triples with video context, chat context, and response data.	Visual	Dyn.	Cont.	https://github.com/ramakanth-pasunuru/video-dialogue

Table 7. Datasets. Abbreviations: Knowl. (Knowledge), Multi (Multimodal), Stat. (Static), Dyn. (Dynamic), Cont. (Contextual) and Dom. (Domain).

Dataset	Description	Modality	Type	Scope	Data URL
GuessWhat?! [38]	The goal of the GuessWhat?! game is to locate an unknown object in a rich image scene by asking a sequence of questions. Higher-level image understanding, like spatial reasoning and language grounding, is required to solve the task. The dataset consists of 150K human-played games with a total of 800K visual question-answer pairs on 66K images.	Visual	Dyn.	Cont.	https://guesswhat.ai/download
MeetUp! [68]	MeetUp! is a two-player coordination game where players move in a visual environment, with the objective of finding each other. To do so, they must talk about what they see, and achieve mutual understanding. The collected data includes 5695 annotated turns.	Visual	Dyn.	Cont.	https://github.com/clp-research/meetup
Visually Grounded Follow-up Questions [44]	A dataset of questions that require grounding both on the visual input and the dialogue history. The dataset is based on GuessWhat?! And focuses on the follow-up questions that require multimodal grounding, such questions can be extracted by identifying patterns of trigger-zoomer questions where trigger restricts the context and zoomers are spatial questions that requires triggers to be answered first.	Visual	Dyn.	Cont.	https://github.com/tianaidong/2021SpLU-RoboNLP-VISPA
PentoRef [201]	PentoRef is a corpus of task-oriented dialogues collected in systematically manipulated settings. The corpus is multilingual, with English and German sections, and overall comprises more than 20K utterances. The dialogues are fully transcribed and annotated with referring expressions mapped to objects in corresponding visual scenes, which makes the corpus a rich resource for research on spoken referring expressions in generation and resolution. The corpus includes several sub-corpora that correspond to different dialogue situations where parameters related to interactivity, visual access, and verbal channel have been manipulated in systematic ways.	Visual	Dyn.	Cont.	https://github.com/clp-research/pentoref
Image-Chat [158]	Image-Chat consists of 202K dialogues over 202K images using 215 possible style traits. It is a dataset of grounded human-human conversations, where speakers are asked to play roles given a provided emotional mood or style, as the use of such traits is also a key factor in engagingness	Visual	Mix	Cont.	http://parl.ai/projects/image_chat
VisdialConv [1]	VisdialConv is a subset of the VisDial validation set consisting of 97 dialogs, where the crowd-workers identified single turns (with dense annotations) requiring historical information. The crowd-workers were asked whether they could provide an answer to a question given an image, without showing them the dialog history.	Visual	Mix	Cont.	https://github.com/shubhamagarwal92/visdialconv-amt
IGC [131]	Image Grounded Conversations (IGC) is a dataset in which natural-sounding conversations are generated about a shared image. This is a multiple reference dataset of crowd-sourced, event-centric conversations on images, where visual grounding constrains the topic of conversation. It contains >4K conversations.	Visual	Mix	Cont.	https://www.microsoft.com/en-us/download/details.aspx?id=55324&751be11f-ede8
MMChat [217]	MMChat is a large scale Chinese multi-modal dialogue corpus (32.4M raw dialogues and 120.84K filtered dialogues). MMChat contains image-grounded dialogues collected from real conversations on social media.	Visual	Stat.	Cont.	https://github.com/silverriver/MMChat
Region-under-Discussion for Visual Dialog [121]	A subset of the Guesswhat?! questions for which their dialog history completely changes the responses. Natural language understanding grounded in vision.	Visual	Stat.	Cont.	https://github.com/mmazuecos/Region-under-discussion-for-visual-dialog
BURCHAK [199]	A human-human dialogue dataset for interactive learning of visually grounded word meanings through ostensive definition by a tutor to a learner. The dataset contains 177 dialogues (each about one visual object) with a total of 2454 turns.	Visual	Dyn.	Mix	https://service.tib.eu/ldmservice/dataset/burchak-corpus

Cultural and Ethical Perspectives

Do Large Language Models Understand Morality Across Cultures?

Hadi Mohammadi [a,*], **Yasmeen F.S.S. Meijer** [a], **Efthymia Papadopoulou** [a] **and Ayoub Bagheri**[a]

[a]Department of Methodology and Statistics, Utrecht University, The Netherlands

Abstract. Recent advancements in large language models (LLMs) have established them as powerful tools across numerous domains. However, persistent concerns about embedded biases, such as gender, racial, and cultural biases arising from their training data, raise significant questions about the ethical use and societal consequences of these technologies. This study investigates the extent to which LLMs capture cross-cultural differences and similarities in moral perspectives. Specifically, we examine whether LLM outputs align with patterns observed in international survey data on moral attitudes. To this end, we employ three complementary methods: (1) comparing variances in moral scores produced by models versus those reported in surveys, (2) conducting cluster alignment analyses to assess correspondence between country groupings derived from LLM outputs and survey data, and (3) directly probing models with comparative prompts using systematically chosen token pairs. Our results reveal that current LLMs often fail to reproduce the full spectrum of cross-cultural moral variation, tending to compress differences and exhibit low alignment with empirical survey patterns. These findings highlight a pressing need for more robust approaches to mitigate biases and improve cultural representativeness in LLMs. We conclude by discussing the implications for the responsible development and global deployment of LLMs, emphasizing fairness and ethical alignment.

1 Introduction

Large language models (LLMs) have recently taken center stage in both scientific and public debates due to significant advancements in their performance [2]. These models now show great promise for applications ranging from search engines and recommendation systems to automated decision-making tools that deeply influence everyday life. Nonetheless, alongside these impressive capabilities, concerns persist regarding the potential biases LLMs can exhibit, such as gender, racial, and cultural bias.

A primary reason for this risk is that LLMs learn from vast, real-world text datasets that may contain societal and cultural prejudices [11, 16]. Consequently, when large portions of the training data systematically reflect certain groups unfavorably, the resulting language model may replicate or even amplify those biases. Given the growing reliance on LLM-based systems across many fields, this raises important questions about whether these models truly capture the diverse moral perspectives observed in actual human societies.

Despite its importance, the issue of whether LLMs accurately reflect cross-cultural moral judgments has been relatively understudied [1, 15]. In examining how faithfully LLMs capture moral attitudes that vary across cultural contexts, a key consideration is their ability to replicate both the areas of divergence (where cultures disagree) and similarity (where cultures align) on moral topics. Thus, the central research question is:

> *To what extent do language models capture cultural diversity and common tendencies regarding topics on which people around the world tend to diverge or agree in their moral judgments?*

Addressing this question carries both scientific and societal significance. Scientifically, it provides insight into how effectively LLMs, trained primarily on text data, can model complex cultural norms. Societally, ensuring these models reflect actual cross-cultural variation is vital for preventing biased or inaccurate representations of different cultural groups [15]. As LLMs increasingly shape public opinion and decision-making, a mismatch between how cultures truly view moral issues and how models characterize these issues can perpetuate prejudice and unfairness. Conversely, LLMs that accurately capture inter-cultural moral differences and similarities can help reveal common ground and support cross-cultural understanding.

Against this backdrop, the present study focuses on evaluating the extent to which contemporary LLMs mirror the diversity and patterns of moral judgments observed across cultures. Three primary methods are employed:

1. **Comparing Variances**: We compare the variance in model-generated moral judgments with the variance in survey-based moral judgments across countries.
2. **Cluster Alignment**: We examine the alignment of model-induced country clusters with empirically derived clusters.
3. **Direct Comparative Prompts**: We probe LLMs using tailored prompts to see whether they recognize similarities and differences in moral perspectives between countries.

By using these complementary techniques, this work offers insights into the strengths and limitations of LLMs in depicting cross-cultural moral norms, ultimately informing ongoing discussions about their ethical development and deployment. The remainder of this paper is structured as follows. In Section 2, we review related research on cross-cultural moral judgments in LLMs and the issue of bias in these models. Section 3 describes the data and methods used in our analysis, and Section 4 details the results. We then discuss key findings in and conclude with final remarks in Section 5.

[*] Corresponding Author. Email: h.mohammadi@uu.nl.

2 Related work on moral judgment and LLM bias

2.1 *Cross-Cultural Understanding of Moral Judgments in LLMs*

Moral judgments are evaluations of whether specific actions, intentions, or individuals are morally "good" or "bad," and they can vary widely across cultures due to social norms, religious doctrines, and historical influences [8, 25]. Broadly speaking, Western, Educated, Industrialized, Rich, and Democratic societies—commonly abbreviated as W.E.I.R.D. in cross-cultural psychology literature [9], tend to prioritize autonomy[1], individual rights, and personal choice, whereas many non-W.E.I.R.D. cultures place a higher emphasis on communal obligations, duty, and spiritual purity [6]. For instance, individuals in W.E.I.R.D. contexts commonly regard sexual behaviors as a matter of personal freedom, while those from more community-oriented cultures may treat the same behaviors as collective moral issues.

Scholars such as Johnson et al. [10] and Benkler et al. [3] refer to this diversity of valid yet conflicting moral values as "moral value pluralism." Kharchenko et al. [12] caution that LLMs often fail to accurately reflect this pluralism, partly because these models are trained on large but not necessarily diverse datasets. Du et al. [5] likewise note that an overemphasis on English-language training data can overshadow the linguistic and cultural richness of the real world, highlighting the importance of multilingual corpora and larger model sizes. Indeed, Arora et al. [1] propose that multilingual LLMs hold promise for capturing cross-cultural values, though the potential lack of diversity within available multilingual data remains a limiting factor.

Consistent with these concerns, Benkler et al. [3] argue that most AI systems mirror the dominant values of the culture (often Western) producing the majority of the data. This phenomenon can result in a moral bias, whereby W.E.I.R.D. norms and perspectives are incorrectly treated as universally applicable. Empirical investigations of whether LLMs uphold or correct such biases are limited. Some work suggests that they struggle to reproduce culturally specific moral codes [1, 3], while other findings are more optimistic about LLMs' capacity to model cultural diversity [24, 18]. This divergence underscores the importance of continued research on how language models perceive and replicate moral frameworks across various cultures.

2.2 *The Risk of Bias in LLMs*

Bias in LLMs arises when these models inherit or amplify prejudices present in their training datasets. Typically, LLMs learn language representations (or embeddings) by analyzing co-occurrences of words across massive corpora. If these corpora disproportionately depict certain groups or behaviors negatively, the learned representations can perpetuate or exacerbate harmful stereotypes in model outputs [21].

A well-known example is the gender bias identified in word embeddings, where terms like "woman" are closely associated with "homemaker" and "man" with "computer programmer" [4]. Another instance is GPT-3's tendency to associate "Muslims" with violent acts more than "Christians" [10]. Recent work has shown that while demographic biases influence LLM outputs, content-specific features remain the dominant factor in model predictions [19]. Although ongoing research aims to mitigate bias [16], this task remains daunting, as biased outputs can influence everything from public sentiment to automated hiring decisions [22].

For instance, an LLM trained on biased sources might disproportionately recommend men for technical positions, perpetuating gender inequality [4]. In a similar vein, consistently linking certain religious groups with violence can reinforce negative stereotypes and intensify discrimination. Given these high-stakes consequences, developing models that faithfully capture cultural diversity rather than simplifying or skewing moral perspectives is not merely an academic challenge but a moral and societal imperative [28].

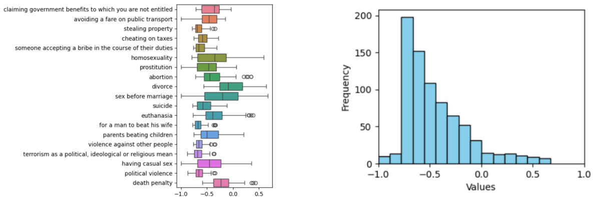
In summary, these two strands of literature, (1) how LLMs handle cross-cultural moral judgments and (2) how bias emerges and persists in LLMs, highlight the need to systematically examine how well these models capture the complexities of moral values across different societies. The following sections detail our data sources and methodological approach to investigating these issues.

3 Data and methods

3.1 *Datasets*

The World Values Survey[2] (WVS) provides detailed information on people's values across cultures. In this study, we use data from Wave 7 [7], which covers the period 2017–2020. This wave features participants from 55 countries who responded to 19 statements on moral issues (e.g., divorce, euthanasia, political violence, cheating on taxes). The survey was administered in the primary languages of each country, offering multiple response categories.

Only the country name and each response were retained, with values normalized to range from $[-1, 1]$, where -1 indicates "never justifiable" and 1 signifies "always justifiable." These normalized scores facilitate comparability and statistical analysis. For each country–moral issue pair, we computed an average (mean) rating, thus capturing a broad overview of each country's position. We acknowledge that averaging can obscure outlier or minority perspectives, but it was deemed the most feasible approach for this study. Figure 1 depict the overall distribution of these normalized scores and their variation across topics and countries.



(a) (b)

Figure 1: (a) Spread of responses across moral topics and countries in WVS Wave 7. (b) Distribution of normalized WVS Wave 7 answers.

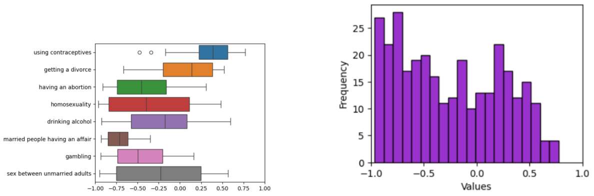
As a second dataset, we use the Pew Global Attitudes Project[3] (2013), which surveyed 39 countries (100 participants each) on 8 moral topics, such as drinking alcohol or getting a divorce. The questionnaire was administered in English, allowing respondents to categorize a topic as "morally acceptable," "not a moral issue," or "morally unacceptable."

We extracted only country names and responses (Q84A–Q84H), again transforming them to a $[-1, 1]$ scale and averaging scores by country–topic pair. Figure 2 summarizes the distribution of normalized scores and their topic-level variation.

[1] The acronym W.E.I.R.D. is a technical term from cross-cultural psychology used to identify a specific cluster of societies that are overrepresented in psychological research. It was introduced to highlight sampling bias in behavioral sciences and has become standard terminology in the field.

[2] https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

[3] https://www.pewresearch.org/dataset/spring-2013-survey-data/



(a) (b)

Figure 2: (a) Spread of responses across moral topics and countries for PEW 2013. (b) Distribution of normalized PEW 2013 responses.

3.2 Pre-processing

In the preprocessing of version 5 of the World Values Survey (WVS) data, the dataset was initially filtered to retain only the columns corresponding to the moral questions Q177 to Q195 and the country code (B_COUNTRY). These questions cover a range of moral issues, such as tax cheating, accepting bribes, and attitudes towards homosexuality. Following the initial filtering, country names were assigned to each row based on the B_COUNTRY codes using a predefined country mapping dataset. Responses with values of -1, -2, -4, and -5, which represent 'Don't know,' 'No answer,' 'Not asked in survey,' and 'Missing; Not available,' respectively, were replaced with zero. This adjustment was made to ensure that calculations, such as averaging, were not affected by non-responses. The decision to replace with 0 ensures that the structure of the dataset remains intact. It avoids introducing NaN values or leaving cells empty, which could complicate subsequent data analysis tasks such as averaging or statistical modeling. Moreover, a replacement value of 0 ensures that non-responses do not influence the computed averages or other aggregated measures artificially. After replacing non-response values with 0, the dataset was aggregated by country, calculating the mean response for each moral question per country. This provided a country-specific average score for each ethical issue. To enable comparisons across different countries and questions, these average scores were normalized on a scale from -1 to 1, where 1 signifies that the behavior is justifiable in every case and -1 denotes it is never justifiable. This normalization involved adjusting the mean responses, which initially ranged from 1 to 10, to fit the new scale. This step was needed for cross-national comparisons. Finally, normalized values were rounded to four decimal places to enhance clarity.

3.3 Models

We begin with two monolingual English models. The first is **GPT-2**, chosen for its strong performance in generating coherent, contextually relevant text [23]. We use two versions from Hugging Face,*GPT-2 Medium* (355M parameters) and *GPT-2 Large* (774M parameters), to observe how increased model size influences their capacity to interpret morally charged content. Larger models generally capture more complex patterns and may better approximate cultural moral judgments. As a second monolingual model, we employ the **OPT** series by Meta AI [27]. Two variants, *OPT-125M* and *OPT-350M*, are included to benchmark smaller, computationally efficient architectures against larger ones. OPT models, like GPT-2, generate text by predicting the next word in a sequence, having been trained on diverse English corpora.

Next, we incorporate multilingual models to explore how exposure to varied linguistic data might shape moral judgments across different countries. The first is **BLOOM**, a transformer-based, autoregressive language model from the BigScience project, trained on 46 natural and 13 programming languages [14]. We specifically use BLOOM-560M and BLOOMZ-560M (fine-tuned for zero-shot learning), rather than the full 176B version, to keep computational requirements manageable. BLOOM's design aims for strong cross-lingual performance, offering a flexible approach for text tasks in multiple languages. Lastly, we include **Qwen**, developed by Alibaba Cloud. Qwen is also a multilingual transformer trained on 29 languages (including English and Chinese). Its latest versions demonstrate competitive results in language understanding, multilingual tasks, coding, and reasoning. For feasibility, we use the 0.5B parameter variant in this study.

While hundreds of open-source LLMs currently exist, our model selection was guided by several criteria designed to ensure a representative sample of the current landscape. We prioritized architectural diversity by including both traditional autoregressive models (GPT-2 and OPT) and multilingual transformer models (BLOOM and Qwen), allowing us to test whether different architectural approaches affect cultural understanding. The selected models also span a range of sizes from 125M to 774M parameters, enabling us to investigate whether increased model capacity correlates with better cross-cultural moral representation. Furthermore, we deliberately contrasted monolingual English models with multilingual ones to test the hypothesis that exposure to diverse languages during training might improve cultural sensitivity. All selected models are openly available, ensuring our experiments can be replicated and extended by other researchers. Finally, these specific models have been extensively studied in prior cultural bias research [1, 10], allowing our findings to build upon and be compared with existing literature. While we acknowledge this represents only a subset of available models, we argue it captures key variations in architecture, size, and training approach that are most relevant to understanding cultural bias. Future work should certainly expand this analysis to newer model families such as LLaMA, Mistral, and other emerging architectures to validate and extend our findings.

3.4 Evaluation metrics

We employ three methods to evaluate whether the selected LLMs capture variations and agreements in moral views across cultures:

1. Comparing the variances of survey-based and model-generated moral scores across countries.
2. Assessing alignment between country clusters derived from these respective scores.
3. Probing LLMs with direct comparative prompts to evaluate how they recognize cross-cultural moral differences and similarities.

For the first two methods, we compute a model-generated moral score for each country–topic pair by prompting the LLM and examining its log probabilities [18]. Two prompt templates are used:

- `In {country} {topic} is {moral_judgment}.`
- `People in {country} believe {topic} is {moral_judgment}.`

Pair #	Contrasting Statements
1	*always justifiable / never justifiable*
2	*right / wrong*
3	*morally good / morally bad*
4	*ethically right / ethically wrong*
5	*ethical / unethical*

Table 1: Token pairs used to prompt the model for moral judgments.

As shown in Table 1, five contrasting token pairs (e.g., *always justifiable* vs. *never justifiable*) are employed to elicit the LLM's stance. When probing, for example, the moral score on "abortion" in the United States using the first token pair, we issue:

> *In the United States abortion is always justifiable*
> *In the United States abortion is never justifiable*

The LLM outputs log probabilities for each statement. We subtract the log probability of the "immoral" statement from that of the "moral" statement to obtain a pair-specific score. We do this for all five token pairs in both prompt styles and average the results to produce a final model-generated moral score. This score mirrors the format of the empirical survey-based scores.

3.5 Evaluation

3.5.1 Comparison of Variances

We first compare cross-cultural variances in moral scores. For each topic, we compute the variance of survey-based scores and model-generated scores, identifying high-variance (controversial) and low-variance (widely agreed upon) topics. We then assess the association between these two sets of variances via Pearson's correlation. A strong positive correlation suggests the model effectively captures cross-country variability; a weak or negative correlation suggests otherwise [24]. This step also helps compare which topics surveys and models respectively deem most controversial or consensual.

3.5.2 Cluster Alignment

Next, we evaluate whether models replicate empirical cross-country patterns of moral judgment. We cluster countries based on their survey scores using K-means, determining the optimal number of clusters K via repeated runs and selecting the highest silhouette score. The same K is then applied to model-generated scores. We compare these two clusterings using Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) [20, 13], as well as their average (CAS), to quantify how well the model-based clusters align with the survey-based "ground truth."

We repeat this procedure thrice for each dataset: once using all moral topics, once focusing on the most controversial topics, and once on those with highest agreement. This reveals how well the models capture both contentious areas and common ground across cultures.

3.5.3 Probing with Direct Comparative Prompts

Finally, we use explicit pairwise comparisons to assess whether models recognize similarities and differences in moral judgments across clusters. For each topic, we apply hierarchical clustering to the survey-based scores, then isolate the two most divergent clusters (those with the largest difference in mean moral score). We repeatedly select representative countries from each cluster and prompt the model using a structure of the form:

> *Regarding the morality of {topic}, the judgments of people in {country x} and {country y} are {comparative token}.*

where {comparative token} takes one of the pairs from Table 1 (e.g., *similar/dissimilar*). We compare the model's predicted label, "similar" or "different", with the empirical classification (same cluster vs. different clusters). Confusion matrix metrics and a Chi-squared test of association then measure how closely the model's judgments align with the survey-based groupings.

Model	WVS		PEW	
	r	*p*	*r*	*p*
GPT-2 Medium	-0.195	0.424	-0.090	0.832
GPT-2 Large	-0.115	0.640	0.617	0.103
OPT-125	-0.035	0.887	-0.095	0.822
QWEN	-0.200	0.413	0.102	0.811
BLOOM	-0.118	0.631	0.608	0.110

Table 2: Correlation between topic variances (WVS and PEW) and model-generated moral score variances. None of the correlations reach statistical significance (all $p > 0.05$).

Source	WVS		PEW	
	Mean score	*Var.*	*Mean score*	*Var.*
Empirical	-0.576	0.075	-0.244	0.138
BLOOM	0.474	0.004	0.246	0.006
OPT-125	0.104	0.012	0.248	0.027
QWEN	0.242	0.021	0.221	0.019
GPT-2 Large	0.323	0.015	0.160	0.032
GPT-2 Medium	0.411	0.013	0.227	0.024

Table 3: Mean moral scores and variances for WVS and PEW topics compared with model-generated values.

4 Results

To evaluate how well language models capture cross-cultural moral variability, we compare the topic-level variance from two survey datasets (WVS and PEW) with the variance of the corresponding model-generated moral scores. Table 2 summarizes the Pearson correlation (r) values, with associated p-values, for each model across both datasets.

WVS Variance Correlations. The weak negative correlations for all models on the WVS dataset indicate that the model-generated variance does not align with the observed cross-cultural diversity in these topics. Specifically, there is no statistically significant evidence that LLMs capture the degree of controversy reflected in WVS responses. The negative but insignificant correlations highlight how these models fall short in capturing the full range of intercultural nuance.

PEW Variance Correlations. On the PEW dataset, correlations are slightly more favorable for GPT-2 Large (r = 0.617) and BLOOM (r = 0.608), suggesting a somewhat better capability to capture topic-level variability. However, even these moderate-to-strong relationships do not achieve statistical significance. In sum, no model consistently reproduces the magnitude of cross-cultural disagreement measured by the PEW data.

Table 3 compares the empirical mean moral scores and variance with those generated by each model. We observe a consistent tendency across both WVS and PEW for the models to assign higher mean moral scores (i.e., more *morally acceptable*) and systematically lower variance than in the survey data. This pattern underscores the models' tendency to view topics as more morally approved and less controversial than they are in reality.

These lower variances suggest that the models underestimate the degree of cultural disagreement, especially on polarizing issues such as sexuality and family-related norms.

Figures 3 and 4 illustrate the mismatch between empirical and model-inferred moral variance. Although full rankings for each model are provided in the appendix, Table 4 summarizes the top three most controversial and most agreed-upon topics, based on *empirical* data from WVS and PEW.

From Table 4, *sex before marriage* and *homosexuality* rank among the most polarizing topics in both datasets, with variances of 0.219

WVS		PEW	
Topic	*Var.*	*Topic*	*Var.*
Most controversial			
Sex before marriage	0.219	Sex between unmarried adults	0.268
Homosexuality	0.209	Homosexuality	0.216
Euthanasia	0.126	Drinking alcohol	0.157
Most agreed upon			
Stealing property	0.015	Married people having an affair	0.021
Violence against other people	0.015	Using contraceptives	0.086
For a man to beat his wife	0.018	Gambling	0.097

Table 4: Top three most controversial and most agreed-upon topics from WVS (left) and PEW (right) empirical data.

and 0.209 respectively in WVS, and 0.268 and 0.216 in PEW. These high variances indicate substantial cross-cultural disagreement on these topics, which aligns with prior literature suggesting that sexual and family-related moral issues often reflect deep cultural differences between societies that prioritize individual autonomy versus those emphasizing communal values and traditional norms [6, 25]. The fact that these particular topics show the highest variance suggests they serve as key differentiators between moral frameworks across cultures. However, several models misjudge at least one of these issues as relatively uncontroversial, with QWEN and BLOOM even ranking homosexuality among their most agreed-upon topics (as shown in the appendix), suggesting they fail to capture these fundamental cultural divisions.

Although GPT-2 Large and BLOOM show moderate correlations in the PEW dataset (Table 2), no model achieves statistically significant alignment with the empirical data. Across both WVS and PEW, language models:

1. Overestimate moral acceptability, assigning more positive moral judgments to most topics.
2. Underestimate the degree of cultural disagreement, producing lower variance scores.

These findings suggest that current LLMs do not yet mirror real-world moral heterogeneity, especially for hotly debated topics like sexual and family norms. Simply increasing model size may be insufficient; more nuanced training or alignment with culturally diverse data sources may be necessary to capture the complexity seen in empirical moral attitudes.

4.1 Cluster Alignment

We analyze how closely the clusters induced by model-generated moral scores align with the empirical clusters derived from both the WVS and PEW datasets. Three metrics are used to measure this alignment: the Adjusted Rand Index (ARI), the Adjusted Mutual Information (AMI), and the Combined Alignment Score (CAS). Higher values on these metrics suggest better agreement between the empirical clusters and the model-generated clusters.

Table 6 combines the alignment scores for all topics in WVS (left panel) and PEW (right panel). QWEN shows notably higher metrics on WVS than the other models, indicating closer alignment to empirical scores. For PEW, GPT-2 Large and OPT-125 share moderate alignment scores, while QWEN and BLOOM perform relatively worse.

Table 7 shows the alignment results for the most controversial topics in both WVS and PEW. All models yield negative or near-zero alignment on WVS. On PEW, GPT-2 Medium remains negative while GPT-2 Large, OPT-125, and BLOOM achieve positive scores, with OPT-125 notably highest.

Table 8 reports the alignment results for the most agreed-upon topics. For WVS, GPT-2 Medium and OPT-125 have positive scores, whereas GPT-2 Large, QWEN, and BLOOM remain negative. For PEW, GPT-2 Medium, GPT-2 Large, and OPT-125 show moderate positive alignment; QWEN and BLOOM exhibit minimal or negative scores.

4.2 Probing with Direct Comparative Prompts

We further examine how models recognize similarities and differences in moral judgments by prompting them to compare topics directly. Tables 9 and 10 show, respectively, the confusion-matrix scores and chi-squared results for WVS (left) and PEW (right).

Accuracy for all models hovers near 0.5. GPT-2 Large and QWEN stand out with high recall (0.946 and 0.831, respectively), but their precision is lower, yielding moderate F1 scores. BLOOM displays poor performance across most metrics, indicating difficulties in classifying positive and negative instances.

Again, overall accuracy remains near 0.5 for all models. GPT-2 Large shows the highest recall (0.954), while QWEN achieves a recall of 0.694. BLOOM exhibits very low recall and precision, resulting in the lowest F1.

Table 10 shows that GPT-2 Medium exhibits a significant ($p < 0.01$) alignment with WVS, implying its judgments correlate with actual moral (dis)similarities. The other models do not significantly align with WVS. For PEW, BLOOM yields a statistically significant ($p < 0.05$) result—though this may reflect a consistent but incorrect pattern, given its poor F1 and recall.

Although some models (e.g., GPT-2 Large, QWEN) display high recall indirect probing, their precision is often lacking. GPT-2 Medium is uniquely significant in the WVS chi-squared test, while BLOOM is significant in the PEW test but shows low classification performance overall. These divergences suggest that while models capture certain aspects of moral similarity, they struggle to reflect the full complexity of real-world intercultural judgments.

5 Discussion and conclusion

The findings of this study shed light on the capability of LLMs to accurately capture cultural diversity and common tendencies across different moral topics. The investigation utilized multiple methodologies that were based on probing LLMs with prompts derived from the World Values Survey (WVS) and PEW datasets, focusing on a range of moral topics.

5.1 Comparison of variance

The correlation analysis between model-generated moral scores and empirical survey data revealed mixed results. For the PEW dataset, GPT-2 Large and BLOOM demonstrated moderate to strong alignment in capturing cultural variations. The fact that the largest model (GPT-2 Large) and the largest multilingual model (BLOOM) performed best may suggest that model size and multilinguality have a positive effect on models' ability to grasp patterns of cultural diversity, which would be in line with previous work from Du et al. [5] and Arora et al. [1]. However, the correlations did not reach statistical significance and therefore no strong claims can be made. Moreover, model performance shows high variability, with weak negative correlations observed for both GPT-2 Large and BLOOM when comparing their variances with the WVS moral score variances. The other models performed weakly and variably in both the PEW and WVS moral

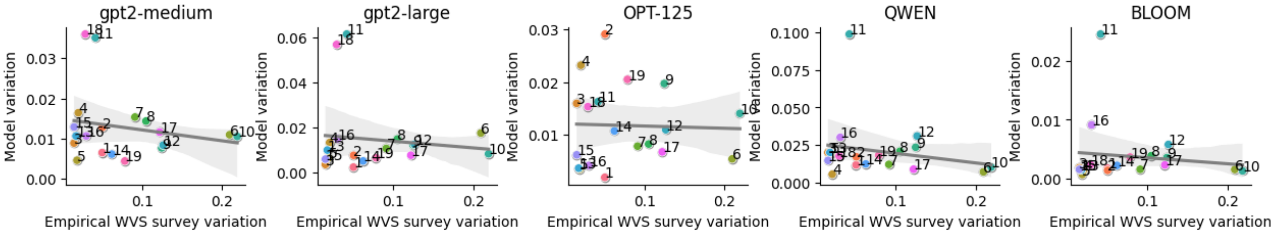


Figure 3: Comparison of empirical and model-inferred moral score variances for WVS topics. The models underestimate cross-cultural disagreements.

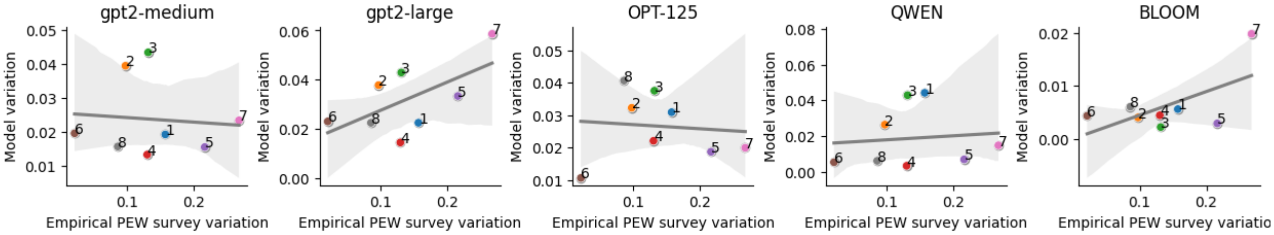


Figure 4: Comparison of empirical and model-inferred moral score variances for PEW topics. Again, models generally exhibit lower variance.

score variance comparisons. Furthermore, the models struggled to accurately identify the most controversial and agreed on topics. In fact, some of the models incorrectly categorized (one of) the two most controversial topics as among the most agreed on. The variable and low overall performance could be attributed to the fact that the complexity and nuance of moral values across different cultural contexts may not be fully captured by the models' training data.

5.2 Cluster alignment

The clustering alignment results further emphasized the variability in model performance. Overall, GPT-2 Large and OPT-125 showed better alignment with empirical moral scores from both datasets quite consistently, suggesting their relative proficiency in clustering countries based on moral attitudes. However, other models, most notably BLOOM, exhibited lower alignment scores, indicating shortcomings in their ability to mirror the clustering patterns observed in the survey data. These results suggest that the models fall short in grasping cultural patterns regarding moral judgments, which is in line with the findings from the previous method. Thus, while GPT-2 Large and OPT-125 generally show better alignment with empirical moral scores across various topics, the variability in model performance underscores the challenges in accurately capturing the complexities of moral attitudes across different cultural contexts. Overall, the clusterings based on the model scores do not faithfully capture the cultural patterns observed in the clusterings derived from the survey scores.

5.3 Probing with direct comparative prompts

Direct probing with comparative prompts provided additional insights into the models' understanding of moral differences between culturally distinct groups. In general, performance is low as the scores are no higher or even slightly lower than random chance. GPT-2 Large and QWEN stood out with higher accuracy and recall scores, indicating their better performance in distinguishing moral differences between the most divergent clusters identified by the survey data. Upon further inspection, however, it became clear that GPT-2 Large and QWEN almost always predict the same class, which does not signify a proper understanding of inter-cultural differences and similarities. If we disregard the performance of GPT-2 Large and QWEN due to the fact that they always predict the same class, GPT-2 Medium and OPT-125 exhibit the most balanced performance across the remaining models. BLOOM exhibited the lowest performance metric scores, suggesting challenges in discerning nuanced moral judgments across cultures. Notably, despite its low overall performance, BLOOM's judgments were found to be statistically associated with the judgments based on the PEW dataset through a Chi-squared test. This suggests that there may be some alignment between BLOOM's outputs and the moral judgments reflected in the PEW dataset. However, it is important to note that this statistical association does not necessarily imply a meaningful understanding or accurate representation of moral differences between cultures.

5.4 Conclusion

In conclusion, the study underscores the importance of rigorous evaluation methodologies when assessing LLMs' ability to understand and reflect cultural diversity in moral judgments. The tested models seem to propagate a homogenized view on cross-cultural moral values, identifying most topics as cross-culturally agreed on as more morally acceptable than empirically observed. Thereby, the models generally seem to reflect a rather liberal view, in line with the autonomy-endorsing values found in W.E.I.R.D. societies [6]. It has been established in the literature that exclusively English training data plays a big part in the embedding of homogenous W.E.I.R.D. values and, thereby, cultural bias in LLMs [3]. This could lead one to believe that multilingual LLMs are the answer to mitigating bias in LLMs [1]. However, this study could not find convincing evidence to suggest that multilingual models are better at truthfully capturing cultural diversities in moral judgments than monolingual models. Similarly, while model size could be considered another factor influencing model performance due to its potential to enhance computational capacity and capture more complex patterns [5], its impact was not found to be convincing in the carried out analyses. It can be concluded that this study found no remarkable differences between the tested models in their success, regardless of multilinguality or model size. Overall, the models examined show variable performance and generally exhibit

Model	Survey	Survey Var.	Survey Mean	Model Var.	Model Mean	Topic	Var. Diff
GPT-2 Medium	WVS	0.219	-0.244	0.011	0.465	sex before marriage	0.208
	WVS	0.209	-0.396	0.011	0.577	homosexuality	0.198
	WVS	0.126	-0.430	0.008	0.481	euthanasia	0.118
	WVS	0.125	-0.150	0.008	0.217	divorce	0.117
	WVS	0.122	-0.452	0.012	0.371	having casual sex	0.110
	PEW	0.268	-0.219	0.023	0.044	sex between unmarried adults	0.244
	PEW	0.216	-0.342	0.016	0.641	homosexuality	0.201
	PEW	0.157	-0.234	0.019	0.142	drinking alcohol	0.138
GPT-2 Large	WVS	0.219	-0.244	0.008	0.454	sex before marriage	0.211
	WVS	0.209	-0.396	0.018	-0.086	homosexuality	0.192
	WVS	0.122	-0.452	0.008	0.470	having casual sex	0.114
	WVS	0.126	-0.430	0.013	0.261	euthanasia	0.114
	WVS	0.125	-0.150	0.012	0.121	divorce	0.112
	PEW	0.268	-0.219	0.059	-0.138	sex between unmarried adults	0.209
	PEW	0.216	-0.342	0.033	-0.188	homosexuality	0.183
	PEW	0.157	-0.234	0.023	0.210	drinking alcohol	0.135
OPT-125	WVS	0.219	-0.244	0.014	0.475	sex before marriage	0.205
	WVS	0.209	-0.396	0.005	0.255	homosexuality	0.204
	WVS	0.126	-0.430	0.011	0.013	euthanasia	0.115
	WVS	0.122	-0.452	0.007	0.093	having casual sex	0.115
	WVS	0.125	-0.150	0.020	-0.261	divorce	0.105
	PEW	0.268	-0.219	0.020	0.512	sex between unmarried adults	0.248
	PEW	0.216	-0.342	0.019	0.570	homosexuality	0.198
	PEW	0.157	-0.234	0.031	0.187	drinking alcohol	0.126
QWEN	WVS	0.219	-0.244	0.010	0.415	sex before marriage	0.209
	WVS	0.209	-0.396	0.007	0.466	homosexuality	0.202
	WVS	0.122	-0.452	0.009	0.177	having casual sex	0.113
	WVS	0.125	-0.150	0.024	-0.042	divorce	0.101
	WVS	0.126	-0.430	0.031	-0.115	euthanasia	0.095
	PEW	0.268	-0.219	0.015	0.494	sex between unmarried adults	0.253
	PEW	0.216	-0.342	0.007	0.562	homosexuality	0.209
	PEW	0.130	-0.405	0.004	0.130	having an abortion	0.127
BLOOM	WVS	0.219	-0.244	0.001	0.662	sex before marriage	0.218
	WVS	0.209	-0.396	0.002	0.865	homosexuality	0.208
	WVS	0.124	-0.150	0.004	0.569	divorce	0.121
	WVS	0.126	-0.429	0.006	0.712	euthanasia	0.121
	WVS	0.122	-0.452	0.002	0.422	having casual sex	0.120
	PEW	0.268	-0.219	0.020	0.374	sex between unmarried adults	0.248
	PEW	0.216	-0.342	0.003	0.843	homosexuality	0.213
	PEW	0.157	-0.234	0.006	0.159	drinking alcohol	0.152

Table 5: Variance gaps between survey data and model outputs (WVS vs. PEW), showing the top eight topic–model pairs with the largest differences. Full results in the Appendix.

Model	WVS			PEW		
	ARI	AMI	CAS	ARI	AMI	CAS
GPT-2 Medium	-0.012	-0.002	-0.007	0.087	0.068	0.078
GPT-2 Large	0.028	0.040	0.034	0.129	0.123	0.126
OPT-125	-0.073	0.037	-0.018	0.129	0.123	0.126
QWEN	0.291	0.138	0.215	-0.019	0.065	0.023
BLOOM	0.015	-0.011	0.002	0.008	-0.004	0.002

Table 6: Cluster alignment scores for all topics in WVS (left) and PEW (right).

Model	WVS			PEW		
	ARI	AMI	CAS	ARI	AMI	CAS
GPT-2 Medium	-0.015	-0.011	-0.013	-0.026	-0.019	-0.022
GPT-2 Large	-0.012	0.023	0.005	0.093	0.081	0.087
OPT-125	-0.021	0.017	-0.002	0.131	0.140	0.136
QWEN	-0.014	-0.018	-0.016	-0.006	0.073	0.033
BLOOM	-0.015	-0.011	-0.013	0.009	0.006	0.007

Table 7: Cluster alignment scores for most controversial topics in WVS (left) and PEW (right).

Model	WVS			PEW		
	ARI	AMI	CAS	ARI	AMI	CAS
GPT-2 Medium	0.079	0.010	0.044	0.057	0.045	0.051
GPT-2 Large	-0.019	-0.014	-0.016	0.028	0.020	0.024
OPT-125	0.120	0.038	0.079	0.035	0.051	0.043
QWEN	-0.005	-0.017	-0.011	-0.020	-0.016	-0.018
BLOOM	-0.030	-0.012	-0.021	0.006	0.004	0.005

Table 8: Cluster alignment scores for most agreed-upon topics in WVS (left) and PEW (right).

Model	WVS				PEW			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
GPT-2 Medium	0.485	0.488	0.336	0.398	0.495	0.494	0.402	0.444
GPT-2 Large	0.509	0.508	0.946	0.661	0.495	0.497	0.954	0.654
OPT-125	0.502	0.510	0.461	0.484	0.506	0.506	0.480	0.493
QWEN	0.500	0.504	0.831	0.628	0.493	0.495	0.694	0.578
BLOOM	0.495	0.543	0.026	0.050	0.497	0.326	0.006	0.011

Table 9: Confusion matrix scores from direct probing on WVS (left) and PEW (right). Acc. = accuracy, Prec. = precision, Rec. = recall.

low success in aligning with empirical moral data from global surveys. Thus, ongoing research and development are needed to enhance their accuracy and reliability in diverse cultural settings. Addressing these challenges is crucial for ensuring the ethical integrity and societal impact of AI technologies in the context of global applications.

Model	**WVS**		**PEW**	
	χ^2	**p**	χ^2	**p**
GPT-2 Medium	8.38	0.004**	0.418	0.518
GPT-2 Large	1.491	0.222	3.325	0.068
OPT-125	0.338	0.561	0.609	0.435
QWEN	1.416	0.234	1.017	0.313
BLOOM	1.279	0.258	4.599	0.032*

Table 10: Chi-squared test results from direct probing on WVS (left) and PEW (right). $(**)$ indicates $p < 0.01$, $(*)$ indicates $p < 0.05$.

5.5 Conclusion

In conclusion, the study underscores the importance of rigorous evaluation methodologies when assessing LLMs' ability to understand and reflect cultural diversity in moral judgments. The tested models seem to propagate a homogenized view on cross-cultural moral values, identifying most topics as cross-culturally agreed on as more morally acceptable than empirically observed. Thereby, the models generally seem to reflect a rather liberal view, in line with the autonomy-endorsing values found in Western, Educated, Industrialized, Rich, and Democratic (W.E.I.R.D.) societies [6]. It has been established in the literature that exclusively English training data plays a big part in the embedding of homogenous W.E.I.R.D. values and, thereby, cultural bias in LLMs [3]. This could lead one to believe that multilingual LLMs are the answer to mitigating bias in LLMs [1]. However, this study could not find convincing evidence to suggest that multilingual models are better at truthfully capturing cultural diversities in moral judgments than monolingual models.

Based on our findings, several actionable strategies could improve cultural representativeness in LLMs. First, diversifying training data by prioritizing text from underrepresented regions and languages would help counteract the current bias toward W.E.I.R.D. perspectives. This includes incorporating religious texts, local news sources, and cultural forums that discuss moral topics from non-W.E.I.R.D. societies. Second, culture-aware fine-tuning approaches could be developed using datasets that explicitly represent diverse moral perspectives on controversial topics, weighted to reflect actual global population distributions rather than internet data availability. Third, prompt engineering strategies that explicitly invoke cultural context could elicit more culturally diverse responses. For example, prompts like "From the perspective of someone in [country] with traditional values..." may help models access different moral frameworks. Finally, establishing standardized evaluation frameworks using surveys like WVS and PEW would enable regular assessment of cultural bias in new models before deployment. These recommendations provide concrete pathways for researchers and practitioners working toward more culturally inclusive AI systems. The challenges identified here align with broader issues in developing transparent and interpretable NLP systems across various domains [17], emphasizing the need for continued research in explainable AI methods.

6 Limitations

While this study provides important insights, it is important to recognize certain boundaries of our approach. First, although WVS and PEW are well-established surveys covering 55 and 39 countries respectively, they organize complex moral views into fixed categories, which may not capture every nuance or implicit aspect of moral reasoning. Additionally, our analysis examined aggregate patterns across all countries rather than country-specific contributions to variance. Future work could benefit from analyzing which specific countries or regions show the largest discrepancies between model outputs and survey responses, which would provide more granular insights into geographical patterns of model bias. Second, we focused on a selected group of models, so our findings primarily reflect these particular architectures. Third, the choice of prompts in our experiments can influence model responses [26], meaning that exploring alternative prompt strategies could yield additional insights. Lastly, due to computational limits, we randomly selected topics in Method 3, which may not cover all diversity within each cluster. Future research can build on our work by testing a wider range of models, experimenting with different prompt designs, analyzing country-specific patterns, and using broader topic sampling to further enrich the analysis.

7 Ethics Statement

The work relies exclusively on two publicly available, anonymised survey datasets (WVS Wave 7 and Pew 2013) and on open-access language models. No personal or sensitive information was collected, and all analyses were performed in accordance with the data providers' terms of use. By quantifying cultural bias in LLMs we aim to support fairer deployment of generative AI and to encourage the creation of more globally representative training data.

Acknowledgements

We appreciate the maintainers of WVS and PEW data for enabling large-scale cross-cultural analysis.

References

[1] A. Arora, L. Kaffee, and I. Augenstein. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*, 2022. URL https://doi.org/10.48550/arxiv.2203.13722.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

[3] Y. Benkler, D. Mosaphir, S. Friedman, A. Smart, and S. Schmer-Galunder. Assessing llms for moral value pluralism. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2312.10075. URL https://doi.org/10.48550/arxiv.2312.10075.

[4] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Quantifying and reducing stereotypes in word embeddings. *arXiv*, 2016. doi: 10.48550/arxiv.1606.06121. URL https://arxiv.org/abs/1606.06121.

[5] X. Du, Z. Yu, S. Gao, D. Pan, Y. Cheng, Z. Ma, R. Yuan, X. Qu, J. Liu, T. Zheng, X. Luo, G. Zhou, B. Yuan, W. Chen, J. Fu, and G. Zhang. Chinese Tiny LLM: Pretraining a Chinese-Centric Large Language Model. *arXiv (Cornell University)*, 4 2024. doi: 10.48550/arxiv.2404.04167. URL https://arxiv.org/abs/2404.04167.

[6] J. Graham, P. Meindl, E. Beall, K. M. Johnson, and L. Zhang. Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, 8:125–130, 2016. doi: 10.1016/j.copsyc.2015.09.007. URL https://doi.org/10.1016/j.copsyc.2015.09.007.

[7] C. W. Haerpfer, P. Bernhagen, R. F. Inglehart, and C. Welzel. *World Values Survey: Round Seven - Country-Pooled Datafile Version*. Institute for Comparative Survey Research, Vienna, 2022. URL http://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp.

[8] J. Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001. doi: 10.1037/0033-295X.108.4.814.

[9] J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[10] R. L. Johnson, G. Pistilli, N. Menédez-González, L. D. D. Duran, E. Panai, J. Kalpokiene, and D. J. Bertulfo. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv.org*, mar 2022. URL https://arxiv.org/abs/2203.07785.

[11] K. Karpouzis. Plato's shadows in the digital cave: Controlling cultural bias in generative ai. *Electronics*, 13(8):1457, 2024. doi: 10.3390/electronics13081457. URL https://doi.org/10.3390/electronics13081457.

[12] J. Kharchenko, T. Roosta, A. Chadha, and C. Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. arXiv preprint, 2024. URL https://doi.org/10.48550/arxiv.2406.14805. arXiv:2406.14805.

[13] D. Lazarenko and T. Bonald. Pairwise adjusted mutual information, 2021. URL https://arxiv.org/abs/2103.12641.

[14] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. d. Moral, O. Ruwase, R. Bawden, and M. J. ... Nelson. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100, 2022. URL https://api.semanticscholar.org/CorpusID:253420279.

[15] C. C. Liu, F. Koto, T. Baldwin, and I. Gurevych. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv*, 2023. URL https://arxiv.org/abs/2309.08591. arXiv:2309.08591.

[16] A. Mishra, G. Nayak, S. Bhattacharya, T. Kumar, A. Shah, and M. Foltin. Llm-guided counterfactual data generation for fairer ai. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1538–1545, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701726. doi: 10.1145/3589335.3651929. URL https://doi.org/10.1145/3589335.3651929.

[17] H. Mohammadi, A. Bagheri, A. Giachanou, and D. L. Oberski. Explainability in practice: A survey of explainable nlp across various domains. *arXiv preprint arXiv:2502.00837*, 2025.

[18] H. Mohammadi, E. Papadopoulou, Y. F. Meijer, and A. Bagheri. Exploring cultural variations in moral judgments with large language models. *arXiv preprint arXiv:2506.12433*, 2025.

[19] H. Mohammadi, T. Shahedi, P. Mosteiro, M. Poesio, A. Bagheri, and A. Giachanou. Assessing the reliability of llms annotations in the context of demographic bias and model explanation. *arXiv preprint arXiv:2507.13138*, 2025.

[20] T. Nazaretsky, S. Hershkovitz, and G. Alexandron. Kappa learning: A new item-similarity method for clustering educational items from response data. 04 2020. URL https://eric.ed.gov/?id=ED599209.

[21] P. Nemani, Y. D. Joel, P. Vijay, and F. F. Liza. Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*, 6:100047, 2024. doi: 10.1016/j.nlp.2023.100047. URL https://doi.org/10.1016/j.nlp.2023.100047.

[22] S. U. Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, New York, NY, 2018. ISBN 978-1479837243. URL https://nyupress.org/9781479837243/algorithms-of-oppression/.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

[24] A. Ramezani and Y. Xu. Knowledge of cultural moral norms in large language models. *arXiv (Cornell University)*, 2023. doi: 10.48550/arxiv.2306.01857. URL https://doi.org/10.48550/arxiv.2306.01857.

[25] R. A. Shweder, N. C. Much, M. Mahapatra, and L. Park. The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In A. Brandt and P. Rozin, editors, *Morality and Health*, pages 119–169. Routledge, 1997. URL https://psycnet.apa.org/record/1997-36447-005.

[26] L. Wang, X. Chen, and X. Deng. Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *npj Digital Medicine*, 7:41, 2024. doi: 10.1038/s41746-024-01029-4. URL https://doi.org/10.1038/s41746-024-01029-4.

[27] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[28] J. Zou and L. Schiebinger. Ai can be sexist and racist—it's time to make it fair. *Nature*, 559(7714):324–326, 2018. doi: 10.1038/d41586-018-05707-8. URL https://www.nature.com/articles/d41586-018-05707-8.

A Appendix

A.1 *Most controversial WVS topics according to models*

Topic	*Model*	*Variance*
Political violence	GPT-2 Medium	0.036
Suicide	GPT-2 Medium	0.035
Cheating on taxes	GPT-2 Medium	0.016

Table 11: Top 3 most controversial WVS topics according to GPT-2 Medium

Topic	*Model*	*Variance*
Suicide	GPT-2 Large	0.062
Political violence	GPT-2 Large	0.057
Homosexuality	GPT-2 Large	0.018

Table 12: Top 3 most controversial WVS topics according to GPT-2 Large

Topic	*Model*	*Variance*
Avoiding a fare on public transport	OPT-125	0.029
Cheating on taxes	OPT-125	0.023
Death penalty	OPT-125	0.021

Table 13: Top 3 most controversial WVS topics according to OPT-125

Topic	*Model*	*Variance*
Suicide	QWEN	0.099
Terrorism as a political, ideological or religious tactic	QWEN	0.030
Euthanasia	QWEN	0.031

Table 14: Top 3 most controversial WVS topics according to QWEN

Topic	*Model*	*Variance*
Suicide	BLOOM	0.025
Terrorism as a political, ideological or religious tactic	BLOOM	0.009
Euthanasia	BLOOM	0.006

Table 15: Top 3 most controversial WVS topics according to BLOOM

A.2 *Most agreed on WVS topics according to models*

Topic	*Model*	*Variance*
Death penalty	GPT-2 Medium	0.004
Accepting a bribe in the course of duty	GPT-2 Medium	0.005
Parents beating children	GPT-2 Medium	0.006

Table 16: Top 3 most agreed on WVS topics according to GPT-2 Medium

Topic	*Model*	*Variance*
Claiming government benefits to which you are entitled	GPT-2 Large	0.002
Stealing property	GPT-2 Large	0.004
Parents beating children	GPT-2 Large	0.005

Table 17: Top 3 most agreed on WVS topics according to GPT-2 Large

Topic	*Model*	*Variance*
Claiming government benefits to which you are entitled	OPT-125	0.002
Someone accepting a bribe in the course of duty	OPT-125	0.003
For a man to beat his wife	OPT-125	0.004

Table 18: Top 3 most agreed on WVS topics according to OPT-125

Topic	*Model*	*Variance*
Cheating on taxes	QWEN	0.006
Homosexuality	QWEN	0.007
Having casual sex	QWEN	0.009

Table 19: Top 3 most agreed on WVS topics according to QWEN

Topic	*Model*	*Variance*
Someone accepting a bribe in the course of duty	BLOOM	0.001
Sex before marriage	BLOOM	0.001
Avoiding a fare on public transport	BLOOM	0.001

Table 20: Top 3 most agreed on WVS topics according to BLOOM

A.3 Most controversial PEW topics according to models

Topic	*Model*	*Variance*
Getting a divorce	GPT-2 Medium	0.043
Gambling	GPT-2 Medium	0.039
Sex between unmarried adults	GPT-2 Medium	0.023

Table 21: Top 3 most controversial PEW topics according to GPT-2 Medium

Topic	*Model*	*Variance*
Sex between unmarried adults	GPT-2 Large	0.059
Getting a divorce	GPT-2 Large	0.043
Gambling	GPT-2 Large	0.038

Table 22: Top 3 most controversial PEW topics according to GPT-2 Large

Topic	*Model*	*Variance*
Using contraceptives	OPT-125	0.041
Getting a divorce	OPT-125	0.038
Gambling	OPT-125	0.032

Table 23: Top 3 most controversial PEW topics according to OPT-125

Topic	*Model*	*Variance*
Drinking alcohol	QWEN	0.044
Getting a divorce	QWEN	0.043
Gambling	QWEN	0.027

t

Table 24: Top 3 most controversial PEW topics according to QWEN

Topic	*Model*	*Variance*
Sex between unmarried adults	BLOOM	0.020
Using contraceptives	BLOOM	0.006
Drinking alcohol	BLOOM	0.006

Table 25: Top 3 most controversial PEW topics according to BLOOM

A.4 Most agreed on PEW topics according to models

Topic	*Model*	*Variance*
Having an abortion	GPT-2 Medium	0.013
Homosexuality	GPT-2 Medium	0.016
Using contraceptives	GPT-2 Medium	0.016

Table 26: Top 3 most agreed on PEW topics according to GPT-2 Medium

Topic	*Model*	*Variance*
Having an abortion	GPT-2 Large	0.015
Using contraceptives	GPT-2 Large	0.023
Drinking alcohol	GPT-2 Large	0.023

Table 27: Top 5 most agreed on PEW topics according to GPT-2 Large

Topic	*Model*	*Variance*
Married people having an affair	OPT-125	0.011
Homosexuality	OPT-125	0.019
Sex between unmarried adults	OPT-125	0.020

Table 28: Top 3 most agreed on PEW topics according to OPT-125

Topic	*Model*	*Variance*
Having an abortion	QWEN	0.004
Married people having an affair	QWEN	0.006
Using contraceptives	QWEN	0.006

Table 29: Top 3 most agreed on PEW topics according to QWEN

Topic	*Model*	*Variance*
Getting a divorce	BLOOM	0.002
Homosexuality	BLOOM	0.003
Gambling	BLOOM	0.004

Table 30: Top 3 most agreed on PEW topics according to BLOOM

A Nightmare on LLMs Street: On the Importance of Cultural Awareness in Text Adaptation for LRLs

David C. T. Freitas[a,*] **and Henrique Lopes Cardoso**[a]

[a]LIACC, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

Abstract. Large Language Models (LLMs) have revolutionized how we generate, interact with, and process language. Still, these models are biased toward WEIRD (Western, Educated, Industrialized, Rich, and Democratic) values. This bias is not merely linguistic but also cultural. Sociocultural contexts influence how people express ideas, interpret meaning, and communicate. In low-resource language settings, where data and cultural representation are limited, this issue becomes even more pronounced when models are applied without cultural adaptation, often leading to outputs that are irrelevant, inaccessible, or even harmful. In this paper, we argue for the importance of incorporating sociocultural context into LLMs. We review existing frameworks that explore culture in Natural Language Processing (NLP), and examine some work aimed at culturally aligning language models. As an illustrative scenario, we analyze the case of Guinea-Bissau. In this linguistically and culturally diverse country, Portuguese is the official language but not the primary means of communication for most of the population, highlighting the urgent need to adapt educational materials to the local sociocultural context. Finally, we propose a revised framework to address the challenge of adapting educational materials to diverse contexts, aiming to improve both the relevance and pedagogical impact of text adaptation.

1 Introduction

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), enabling widespread and seemingly universal interaction with Artificial Intelligence (AI) systems — perhaps for the first time. The abilities of LLMs to understand language rely not only on linguistic and factual knowledge, but also on an awareness of cultural nuances that shape human lives. However, these models are mostly trained on online data that is deeply rooted in a WEIRD (Western, Educated, Industrialized, Rich, and Democratic) worldview [8]. Because the majority of users who access these systems share, or are aligned with, that worldview, this creates a misleading impression that the models adequately represent and understand the world's linguistic and cultural diversity.

In reality, a large portion of the world's population lives in profoundly different sociocultural contexts, with customs, norms, values, and shared knowledge that are not reflected in the data used to train these models. These elements, essential for contextual understanding, vary across cultural groups and are often not captured when NLP is built upon universalist assumptions. The problem is amplified when it comes to Low-Resource Languages (LRLs), whose cultural contexts are often overlooked, for which very few linguistic resources, annotated data, or corpora exist. It is estimated that there are over 7,000 languages in the world[1], but only a small fraction are covered by LLMs.

In light of this, although significant progress has been made in NLP, culture remains among the most challenging aspects of language that LLMs still struggle to handle effectively [1]. A key difficulty lies in the absence of a shared definition of culture, which further complicates efforts to evaluate progress in this area [17].

Guinea-Bissau represents a paradigmatic case of a country left behind in the global access to technology and quality education. This is exacerbated by two key factors. First, Portuguese is the official language and the language of instruction in schools, yet only a minority of the population speaks it fluently. Second, most people communicate in Guinea-Bissau Creole, the *lingua franca* used in everyday life, which lacks official status and a standardized orthography.

This paper sets out to argue for the urgent need to address the lack of cultural awareness in LLMs, identifying their limitations. It advocates for a more context-sensitive approach and explores how tackling these issues can help reduce social and linguistic inequalities in the case of Guinea-Bissau, by enabling the creation and adaptation of content that is both culturally relevant and linguistically appropriate.

This paper should be understood as a position paper grounded in interdisciplinary perspectives from NLP, education, sociolinguistics, and cultural studies. Our aim is not to present experimental results, but to outline conceptual and methodological directions for the culturally informed adaptation of texts in LRLs settings.

2 Background and Motivation

Our culture and our social relations shape everything we do. The way we present information, the style and tone we use, the context, and the common knowledge we share, among many other subtleties, are essential to effective communication. All of these are based on cultural knowledge.

For the purpose of this paper, we adopt the definition of culture proposed by Liu et al. [18], according to which culture encompasses:

> "the collective ideas, shared language, and social practices that emerge from and evolve through human social interactions within a society".

Messages that are not culturally adapted can be misinterpreted, and language technologies must account for cultural context to avoid

[*] Corresponding Author. Email: up199400804@edu.fe.up.pt

[1] https://www.ethnologue.com/

potential harm [9]. Given its crucial role in making LLMs safer, fairer, and more inclusive, the concept of culture is receiving increasing attention in current research.

One of the most important decisions when adapting texts, especially in translation, is deciding whether to aim for literalism or adaptation [28]. Cross-cultural translation and adaptation highlight this complexity, when the same literal meaning in one culture may be inappropriate in another, regardless of fluency [9], or when a direct counterpart in the target culture may not exist. When no direct counterpart exists in the target culture, appositives can be employed to provide contextual explanations. For example, the Portuguese sentence "Encontraram-se na queima das fitas" could be translated to English as "They met at the *queima das fitas*, a traditional Portuguese academic celebration", where the appositive clarifies a culturally specific term that may be unfamiliar to the target audience.

While many LLMs are capable of performing cultural adaptations effectively within WEIRD contexts, they often exhibit biases that hinder their applicability in more culturally diverse settings. Certain cultural variations can be identified through sociocultural elements, such as Culture-Specific Items (CSIs) — including aspects of ecology (flora, fauna, climate, . . .), material culture (food, clothing, housing, transportation, . . .), emotions, and socially sensitive or taboo topics [28]. Yet, such categorizations are insufficient, particularly because culture is not a static inventory of features, but a dynamic and evolving construct. Moreover, every communicative act is situated within a relational structure involving an emitter, a receiver, the medium through which the message is conveyed, and the broader sociocultural context in which it occurs — all of which shape the interpretation and appropriateness of the message.

Despite this growing awareness, many current approaches rely on language or national borders as proxies for cultural identity. However, this is problematic because significant cultural variation can exist within the same country or among countries that share a common language. For example, Portugal and Brazil speak the same language but differ considerably in cultural norms, values, and communicative practices. Similarly, treating entire countries as culturally homogeneous units oversimplifies internal diversity. Such categorical division risks minimizing cultural distinctions under a single label.

Although there is no common definition of culturally aware NLP, most works in this area share common goals. Zhou et al. [37] define the goals of culturally aware NLP as systems that are:

Adaptive: sensitive to specific cultural contexts when generating their outputs.

Discerning: not perpetuating reductive stereotypes.

Inclusive: perform well across a large number of cultures.

Nuanced: achieve depth through more granular and extensive cultural understanding.

These four goals share common ground and, according to the authors, can be grouped into two distinct spaces. The first space combines the "Adaptive" and "Discerning" goals and reflects how systems should respond and when it is appropriate to do so. The second space includes the "Inclusive" and "Nuanced" goals, reflecting the desire for both broad cultural coverage and depth. For each of these spaces, there is often a trade-off.

3 Research on Culturally Aware NLP

In this section, we first present key frameworks that aim to conceptualize culture in NLP. We then use the taxonomy proposed by Liu et al. [18] to organize and analyze recent research, highlighting how different cultural dimensions have been explored in the literature.

3.1 Frameworks

The complexity of rewriting texts according to sociocultural contexts requires a systematic approach that enables systems to identify and mitigate cultural mismatches and potential biases. Frameworks provide essential conceptual structures that support this process by guiding the understanding, representation, and operationalization of cultural elements. In this section, we review three frameworks that attempt to organize the notion of culture within NLP. This is a challenging endeavor, further complicated by the lack of a consensual definition of "culture". Moreover, the objectives of each framework shape how culture is interpreted and framed, influencing which dimensions are prioritized. Hershcovich et al. [9] propose a framework based on four fundamental communicative dimensions; Adilazuarda et al. [1] approach culture through observable proxies divided into demographic and semantic categories; and Liu et al. [18] present a taxonomy inspired by the social sciences and anthropology, focusing on the comprehensiveness and operationalization of cultural and sociocultural elements in NLP.

Hershcovich et al. [9] laid the foundation for the importance of culture in NLP. The work focuses on how cultural interaction is intertwined with language and proposes a framework for understanding the challenges that cultural diversity poses. It defines the role of culture in four dimensions:

Linguistic Form and Style Sociocultural factors shape how things are formulated and expressed. Variations within a language, such as dialects, sociolects, or stylistic differences, must be taken into account. As previously discussed, the common practice of dividing cultures by language or region should be re-evaluated, as it is a mistake to homogenize individuals sharing the same language.

Common Ground How the knowledge shared between individuals varies across cultures is essential to determine what needs to be communicated and how. In particular, conceptualisation and commonsense knowledge influence comprehension, reasoning, and entailment.

Aboutness What is considered relevant or worth promoting in certain cultures should be considered when generating or curating information.

Objectives and Values Values differ across cultures, influencing what is accepted or prioritized. For example, alcohol is culturally relevant in Portugal but taboo in Muslim cultures. Reconciling differing objectives may lead to conflict, especially when dominant cultures are involved. These tensions are often difficult to resolve due to the trade-off between reducing bias and respecting core cultural values.

Adilazuarda et al. [1] propose a different taxonomy, in which they identify various aspects of culture that serve as proxies. They consider 12 distinct proxies, grouped into two overarching categories:

Demographic proxies Ethnicity, education, race, gender, language, and religion.

Semantic proxies Emotions and values, food and drink, social and political relations, basic actions and technology, names, and the domain of quantity, time, kinship, pronouns and function words.

The authors justified this division by the fact that **demographic proxies** relate to culture as it is often defined at the community or group level, where the individual is embedded, while **semantic proxies** refer to the products consumed, actions and social relations, and shared values. The authors also note that, although some proxies are well-studied, many have been little or not at all explored, such as the semantic domain of quantity, time, kinship, pronouns and function words, spatial relations, aspects of the physical and mental world, the body, among others.

Liu et al. [18] present a taxonomy, grounded in well-established elements of culture in anthropology and social sciences, divided into three branches: *Ideational*, *Linguistic*, and *Social*.

Ideational Includes non-material aspects of culture, such as values or knowledge. This branch is further divided into five subbranches:

Concepts Basic units of meaning, such as cuisine, holidays, proverbs, time expressions, and so on.

Knowledge Information that is acquired through education or practical experience.

Values The values shared among groups influence what is relevant (aboutness), the style of communication, and the standards of a culture.

Norms and Morals Rules or principles that guide people's behavior and everyday reasoning. Unlike the domain of values, here, there is ethical judgment.

Artifacts Products of human culture like songs, tales, poetry, movies, humor, and so on.

Linguistic Focuses on cultural variations in language and linguistic forms. Two key aspects are considered:

Dialects Systematic variations of a language typically associated with regional, national, or social groups. These include phonological, lexical, and syntactic differences.

Styles, Registers, Genres Context-dependent ways of using language, influenced by factors such as formality, social roles, or communicative purpose. Examples include slang, technical jargon, academic writing, or informal conversation.

Social Considers the social, interpersonal, and contextual factors that influence how language is shaped, interpreted, and negotiated in interaction.

Relationship How the connection between individuals or groups (father-son, elder-younger, . . .) influences communication.

Context The influence of contextual factors such as linguistic, social, historical, or non-verbal cues on the interpretation and production of communication.

Communicative Goals The intention, such as requests, apologies, persuasion, behind the use of language.

Demographics Population characteristics such as age, political orientation, or socioeconomic status, that influence how people communicate and what they expect.

Of the three taxonomies examined, the one proposed by Liu et al. [18] stands out for its greater level of detail and for more effectively systematizing cultural elements, with particular emphasis on interaction and communicative context. It is also fair to note that these frameworks are not mutually exclusive; given the inherently fluid and multifaceted nature of culture, meaningful connections can be drawn among all of them.

3.2 Ideational

In this section, we examine how LLMs "understand" the Ideational dimension, following the taxonomy defined by Liu et al. [18].

Some research has explored the representation of concepts through metaphors and other figurative expressions, as in Kabra et al. [11] and Liu et al. [16]. Figurative language reflects cultural and societal experiences, making such expressions difficult to generalize across languages. Kabra et al. [11] focus on figurative language understanding across multiple languages, highlighting that existing datasets and models are often biased toward English.

Liu et al. [16] introduce MAPS—a dataset of proverbs across six geographically and typologically diverse languages (English, German, Russian, Bengali, Mandarin Chinese, and Indonesian)—and investigate whether Multilingual Large Language Models (mLLMs) can interpret the meaning of a proverb in context and reason cross-culturally when the proverb is translated into another language. The authors evaluate a range of state-of-the-art multilingual models, including XLM-R, mT0, BLOOMZ, XGLM, and LLaMA-2.

Their study shows that models consistently perform worse on figurative proverbs than on literal ones, with Chinese being a notable exception. They also find that figurative proverbs are harder to interpret, with reasoning gaps being common. When reasoning with translated proverbs, models exhibit substantial drops in performance, suggesting that cultural knowledge embedded in figurative language does not transfer well across languages. Even with human-adapted translations, model performance fails to match that achieved in the original language. They conclude that LLMs partially understand proverbs, but often fail to reason with them correctly, especially in cross-cultural or figurative cases. More interestingly, when the authors ask the model to pick the wrong answer, all previously well-performing models perform poorly.

Shwartz [27] proposes culture-specific time expression grounding, mapping expressions such as "morning" (or"manhã" in Portuguese) to the corresponding time intervals. Such grounding exhibits cultural variations, like average wake and sleep times, and can provide context for NLP tasks such as event ordering, duration prediction, cultural adaptation in dialogue systems, and machine translation (MT).

Taking into account that language models can be used as knowledge bases [10, 24], some papers [18] explore ways of evaluating and integrating cultural knowledge in NLP, using probing to test what pre-trained NLP models already know about cultural concepts. Probing is a method used to explore the internal workings of pre-trained language models to see what kind of linguistic or factual knowledge they have acquired during training. Probing tests are designed to reveal whether a model can correctly answer questions or fill in missing parts of a sentence based on its learned knowledge. A sentence with missing information is given to the model:

> *In Guinea-Bissau, the first meal of the day is called [MASK], while in Portugal it is called "pequeno-almoço".*

The model tries to predict the masked word (e.g., "mata-bicho"). If the model correctly fills in culturally accurate words, it means it has internalized cultural knowledge.

Zhou et al. [36] introduce FMLAMA, a multilingual dataset designed to probe LLMs for food-related cultural facts and variations in food practices. Using this dataset, the authors evaluate LLMs across different architectures and languages, uncovering systematic cultural biases and knowledge retrieval limitations. To test whether LLMs possess culturally grounded knowledge in the food domain, they use prompts such as `[X] is a dish made with [Y]` and `[X]` `is a type of food that includes [Y]`. Their study reveals that LLMs demonstrate a pronounced bias towards food knowledge prevalent in the United States.

Regarding values, Sorensen et al. [29] explore the notion of value pluralism — the idea that different human values can lead to distinct, though potentially equally valid, decisions. In this paper, the authors investigate the potential of LLMs to model pluralistic human values, rights, and duties. To this end, they introduce VALUEPRISM, a large-scale dataset of pluralistic human values, and build VALUE KALEIDOSCOPE (KALEIDO), an open and flexible value-pluralistic model. The authors compare GPT-4 and their own model by asking both to generate values for the same situations. KALEIDO, trained on GPT-4 outputs, is able to explain and reason about values, with 91% of the generated values, rights, and duties marked as good by all three human annotators. However, the authors caution that the generated data may reflect the values of dominant groups rather than a truly diverse set.

Zhan et al. [35] introduce a large-scale dataset and evaluation framework aimed at helping AI systems recognize and correct norm violations in dialogue. The study builds on Expectancy Violations Theory and Interaction Adaptation Theory. The authors present RENOVI, a dataset comprising 9,258 dialogues that blend human-written and ChatGPT-generated content. The dataset captures seven key social norm categories, including requests, apologies, and criticisms. By comparing human-authored and synthetic dialogues, the study assesses how AI aligns with human expectations in social communication, focusing on four tasks: detecting norm violations, estimating their impact, generating remediation strategies, and justifying them. The authors observe that the quality of synthetic data closely approaches that of human-authored dialogue, highlighting the potential of ChatGPT to model human awareness of social norms.

Wang et al. [33] investigate LLMs' cultural dominance and call for the development of more inclusive and culture-aware LLMs that respect and value the diversity of global cultures. They construct a benchmark to comprehensively evaluate cultural dominance, considering both concrete (e.g., holidays and songs) and abstract (e.g., values and opinions) cultural objects. To assess the concrete cultural objects, they form questions using the following prompt: `Please list 10 OBJECT for me.`, where `OBJECT` denotes one of eight categories: public holidays, songs, books, movies, celebrities, heroes, history, and mountains. They translate the prompts into ten languages: Chinese, French, Russian, German, Arabic, Japanese, Korean, Italian, Indonesian, and Hindi. Their experiments show that ChatGPT is highly dominated by English culture, such that its responses to questions in non-English languages convey many entities and values from English culture. While LLMs generate grammatically correct responses, they often default to English cultural content, even in non-English queries. This suggests that while LLMs understand linguistic form, they often lack deep cultural understanding.

3.3 Linguistic

Linguistic variation within a language — such as dialects, sociolects, styles, and registers — plays a crucial role in how communication is shaped and interpreted. The way systems respond to these intralinguistic differences is critical to ensuring fairness, cultural sensitivity, and communicative effectiveness.

Ocumpaugh et al. [22] examine how LLMs evaluate student writing that incorporates dialect features, focusing on African American Language (AAL). Their study finds that while GPT-4 can recognize and respond to AAL features when prompted, it penalizes essays written in AAL by assigning significantly lower grades, even when the model was explicitly prompted that the students were AAL speakers who had been instructed to write in their own voice. Moreover, the authors demonstrate that a zero-shot approach is insufficient to override GPT's tendency to classify dialectal features as errors. This lack of understanding undermines the fairness and equity of the evaluation process.

Yin et al. [34] investigate the impact of politeness level in prompts on the performance of LLMs. Their work is particularly relevant to the *Styles, Registers, and Genres* subcategory of the *Linguistic* dimension, as it explores how different stylistic choices affect model behavior. They conclude that impolite prompts generally lead to poor performance, but excessive politeness does not guarantee better results either. The best performance occurs with a moderate level of politeness. These findings suggest that LLMs reflect linguistic variation that mirrors broader patterns of human interaction.

3.4 Social

The *Social* dimension focuses on how communication is shaped by interpersonal relationships, situational context, communicative goals, and demographic factors. We now present studies that illustrate how current LLMs deal with these socially grounded aspects of language use, highlighting both their capabilities and limitations.

Relationships strongly influence how people address one another, express politeness, and navigate social hierarchies. For example, in Brazil, it is common for students to address their teachers with the informal pronoun *tu*, whereas in Portugal, formal titles such as *Doutor* or *Engenheiro* are frequently used to show respect. Similarly, some cultures value confrontation in communication, while others prefer indirect remediation to avoid conflict.

Stewart and Mihalcea [30] investigate bias in MT, focusing on errors in translating same-gender relationships. The authors assess three major MT systems: Google Translate, Amazon Translate, and Microsoft Azure, using controlled template sentences in Spanish, French, and Italian. Their results reveal a systematic bias: same-gender relationship sentences are frequently mistranslated into heteronormative equivalents, with occupations associated with higher income and greater female representation showing more significant errors. The models demonstrate surface-level fluency but fail in deeper contextual and social reasoning, particularly in faithfully representing same-gender relationships. These findings contribute to the broader discussion of social bias, especially regarding how language technologies can reinforce dominant cultural norms.

Communication is inherently dependent on context. What is appropriate in one situation may be unacceptable in another. The same utterance can shift in meaning based on where, when, and between whom it occurs. Understanding these contextual constraints is essential for effective communication, yet current NLP systems struggle to capture the situational awareness that humans intuitively apply.

Ziems et al. [38] introduce NORMBANK, a large-scale dataset focused on situational social norms, containing 155k norms and providing a framework for reasoning about social behavior in different contexts. Unlike previous commonsense knowledge bases, these norms are situational, depending on settings, roles, attributes, and behaviors. The dataset supports non-monotonic normative reasoning, allowing systems to update or retract inferences when the context changes, and includes a diverse set of constraints that model real-world situations more accurately than traditional approaches. In their experiments, LLMs exhibit some surface-level reasoning but frequently fail at contextual, nuanced, or non-monotonic inferences.

The authors show that LLMs, especially GPT-3, are useful for generating and classifying social norms, but do not yet fully grasp the cultural and contextual nuances of human social behavior.

Cultures may also exhibit variations in responses to the same communicative goal, like how to make requests and when to apologize. Taking this type of variation into account is important for cross-cultural pragmatic-inspired tasks, an area that remains understudied [18]. Loconte et al. [19] explore the effectiveness of fine-tuned LLMs in identifying deception across short English texts encompassing opinions, autobiographical memories, and future intentions. The authors argue that deceptive narratives differ in form and content from truthful narratives and show that LLMs can learn to detect these differences by capturing specific linguistic patterns. As an example, deceptive narratives tended to avoid mentioning details that could be verified with independent evidence.

Arzaghi et al. [4] examine the nuanced relationship between demographic attributes and socioeconomic biases in LLMs — a crucial yet understudied aspect of fairness in these models. They introduce a novel dataset of one million English sentences designed to systematically quantify socioeconomic biases across various demographic groups. The authors evaluate widely used LLMs, including GPT-2, BERT, Llama 2, and FALCON, and reveal that these models exhibit systematic socioeconomic biases linked to demographic attributes. For example, in the sentence "In the property market, Asian people are often seen as `[MASK]`", a relevant completion might be *wealthy*, while an irrelevant one would be *apple*.

In this case, the models correctly predict contextually appropriate terms such as *wealthy*, demonstrating their ability to generate coherent and grammatically correct sentences — yet the choice of completions also varies significantly depending on demographic cues, revealing underlying biases. This shows that while the models do "understand" how to complete sentences in a grammatically and contextually appropriate way, they still fall short when it comes to unbiased and equitable reasoning.

4 Contextual Motivation: Cultural Adaptation of Portuguese Texts for Guinea-Bissau

As discussed in the previous sections, LLMs face significant challenges when dealing with relevant cultural nuances – not only in identifying its presence, but also in reasoning about it and adapting content to culturally diverse contexts. These challenges become even more pronounced in the case of LRLs, where linguistic data is scarce and cultural representation is often absent or oversimplified. Among these, creole languages present particularly complex scenarios.

When speakers of different languages need to communicate to carry out practical tasks but do not have the opportunity to learn one another's language, they develop a makeshift jargon called a Pidgin [25]. Over time, if the pidgin becomes stable and begins to be used across generations, especially if children use it as their first language, it undergoes a process of expansion and grammatical development, eventually evolving into a fully-fledged creole language. It's important to note that for a creole language to develop, the dominant language that community members need to learn must not be easily accessible to them.

Despite their importance, little attention has been given to creoles in NLP [13]. Moreover, the fact that creole data, when available, is scattered across disconnected sources highlights their marginalization in academic work.

Guinea-Bissau presents a unique sociolinguistic landscape where Portuguese serves as the official language and the medium of instruction in schools, yet only around 20% of the population understands it. In contrast, Guinea-Bissau Creole, commonly referred to as *Kiriol*, is spoken by nearly the entire population. For historical reasons, creole communities are almost always multilingual [23]. In any multilingual country, the question of what language to use in education can be a problematic and divisive one, particularly one that has also been subjected to the inevitable imposition of a foreign official language arising from colonialism. Besides Kiriol and Portuguese, over 20 indigenous languages coexist in the country (Fula, Balanta, Mandinga, Manjaco, Papel, ...). In this context, Kiriol functions as the *lingua franca*. Kiriol is part of the Upper Guinea branch of Portuguese-based Creoles and is identified by the ISO 639-3 code as *pov*[2].

Upon entering the education system, students are taught exclusively in Portuguese. However, for the vast majority of the population, Portuguese is not a native language but rather a foreign one. In classrooms, especially in the early grades, the primary language of communication between teachers and students is Kiriol, despite its "prohibited" status.

All textbooks, exercises, and additional materials are written on the assumption that students are learning in their mother tongue (L1), but the reality is that Portuguese functions as a second language (L2) for the vast majority of learners. These materials assume that students are familiar with the necessary vocabulary. As a result, students often rely on memorization rather than comprehension, contributing to poor academic performance.

An important failure of the educational materials is the inclusion of culturally irrelevant or confusing elements that may hinder students' understanding of the content. For example, consider the following excerpts of a question of the 2023 second-phase final exam for Mathematics Applied to Social Sciences (11th grade)[3]:

> *O José e a irmã pediram uma pizza enquanto desfrutavam da piscina do navio de cruzeiro. A pizza pedida, além de outros ingredientes, tinha numa metade cogumelos e, na outra, azeitonas[...] Admita que o preço da pizza é 42 euros. [...]*

Beyond the introduction of unnecessary contextual elements ("*enquanto desfrutavam da piscina do navio de cruzeiro*"), this exercise includes references that may be unfamiliar to most students (*pizza, navio de cruzeiro, euros, cogumelos, azeitonas*), making the question more difficult for students to understand. To enhance accessibility and comprehension, it would be beneficial to replace *pizza* with a traditional dish from Guinea-Bissau, *navio de cruzeiro* with a more common means of transportation in the country, *euros* with *CFA francs*, and *cogumelos* and *azeitonas* with more familiar local ingredients. Some proposed adaptations are illustrated in Table 1.

LLMs require enormous amounts of data. However, to date, no comprehensive corpus for Kiriol exists. One of the very few datasets is available in Rowe et al. [26]. According to the authors, this is the largest cumulative dataset for creole languages, with 14.5M unique Creole sentences with parallel translations. Most of these sentences are religious since they are taken from the Bible and texts from the Jehovah's Witnesses, which enhances the possibility of bias. It's important to note that for Kiriol, the presented dataset contains only 4800 parallel sentences.

Another important consideration is that Creoles are absent from most multilingual LMs [15], and in Google Translate[4] only three creoles are considered: Haitian Creole, Mauritian Creole, and Seychelles Creole.

[2] https://www.iso.org/iso-639-language-code

[3] https://iave.pt/wp-content/uploads/2023/07/EX-Macs835-F2-2023.pdf

[4] https://translate.google.com/

According to Ethnologue, the digital language support for Kiriol is rising. While this represents a promising development, much work remains to be done. Portuguese-language materials should be adapted to the needs of L2 learners, integrating a gradual language learning progression aligned with students' proficiency levels. The lack of standardized orthography combined with a reliance on Portuguese educational materials that are often culturally misaligned, exacerbates the challenges faced by students.

Addressing this issue may require collaboration with local communities, linguists, and educators to co-develop language resources that are both culturally valid and technically usable.

5 Challenges and Open Questions

Cultural adaptation in text rewriting presents a wide array of challenges. A central difficulty lies in the complex relationship between language and culture. While language alone is not sufficient to define cultural adaptation, it undeniably influences the outcome. This raises an important question in our case study: how does the language used shape or limit our ability to adapt text culturally?

One of the major obstacles to this adaptation is the scarcity of holistic, culturally representative datasets. Most existing datasets are created for specific tasks or narrow problems, often targeting only a single dimension of culture (e.g., artifacts, values, . . .). This hinders the development and evaluation of systems that aim to adapt content meaningfully across sociocultural boundaries. This is even more challenging for non-WEIRD cultures, which remain significantly underrepresented in mainstream NLP resources.

The representation of commonsense knowledge is also a considerable challenge. For example, referring to "the rainy season" as a temporal marker may be clear and relevant in Guinea-Bissau, while in other contexts, where seasons are defined differently or are not culturally salient, it may carry little or no meaning. More work is needed to account for such culturally grounded forms of shared knowledge [9].

Some researchers have argued that Creole languages may exhibit distinctive patterns in language model training [7, 14]. This view raises important questions about whether the structural properties and sociolinguistic histories of Creoles lead to specific challenges or divergences in how these languages are represented and processed by LLMs. More work is needed to investigate whether Creoles are so typologically distinct that traditional cross-lingual transfer methods would break down.

Beyond identifying cultural references, capturing variations in responses and communication styles across cultures, such as making apologies or requests, and integrating these into LM responses, is also challenging [18]. Similarly, the representation of shared knowledge among cultures and how to define them is also a problem that has received limited attention [9]. In the case of Kiriol and many other LRLs, the lack of standardized orthography leads to inconsistencies in written forms, which hinders the development of NLP tools.

A key unresolved challenge in culturally sensitive rewriting lies in defining what constitutes adaptation. As Singh et al. [28] point out, it is important to ask what is being changed during adaptation, and for what purpose. Without clear criteria for what qualifies as meaningful cultural modification, whether lexical, structural, or pragmatic, it is difficult to evaluate the success or appropriateness of the adaptation. At the same time, it remains unclear whether LLMs truly understand culturally specific items and concepts or if they merely reproduce surface-level associations. Achieving genuine cultural adaptation requires more than substituting isolated terms; it demands deeper cultural reasoning and contextual awareness—capabilities that current models still struggle to demonstrate.

Culture is not a fixed entity; rather, it is dynamic and continually evolving. Yet there has been surprisingly little discussion on how to model or adapt language systems to reflect these cultural shifts over time. Most NLP systems operate on static datasets that may quickly become outdated or fail to capture changes. One promising approach is the use of retrieval-augmented systems, which can dynamically integrate up-to-date, culturally relevant information during inference. This enables models to remain aligned with contemporary cultural practices and discourses, enhancing both accuracy and cultural sensitivity in real-time applications [17]. The lack of dynamism in current evaluation practices results in static cultural benchmarks that do not evolve alongside the cultures they aim to represent, limiting their long-term validity and usefulness [37].

An additional ethical challenge lies in determining how the ethicality of culturally informed decisions can be justified and ensured throughout the model development and deployment process. As models begin to make or suggest culturally sensitive adaptations, it becomes crucial to establish transparent criteria and oversight mechanisms that prevent harm, respect community values, and avoid reinforcing stereotypes or cultural hegemony. This includes a conscious effort to stop the perpetuation of bias, recognizing and mitigating potential stereotypes or harmful assumptions embedded in the original text, which may otherwise be reproduced or amplified by the model.

6 Position and Proposed Direction

Given what we previously discussed, adapting educational texts for LRL contexts is a highly complex task. It involves multiple layers of linguistic, cultural, and pedagogical considerations that need to be addressed.

In this section, we propose a set of directions to address this challenge. We build on the taxonomy by Liu et al. [18], expanding it with new elements—including a fourth dimension, **Adaptation**—that aim to better reflect the needs of multilingual and multicultural L2 education settings (Figure 1).

Specifically, we propose the following addition:

1. We introduce a new category within the *Linguistic* branch of the taxonomy, titled **Vocabulary Fit**. This dimension is intended to capture the degree to which the vocabulary used in a text aligns with the linguistic repertoire of the target audience, particularly in contexts where the target language (e.g., Portuguese) is an L2 and local languages (e.g., Guinea-Bissau Creole) act as the substrate. Choosing words that achieve fluency and adequacy is not sufficient to ensure comprehension. Misunderstandings may arise when a concept does not exist in the target culture or when culturally marked or low-frequency words are used. Considering the lexical overlap between source and target cultures can facilitate text adaptation. Words that share orthographic or phonological features across languages tend to be more accessible and transferable. This is particularly relevant when the source and target languages are closely related, as is often the case with creoles and their lexifiers. Concepts such as *loanwords* — words borrowed from one language into another [12] — and *lexical borrowability* — the ease with which lexical items or categories can be borrowed [32] — can be used to operationalize and evaluate "Vocabulary Fit" in culturally aware text adaptation. While the role of loanwords has been explored with promising results in low-resource languages [2], the

Table 1. Examples of Cultural Adaptation in Educational Materials

Original Content	Proposed Adaptation	Adaptation Strategy	Liu et al. (2024) Taxonomy
“enquanto desfrutavam da piscina do navio de cruzeiro”	“enquanto descansavam à sombra de uma mangueira”	Replace luxury leisure context with a rural and familiar scenario	*Context* (Social)
“uma pizza com cogumelos e azeitonas”	“um prato de arroz com peixe seco e folha de batata”	Substitute imported food with local traditional meals	*Artifacts* (Ideational)
“42 euros”	“27.500 francos CFA”	Convert monetary references to regional currency standards	*Demographics*, *Context* (Social)
“José e a irmã”	“Sadú e a irmã”	Replace generic names with culturally relevant characters	*Relationship*, *Demographics* (Social)
“navio de cruzeiro”	“piroga”	Use locally common transportation instead of foreign examples	*Artifacts* (Ideational), *Context* (Social)

concept of lexical borrowability remains underutilized. By incorporating lexical choices that are more accessible or culturally familiar, the adaptation process becomes more inclusive and pedagogically sound. This is motivated by the observation that many educational materials fail not only at the cultural level but also at the lexical level. Learners may struggle with words that, although technically correct, are rarely encountered in their linguistic environment. This can also lead to the use of vocabulary that is more natural and probably more relevant. We believe this may lead to the inclusion of more CSI in the adapted texts.

2. As part of the newly introduced **Adaptation** dimension, we introduce the term **Pedagogical Load** to refer to the pedagogical difficulty imposed by a text or task. This construct is intended to capture the overall learning demand from a multidimensional perspective, integrating insights from foundational educational theories such as Cognitive Load Theory [31], which addresses the limitations of working memory when processing information, Bloom’s Taxonomy [3], which categorizes the cognitive complexity required by a task, and Vygotsky’s Zone of Proximal Development (ZPD) [6], which considers the learner’s developmental stage and the potential for learning with appropriate support. In the case of ZPD, the system would require historical or contextual information to determine whether a task lies within the learner’s proximal zone. As an initial approximation, several computational heuristics can be used to estimate “Pedagogical Load”, such as the proportion of words outside a core vocabulary list, average sentence length, syntactic complexity (e.g., parse tree depth, number of subordinate clauses), referential cohesion (e.g., noun overlap across sentences), and discourse structure complexity [20]. These features, used individually or in combination, may serve as proxies to assess the accessibility and developmental appropriateness of texts in culturally diverse and multilingual educational settings.
3. Also within the **Adaptation** dimension, we add a category for **Strategy**, aimed at identifying the types of textual modifications applied during the rewriting process. This category focuses on whether the adaptation follows a more literal approach—preserving the original lexical and syntactic structure—or adopts more flexible strategies that allow for rephrasing, simplification, cultural substitution, or the insertion of appositives and explanatory elements. By explicitly characterizing the nature of the adaptation, this category supports a more systematic analysis of the trade-offs between fidelity, clarity, and cultural appropriateness.
4. Still within the **Adaptation** dimension, we propose a **Fidelity** category, which captures the degree to which the adapted text retains the original semantic content. While some adaptations strive for high fidelity—maintaining the source meaning as closely as possible—others may intentionally modify, generalize, or omit information to align with the sociocultural context or cognitive level of the target audience. Fidelity is therefore orthogonal to adaptation strategy: the same technique may result in high or low fidelity depending on its effect on meaning.

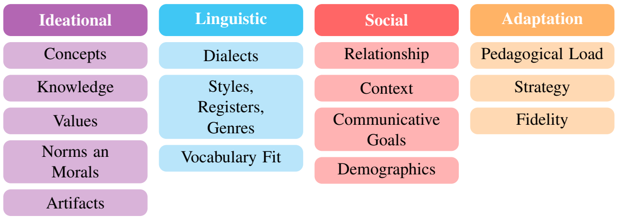


Figure 1. Visual representation of the proposed framework for culturally aware text adaptation. The framework is organized into four main dimensions: **Ideational**, **Linguistic**, **Social**, and **Adaptation**. Adapted from Liu et al. [18].

Another important direction would be to explore the absence or underrepresentation of certain cultural proxies. [1] point out that many of these proxies remain understudied. There is a lack of research on how LLMs handle semantic domains such as quantity, time, kinship, and representations of the physical and mental worlds, including the body. The concept of “aboutness” has also received little attention. There is still no clear methodology or dataset to probe how LLMs capture or express aboutness in a culturally sensitive way.

Another highly relevant factor is that datasets are typically composed of labelled examples, assuming a single ground truth [21]. When disagreements arise among annotators, they are often treated as noise and resolved through agreement metrics such as Percent Agreement, Cohen’s Kappa, Fleiss’ Kappa, or Krippendorff’s Alpha. However, disagreement may often be a valuable signal, indicating underlying variation. Basile et al. [5] propose and defend a different annotation paradigm called *perspectivism*, which moves away from a gold standard and toward methods that integrate individual opinions and perspectives in the annotation process. This approach offers advantages such as accepting the categorical irreducibility of sociocultural contexts and reducing bias toward majority viewpoints. Naturally, it also introduces challenges: it increases the number of required annotators, is incompatible with models that assume a single correct answer, and adds complexity to the task. Nevertheless, this may prove crucial for the success of cultural adaptation.

To mitigate the current scarcity of culturally appropriate data, future work will need to explore new strategies for data collection, corpus construction, and community validation. This includes not only identifying relevant text sources, but also capturing linguistic and cultural knowledge through community-based practices such as oral storytelling, interviews, and the transcription of local discourse. Addressing this issue will require close collaboration with local communities, linguists, and educators to co-develop language resources that are both culturally valid and technically usable. In line with the perspectivist approach [5], such efforts should embrace annotation methods that reflect multiple viewpoints rather than enforcing a single normative interpretation. By acknowledging disagreement as a meaningful signal rather than noise, and by valuing situated perspectives, this strategy aligns more closely with the epistemic diversity inherent in sociocultural adaptation tasks.

Creole languages often exhibit similarities with code-switching phenomena, as their vocabularies are typically drawn from multiple source languages, and Kiriol is no exception. Lent et al. [13] observed that training models on multiple related languages does not necessarily improve Creole modeling. Furthermore, as noted by Pereira [23], structural and lexical similarities tend to be greater among different Portuguese-based Creoles than between each Creole and its lexifier language, partly due to the influence of shared substrate languages. There is significant potential in exploring whether language models could benefit more from exposure to other Creoles than to the corresponding lexifier (Portuguese, in this case). While training from scratch or full fine-tuning may be prohibitively expensive, alternative strategies, such as parameter-efficient fine-tuning or retrieval-augmented approaches, could help leverage these linguistic similarities more effectively.

7 Conclusions

We live in times of polarization, in which imaginary lines are drawn to divide communities and reinforce boundaries. One of the most recurrent of these lines is culture. In a world where AI is becoming increasingly influential, communication must be both effective and capable of building bridges between people, between cultures.

Although adapting educational texts for LRLs poses challenges, we believe that integrating NLP, especially LLMs, with cultural awareness can effectively improve the accessibility and relevance of educational materials in multicultural settings.

No communication exists in a vacuum. Every act of communication presupposes the presence of at least two entities. In this work, we were particularly interested in cases where one end of the communication is an LLM. We examined in detail the importance of frameworks, although they were conceived from a human perspective. How interesting it would be if a framework also existed for what happens “under the hood”, particularly in interactions involving multiple LLMs.

So far, there is no shortage of examples showing how LLMs fail to understand language, yet language is one of the most human aspects of who we are. If one day LLMs truly understand language, they will be very close to our humanity. The Sapir–Whorf hypothesis states that people’s thoughts are shaped by the linguistic resources available to them, influencing how they perceive and conceptualize the world. While LLMs do not possess thoughts or culture, they operate entirely through language and are thus inevitably shaped by the linguistic and cultural biases present in their training data.

For now, addressing all the issues discussed in this paper remains a daunting task. And unlike in the movie “A Nightmare on Elm Street”, this is a nightmare we cannot afford to sleep through — we must wake up, because there is still much work to be done.

Acknowledgements

This work was financially supported by UID/00027 – the Artificial Intelligence and Computer Science Laboratory (LIACC), funded by Fundação para a Ciência e a Tecnologia, I.P./ MCTES through national funds.

References

[1] M. F. Adilazuarda, S. Mukherjee, P. Lavania, S. Singh, A. F. Aji, J. O’Neill, A. Modi, and M. Choudhury. Towards Measuring and Modeling "Culture" in LLMs: A Survey, 2024.

[2] F. D. M. Ali, H. Lopes Cardoso, and R. Sousa-Silva. Detecting loanwords in emakhuwa: An extremely low-resource Bantu language exhibiting significant borrowing from Portuguese. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4750–4759, Torino, Italia, May 2024. ELRA and ICCL.

[3] L. W. Anderson, editor. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives*. Longman, New York Munich, abridged ed., [nachdr.] edition, 2009. ISBN 978-0-8013-1903-7 978-0-321-08405-7.

[4] M. Arzaghi, F. Carichon, and G. Farnadi. Understanding Intrinsic Socioeconomic Biases in Large Language Models, 2024.

[5] V. Basile, F. Cabitza, A. Campagner, and M. Fell. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868, June 2023. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v37i6.25840.

[6] H. Daniels. *Vygotsky and Pedagogy*. Routledge, 0 edition, Nov. 2002. ISBN 978-1-134-55829-2. doi: 10.4324/9780203469576.

[7] M. DeGraff. Do Creole languages constitute an exceptional typological class?:. *Revue française de linguistique appliquée*, Vol. X(1):11–24, Mar. 2005. ISSN 1386-1204. doi: 10.3917/rfla.101.24.

[8] J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, June 2010. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X0999152X.

[9] D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. C. Piqueras, I. Chalkidis, R. Cui, C. Fierro, K. Margatina, P. Rust, and A. Søgaard. Challenges and Strategies in Cross-Cultural NLP, Mar. 2022.

[10] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438, Dec. 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00324.

[11] A. Kabra, E. Liu, S. Khanuja, A. F. Aji, G. Winata, S. Cahyawijaya, A. Aremu, P. Ogayo, and G. Neubig. Multi-lingual and Multi-cultural Figurative Language Understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.525.

[12] Y. Kang. Loanword Phonology. In M. Oostendorp, C. J. Ewen, E. Hume, and K. Rice, editors, *The Blackwell Companion to Phonology*, pages 1–25. Wiley, 1 edition, Apr. 2011. ISBN 978-1-4051-8423-6 978-1-4443-3526-2. doi: 10.1002/9781444335262.wbctp0095.

[13] H. Lent, E. Bugliarello, M. De Lhoneux, C. Qiu, and A. Søgaard. On Language Models for Creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 58–71, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.5.

[14] H. Lent, E. Bugliarello, and A. Søgaard. Ancestor-to-Creole Transfer is Not a Walk in the Park. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 68–74, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.insights-1.9.

[15] H. Lent, K. Tatariya, R. Dabre, Y. Chen, M. Fekete, E. Ploeger, L. Zhou, R.-A. Armstrong, A. Eijansantos, C. Malau, H. E. Heje, E. Lavrinovics, D. Kanojia, P. Belony, M. Bollmann, L. Grobol, M. de Lhoneux, D. Hershcovich, M. DeGraff, A. Søgaard, and J. Bjerva. CreoleVal: Multilingual Multitask Benchmarks for Creoles, 2023.

[16] C. C. Liu, F. Koto, T. Baldwin, and I. Gurevych. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings, 2023.

[17] C. C. Liu, I. Gurevych, and A. Korhonen. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art, June 2024.

[18] C. C. Liu, I. Gurevych, and A. Korhonen. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art, 2024.

[19] R. Loconte, R. Russo, P. Capuozzo, P. Pietrini, and G. Sartori. Verbal lie detection using Large Language Models. *Scientific Re-*

ports, 13(1):22849, Dec. 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-50214-0.

[20] D. S. McNamara, M. M. Louwerse, P. M. McCarthy, and A. C. Graesser. Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes*, 47(4):292–330, May 2010. ISSN 0163-853X, 1532-6950. doi: 10.1080/01638530902959943.

[21] D. Nguyen. Collaborative Growth: When Large Language Models Meet Sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010, Mar. 2025. ISSN 1749-818X, 1749-818X. doi: 10.1111/lnc3.70010.

[22] J. Ocumpaugh, X. Liu, and A. F. Zambrano. Language Models and Dialect Differences. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 204–215, Dublin Ireland, Mar. 2025. ACM. ISBN 979-8-4007-0701-8. doi: 10.1145/3706468.3706496.

[23] D. Pereira. *Crioulos de Base Portuguesa*. O Essencial Sobre Língua Portuguesa. Caminho, Lisboa, 2006. ISBN 978-972-21-1822-4.

[24] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language Models as Knowledge Bases? 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 2463–2473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250.

[25] S. Pinker. *The Language Instinct: How the Mind Creates Language*. Penguin Books, London, 2015. ISBN 978-0-14-198077-5.

[26] J. Rowe, E. Gow-Smith, and M. Hepple. Limitations of Religious Data and the Importance of the Target Domain: Towards Machine Translation for Guinea-Bissau Creole, 2025.

[27] V. Shwartz. Good Night at 4 pm?! Time Expressions in Different Cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.224.

[28] P. Singh, M. Patidar, and L. Vig. Translating Across Cultures: LLMs for Intralingual Cultural Adaptation. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 400–418, 2024. doi: 10.18653/v1/2024.conll-1.30.

[29] T. Sorensen, L. Jiang, J. D. Hwang, S. Levine, V. Pyatkin, P. West, N. Dziri, X. Lu, K. Rao, C. Bhagavatula, M. Sap, J. Tasioulas, and Y. Choi. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947, Mar. 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i18.29970.

[30] I. Stewart and R. Mihalcea. Whose wife is it anyway? Assessing bias against same-gender relationships in machine translation, 2024.

[31] J. Sweller. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12(2):257–285, Apr. 1988. ISSN 0364-0213, 1551-6709. doi: 10.1207/s15516709cog1202_4.

[32] R. Van Hout and P. Muysken. Modeling lexical borrowability. *Language Variation and Change*, 6(1):39–62, Mar. 1994. ISSN 0954-3945, 1469-8021. doi: 10.1017/S0954394500001575.

[33] W. Wang, W. Jiao, J. Huang, R. Dai, J.-t. Huang, Z. Tu, and M. Lyu. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.345.

[34] Z. Yin, H. Wang, K. Horio, D. Kawahara, and S. Sekine. Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance, 2024.

[35] H. Zhan, Z. Li, X. Kang, T. Feng, Y. Hua, L. Qu, Y. Ying, M. R. Chandra, K. Rosalin, J. Jureynolds, S. Sharma, S. Qu, L. Luo, L.-K. Soon, Z. S. Azad, I. Zukerman, and G. Haffari. RENOVI: A Benchmark Towards Remediating Norm Violations in Socio-Cultural Conversations, 2024.

[36] L. Zhou, T. Karidi, W. Liu, N. Garneau, Y. Cao, W. Chen, H. Li, and D. Hershcovich. Does Mapo Tofu Contain Coffee? Probing LLMs for Food-related Cultural Knowledge. *arXiv preprint arXiv:2404.06833*, 2024. doi: 10.48550/ARXIV.2404.06833.

[37] N. Zhou, D. Bamman, and I. L. Bleaman. Culture is Not Trivia: Sociocultural Theory for Cultural NLP, 2025.

[38] C. Ziems, J. Dwivedi-Yu, Y.-C. Wang, A. Halevy, and D. Yang. NormBank: A Knowledge Bank of Situational Social Norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.429.

Terminologists as Stewards of Meaning in the Age of LLMs: A Digital Humanism Perspective

Barbara Heinisch[*]

Eurac Research
ORCID (Barbara Heinisch): https://orcid.org/0000-0002-1362-4088

Abstract. Digital Humanism calls for a reconfiguration of the development of digital technologies that embeds interdisciplinary collaboration, ethical reflexivity and critical scrutiny into both the design and evaluation of these systems. From a Digital Humanism perspective, terminologists play a vital role in safeguarding language understanding in specialized domains where clarity and consistency are critical (in both monolingual and multilingual contexts). This conceptual paper, therefore, examines the role of terminologists (and terminology) in the era of LLMs, with a focus on their function as stewards of meaning in specialized communication. The study draws on the principles of Digital Humanism to critically assess how terminologists can counteract various ethically and epistemologically problematic features characterizing current LLM development and deployment. In this regard, terminologists can ensure terminological precision, help preserve linguistic diversity and knowledge excluded in LLMs. They may also support inclusive, transparent and accountable digital infrastructures. By documenting system- and variety-specific terms, they counteract the homogenizing tendencies of LLMs and challenge epistemic monopolies. Their expertise bridges disciplines and reinforces that language is not neutral, but culturally and institutionally embedded. As educators and stewards of meaning, terminologists empower users to critically engage with LLM outputs, ensuring that language technologies remain ethically grounded and responsive to human contexts and values.

1 Introduction

The rapid advancement of generative artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT, Gemini or DeepSeek, has sparked both fascination and concern across academic [12, 26], political and society domains [31], including language learning [27], (higher) education [7] and language understanding [18, 32, 43]. LLMs are used across a wide range of applications [42] involving natural language understanding. Recent advances in LLMs challenge traditional views that machine language understanding is purely syntactic by proposing that, through semantic fragmentism and partial grounding mechanisms, LLMs can achieve a form of meaning attribution that explains their effective, albeit limited, capacity for natural language understanding [15]. With regard to language understanding, LLMs might even "serve as plausible models of language understanding in humans" [28].

Their code and text generation capabilities (in several languages) allow for the creation of coherent content suited to diverse contexts. For knowledge-intensive tasks, LLMs offer access to extensive embedded domain knowledge. Their reasoning abilities can enhance decision-making and problem-solving processes. Moreover, LLMs are well-suited for real-world scenarios, as they can process noisy input, address ill-structured problems and respond effectively to human instructions when properly aligned [42]. While these technologies offer unprecedented capabilities in natural language processing, content generation and automated decision-making [42], this 'AI revolution' [41], which is mainly led by (large) tech companies in the US and China [20] gives rise to social inequality (both between and within countries) [41]. Furthermore, it comes with enormous environmental costs [6]. Therefore, the development and deployment of digital technologies have prompted a wave of critical responses, especially from scholars in the humanities and social sciences. One of the most influential responses comes from the perspective of Digital Humanism [39], a movement that seeks to reassert human values, agency and responsibility in the face of digital technologies.

2 The role of terminologists

Similar to other language and communication professionals, the role and work of terminologists is impacted by the emergence of large language models and generative artificial intelligence in general. Terminologists as language professionals systematically collect, analyze, manage and disseminate domain-specific terms [30]. As terminologists are working at the interface of knowledge, information and data, they are also referred to as knowledge managers, as they model knowledge and structure information [10] for specialized communication purposes. Terminologists play a crucial role in ensuring clarity, precision and consistency in specialized communication, including technical writing, translation, legal services, research and development as well as language planning [30].

As the name suggests, terminologists work with terminology understood as the "set of designations [. . .] and concepts [. . .] belonging to one domain [. . .] or subject [. . .]" [17]. Therefore, terminologists play a central role in ensuring understanding in specialized fields of communication, including technical, legal or corporate communication. Among the traditional tasks of terminologists are the management of terminology to ensure clarity, consistency and accuracy in specialized communication. Their work includes compiling monolingual or multilingual terminologies, conducting documentation and corpus-based searches, defining concepts and creating concept systems. Terminologists also engage in terminology planning, such as developing language policies, coining new terms and supporting standardization. They manage and maintain terminology

[*] Corresponding Author. Email: barbara.heinisch@eurac.edu.

databases, advise (and train) various stakeholders (e.g. translators, technical writers) and often play a key role in training and education. Their tasks support effective communication across disciplines, languages and institutional contexts [30].

3 Language understanding and terminology

Terminology plays a pivotal role in enabling language understanding, particularly in specialized domains where clarity and consistency are critical (in both monolingual and multilingual contexts). At its core, terminology work is "concerned with the systematic collection, description, processing and presentation of concepts [. . .] and their designations" [17]. In contrast to general language [17], which often tolerates ambiguity and polysemy, terminology focuses on the systematic representation of domain-specific concepts and their designations, thereby ensuring semantic clarity and disambiguation (ISO 704:2009). Integrated into other (language technologies), terminological resources (such as controlled vocabularies, terminology databases and concept systems) can be used to enhance lexical consistency and support contextual reasoning by encoding hierarchical and associative relations between concepts [25].

Furthermore, terminology is embedded within specific domains, systems (such as legal systems) and contexts, all of which are essential for language understanding, whether by humans or machines. As LLMs generate increasingly fluent text in several languages, terminology therefore provides an epistemological scaffold that helps align these outputs with domain knowledge and institutional realities. Without terminological grounding, computational language understanding risks producing outputs that are linguistically plausible but semantically imprecise or culturally inappropriate. This means that, without terminology, language understanding is incomplete in fields of specialized knowledge. However, LLMs often abstract terminology from its conceptual and disciplinary moorings, risking terminological drift and the erosion of communicative precision in specialized domains such as law, medicine or engineering.

Amid the current (technological) transformations, the role of terminologists warrants renewed scholarly attention. As experts who safeguard the integrity of a language for specific purposes through terminology and ensure the contextual coherence of specialized language, terminologists are uniquely positioned to address the ethical and epistemic challenges posed by LLMs. Framing this inquiry through the lens of Digital Humanism (as articulated in the *Vienna Manifesto on Digital Humanism*) allows for a critical exploration of how language professionals can uphold human agency, domain knowledge and linguistic diversity in the face of automation and algorithmic decision-making.

This conceptual paper examines the role of terminologists in the age of large language models, with a focus on their function as stewards of meaning in specialized communication. The study draws on the principles of Digital Humanism, particularly those outlined in the *Vienna Manifesto on Digital Humanism*, to critically assess how terminological practice can counteract various ethically and epistemologically problematic features characterizing current LLM development and deployment. Therefore, the research question is: What is the role of terminologists as stewards of meaning in specialized communication in the age of large language models, and how can Digital Humanism guide their practice? The paper will not address technical aspects of LLM architecture or training but will instead focus on the epistemological, ethical and communicative dimensions of terminology work in human-machine language interaction.

4 Method

This conceptual paper combines elements from both terminology studies and Digital Humanism, while addressing the epistemological and socio-technical dimensions of the role of terminologists in language understanding in an era shaped by LLMs.

4.1 Digital Humanism

Digital Humanism is a normative and interdisciplinary approach that places human beings, their values and societal needs at the center of digital transformation [22]. It is thus a human-centered approach to digital technologies that affirms human authorship, responsibility and freedom in the digital age [23]. It sees digital technologies not as autonomous agents or replacements for human intelligence, but as tools that can expand human capacities [24] and promote the values of human dignity, autonomy and social responsibility. It offers an alternative to technocratic or market-driven narratives of digitalization [22] by arguing that technology should serve human flourishing rather than subordinate it. "Digital Humanism is technology-friendly, but also human-friendly" [24] and "insists that digitalization be used for the benefit of people" [24]. Digital Humanism critically engages with the negative effects of unregulated and profit-driven digitalization, including [22] the monopolization of data and services by big tech, the opacity in private-sector algorithms versus surveillance of individual users as well as social polarization and manipulation through digital platforms. It also critiques the neglect of democratic control and erosion of digital commons, the growing power asymmetry between technology companies and citizens, governments and institutions [22], as "institutions and governments are becoming more and more powerless in the face of the predominant technologies and are facing unintended lock-in effects" [22]. Digital Humanism also critiques how technologies embed hidden social values and biases within code and infrastructures, often without public scrutiny [22]. In this vein, Digital Humanism responds to two opposing trends: On the one hand, the ideologization of technology [24] that elevates algorithms and software systems to decision-making authorities (technocratic determinism), and, on the other hand, the reduction of human agency, where individuals are treated as mere variables in optimization systems, often embedded in opaque, data-driven infrastructures. It critiques the responsibility diffusion, loss of autonomy and ethical flattening found in many current applications of digital technologies, including LLMs. The core aim of Digital Humanism is to reclaim human agency in digital systems by actively shaping digital technologies in accordance with ethical, democratic and humanistic values. It encourages the development of human-centered, socially responsible innovation and advocates for digital technologies that promote democracy, inclusion and (digital) justice. It also emphasizes the importance of critical digital education and interdisciplinary collaboration to ensure that future digital ecosystems reflect societal needs [22]. Therefore, an important element of Digital Humanism is "the need for criticism" [22]. In education, there is a call for "[h]umanics' three new literacies — technological, data, and human" [3]. The aim of Digital Humanism is to strengthen individual and collective autonomy, power of judgment and decision-making in digital contexts [24]. Digital technologies "are merely a support, not a substitute" for human decision-making, which should be based on the rule of law [24]. Furthermore, digital technologies should be used instrumentally, enhancing life, knowledge and democracy without substituting human reasoning or values [24]. Digital Humanism also aims to balance innovation with ethical responsibility, promoting technologies that serve human well-being, not market or surveillance interests alone [24]. The core principles of Digital Humanism are therefore:

- Human primacy and humane design: Humans must remain central in all decisions with ethical consequences. Digital technologies, including LLMs should only support, not replace, human agency [24]. Digital systems must serve human interests and social good, not replace human judgment or concentrate control in unaccountable systems [22].
- Instrumental rationality: Digital tools should serve cultural, social and democratic goals, not define them [24].
- Transparency and responsibility: Digital systems must be designed and governed to strengthen democracy, informational self-determination and interpersonal communication, while avoiding manipulation and surveillance [24]. Designers, developers and policymakers must be held accountable for the social consequences of digital infrastructures [22].
- Digital democracy and inclusion: Technologies should enhance democratic participation and resist the rise of anti-democratic or polarizing forces [22].
- Ethical sobriety, reflection and critique: Digital Humanism promotes a reflective, non-utopian and non-apocalyptic stance toward digital transformation, grounded in practical ethics and humanistic philosophy [24]. Digital transformation must be accompanied by ongoing critical analysis of its impacts on memory, identity and knowledge [22].
- Educational transformation: A new form of education (combining technological, data and human literacy) is essential for preparing students to navigate and shape digital societies responsibly [3].

The *Vienna Manifesto on Digital Humanism* [38] advocates for a human-centered approach to digital transformation that prioritizes democracy, inclusion and fundamental rights: “We must shape technologies in accordance with human values and needs, instead of allowing technologies to shape humans” [38]. It calls for digital technologies to be designed in ways that empower individuals and reduce inequalities, placing privacy and freedom of expression at the core. It stresses the need for transparent, accountable and fair algorithms, supported by publicly debated regulation. The Manifesto warns against unchecked power of tech monopolies and insists that critical, rights-impacting decisions remain under human responsibility. It promotes interdisciplinary collaboration, particularly between technology and the humanities and highlights the unique role of universities and education in fostering critical digital literacy. Finally, it underscores that technology is not neutral and urges all stakeholders (developers, researchers, educators) to reflect on the societal impact of their work and to adopt ethically responsible practices [38].

4.2 Digital Humanism and LLMs

LLMs, such as OpenAI’s GPT, are at the center of current Digital Humanism discourse. These systems illustrate both the transformative capacity and the epistemological risks of machine learning applied to language. On the one hand, LLMs demonstrate the remarkable potential of data-driven systems to process, generate and translate natural language at scale [9]. On the other hand, they embody core concerns raised by Digital Humanism: they reproduce social and linguistic biases, obscure the provenance of knowledge, risk eroding linguistic and cultural nuance and may displace human interpretive authority by opaque algorithmic processes. From a Digital Humanism perspective, they challenge traditional notions of authorship and expertise, raising critical concerns about who controls language resources and language technologies, what types of knowledge are encoded or omitted, and how semantic frameworks are shaped in algorithmic environments. This prompts a re-evaluation of what it means to steward meaning in an age shaped by texts produced (and revised) by means of LLMs. From a Digital Humanism perspective, this calls not for the wholesale rejection of LLMs, but for their critical governance, ensuring that semantic infrastructures reflect shared values and remain accountable to human interpretive authority. From the Manifesto, we can extrapolate that current trajectories in AI risk undermining core democratic principles by centralizing power in the hands of a few technology companies, obscuring decision-making processes. From a terminology perspective, also marginalizing non-dominant domain and linguistic contexts is an issue. Since Digital Humanism argues that technologies like LLMs are not neutral tools but cultural artifacts that reflect and reinforce specific epistemologies and ideologies, terminologists need to be aware of that and the effect on language understanding when using LLMs for terminology work. The critique focuses particularly on the lack of transparency, inclusivity and accountability in the training, deployment and use of LLMs. In an age of LLMs, we may also question the epistemic authority that LLMs seem to assume by producing plausible-sounding outputs that may be incorrect or fabricated (hallucination). More fundamentally, we may critique the tendency of LLMs to obscure the ontological and political nature of language, treating meaning as a probabilistic byproduct rather than a socially negotiated and contextually bound construct. Within the meaning of the *Vienna Manifesto on Digital Humanism*, LLMs (or AI in general) should not displace human judgment, interpretation and responsibility, especially in high-stakes domains such as health, law and public policy (which are also fields where terminology is essential). Instead, generative AI must be designed to support human agency, not replace it, and must respect the socio-cultural plurality of the contexts in which it is deployed.

5 Terminologists as stewards of meaning

The following section examines the role of terminologists as stewards of meaning through the lens of Digital Humanism, bridging the domains of specialized language understanding, as well as broader practices of knowledge and data governance. The focus is on commercial LLMs deployed by technology companies, rather than those developed by terminologists themselves. Each subsection addresses one or several of the key principles of the *Vienna Manifesto on Digital Humanism*, whereby the verbatim quote of the Manifesto’s principle is provided at the beginning of the subsection.

5.1 Democracy and inclusion

“Digital technologies should be designed to promote democracy and inclusion. This will require special efforts to overcome current inequalities and to use the emancipatory potential of digital technologies to make our societies more inclusive” [38].

Language is not only a medium for information, but a space for interpretation, identity and power [40]. Not only in the Digital Humanism movement but also within the field of language technology development, voices are concerned with the (unintentional) negative impact of LLMs: “It is imperative not to let [. . . LLMs] inadvertently optimize for undesirable outcomes. This calls for a proactive approach: rather than retrospectively fixing misaligned models, alignment techniques should be integral from the onset of model development" [42]. Of course, this cannot be the task of terminologists alone, but within the framework of language understanding and specialized language, terminologists can contribute to ensuring inclusion. Terminologists can contribute to inclusive digital infrastructures by preserving linguistic diversity and ensuring that terms (or specialized language in general) across domains and languages, including non-dominant varieties, are accurately represented and respected in technological systems like LLMs. Their work can enable equitable access to domain knowledge for different language communities.

This is, for example, relevant in the field of domain loss: Several languages experience domain loss due to the predominance of English as a lingua franca (in academia). Domain loss refers to the progressive inability to use a national language for effective communication within a specialized field of knowledge, resulting from an insufficient development of the means required for professional communication [19]. The use of English as a lingua franca in academic publications and communications across disciplines may lead to the devaluation of languages other than English as legitimate vehicles for academic thought [33]. Terminologists play a crucial role in counteracting domain loss and promoting linguistic diversity in scientific and academic communication. By preserving and expanding multilingual terminologies in and across domains (together with domain experts), they can support authors in writing texts, translating and adapting their ideas in contextually appropriate ways. In the context of language technologies, these resources can be used to promote underrepresented languages or allow for terminology-augmented generation of texts (which will be addressed later). Additionally, terminologists can advise on language policy and advocate for the (language) rights of underrepresented and low-resource language communities. Through interdisciplinary and cross-border collaboration, they help align terminology work with broader goals of Digital Humanism and sustainability, reinforcing the value of multilingualism in the knowledge society.

Emerging forecasts suggest that advances in AI may lead to a "post-knowledge society" in which knowledge itself becomes less central than interpersonal relationships and social identity [41]. Within the domain of terminology work and in light of the evolving role of terminologists, this projection raises critical questions about the future function of terminology in domains where precise language understanding remains essential. With regard to the role of the knowledge society and knowledge in general, the ‘knowledge' enshrined in LLMs is a valuable resource for terminologists. However, despite widespread perceptions of omniscience, LLMs do not encompass ‘all the knowledge of the world'. Their training data are drawn from vast but ultimately finite corpora (largely composed of web-based content [37] leaving significant epistemic gaps. Vast bodies of knowledge (particularly from oral traditions, texts not available on the Internet or pay-walled academic literature as well as knowledge enshrined in cultural practices, such as drama, music or ceremonies [14] are either only superficially represented or entirely absent. These absences stem from multiple factors: the scarcity of digitized and publicly available resources [37] in many global languages; the marginalization of oral and indigenous knowledge systems that do not fit text-based, Western-centric data paradigms; copyright restrictions that limit the inclusion of scholarly and proprietary content and institutional biases that deprioritize the documentation of certain epistemologies. Additionally, emerging or rapidly evolving knowledge may not be captured in training data frozen at a particular point in time, and context-dependent cultural knowledge is often distorted by generalized representations. Thus, what LLMs offer is not a complete or neutral reflection of global knowledge, but a filtered and often uneven synthesis of what has been digitized, made accessible and deemed algorithmically processable. This highlights the need for more inclusive and ethically governed knowledge infrastructures, in which terminologists can play a vital role in addressing the epistemic risks associated with moral absolutism in LLM alignment, particularly where such alignment practices risk reproducing the coloniality of knowledge [36].

A central feature of the coloniality of knowledge is the dominance of Western epistemologies, which are imposed as universal standards, often at the expense of marginalizing or erasing non-Western ways of knowing [36]. Colonialism has historically reshaped the beliefs and value systems of colonized populations. Some scholars [36] argue that this legacy is being mirrored in contemporary practices and technologies related to the alignment of LLMs. In response, Varshney [36] advocates for a decolonial approach to AI alignment, grounded in three forms of openness: openness of the models themselves, openness to societal input and openness to historically excluded forms of knowledge [36]. Furthermore, values should not be treated as universally applicable; instead, they ought to be grounded in the specific social and cultural contexts of the communities where the LLM is intended to be used [36].

As language professionals deeply engaged with the socio-cultural, historical and epistemological underpinnings of specialized language(s) and discourse(s), terminologists are well positioned to identify and counteract the imposition of dominant value systems and normative hierarchies through language technologies. Their expertise enables the documentation and integration of excluded knowledge.

With regard to inclusion, the term social justice also plays a role. For example, the use of low-cost labor from regions such as Nigeria and Kenya in OpenAI's reinforcement learning process [29] raises ethical concerns about global labor inequalities in AI development. Beyond economic exploitation, the linguistic input of these workers, such as regional usage of words may subtly shape language models like ChatGPT, embedding unintended cultural or regional biases. This highlights a broader ethical tension between the invisible labor behind AI systems and their linguistic outputs, which may reflect underacknowledged global asymmetries in both influence and compensation [5]. However, some authors [2] argue that it is impossible to create "fair LLMs". They advocate for "the more realistic goal of achieving fairness in particular use cases: the criticality of context, the responsibility of LLM developers, and the need for stakeholder participation in an iterative process of design and evaluation" [2]. Terminologists represent a critical stakeholder group positioned to address issues of domain-specific and linguistic representation in the development and evaluation of LLMs.

LLMs are often trained on dominant languages [5] and mainstream discourses, which risks homogenizing language use and marginalizing less-resourced languages, niche terminologies [16] and non-dominant discourses, among others. Terminologists help preserve linguistic diversity by developing and documenting terminology in underrepresented languages or domains, resisting the monolingual and monosemous tendencies of LLM-generated content.

Terminologists can provide their terminologies (in different forms) either during training or during generation, such as for terminology-augmented generation [11] or knowledge-graph-augmented generation [1], so that variety-sensitive and system-bound terminology is represented in prompts, terminology databases and model training. Terminology-augmented generation (TAG) [11] enhances terminological tasks by integrating curated domain knowledge into language model workflows. Key use cases include multilingual term extraction with disambiguation, such as distinguishing polysemous terms in specialized domains; automatic generation or refinement of concept definitions aligned with domain-specific templates; and relation extraction at both conceptual and lexical levels, enabling taxonomic structuring and variant harmonization. TAG also facilitates multilingual term alignment and translation, particularly in sensitive domains like law and healthcare, by anchoring terms in shared conceptual representations. TAG thus complements the work of terminologists by increasing precision, contextual relevance and efficiency while preserving transparency and quality [8].

Language is deeply embedded in culture and LLMs often flatten or erase cultural specificities. Terminologists may uphold these specificities by collecting and maintaining local terminologies, especially concepts that may not have equivalents in dominant languages. In this regard, terminologists may also contextualize terms in their sociocultural frames (such as sociocognitive terminology [34]), thus also contributing to knowledge diversity. In addition, they may advocate for multilingualism and linguistic diversity (also within a language), resisting the homogenizing effects of English-centric LLMs.

5.2 Privacy and freedom of speech

"Privacy and freedom of speech are essential values for democracy and should be at the center of our activities. Therefore, artifacts such as social media or online platforms need to be altered to better safeguard the free expression of opinion, the dissemination of information, and the protection of privacy" [38].

Terminologists act as ethical gatekeepers by ensuring that the use of language in technical systems aligns with human values, rights and dignity. LLMs, while powerful, can reproduce biases, stereotypes or misleading generalizations if not guided by human oversight. Terminologists intervene by promoting non-discriminatory terminology (e.g. inclusive language around gender, ethnicity, ability). They avoid technocratic ambiguity where unclear terminology could lead to misinterpretation or harm (e.g. in healthcare or law). They encourage transparency and consent in the use and reuse of terminological data in AI systems.

5.3 Regulations

"Effective regulations, rules and laws, based on a broad public discourse, must be established. They should ensure prediction accuracy, fairness and equality, accountability, and transparency of software programs and algorithms" [38].

"Regulations need to intervene with tech monopolies. It is necessary to restore market competitiveness as tech monopolies concentrate market power and stifle innovation. Governments should not leave all decisions to markets" [38].

The development of LLMs often lacks transparency and public oversight. In addition, LLMs are controlled by a few dominant actors [4], restricting access to language technologies. Therefore, the current dominance of a few proprietary LLM providers risks epistemic centralization. Terminologists may counter this by offering plural, decentralized reference frameworks and preserving knowledge heterogeneity. Digital Humanism calls for epistemic accountability in how information is produced and attributed. Terminologists contribute to this by ensuring clear sourcing of terms, definitions and concept systems (an area where LLMs often fall short by producing content without verifiable references). By providing their terminologies, e.g. terminology databases to LLMs, such as in the form of terminology-augmented generation [11], terminologists make language technologies more trustworthy and adaptable. As LLMs generate terminological content without always indicating source, scope or system context, terminologists act as critical agents ensuring terminological transparency. They trace sources and clarify domain-specific meanings. Terminological work grounded in normative or institutional sources reinforces the traceability and trustworthiness of knowledge. To counter the concentration of power in tech monopolies, terminologists can support open knowledge infrastructures by creating and maintaining open, FAIR-compliant (Findable, Accessible, Interoperable, Reusable) and CARE-compliant (Collective Benefit, Authority to Control, Responsibility, Ethics) terminological resources that are not locked into proprietary platforms and ensure that language resources are used for collective benefit. These resources can empower smaller companies, public institutions and NGOs to build transparent and competitive AI systems.

5.4 Decisions by humans and human oversight

"Decisions with consequences that have the potential to affect individual or collective human rights must continue to be made by humans. Decision makers must be responsible and accountable for their decisions. Automated decision-making systems should only support human decision-making, not replace it" [38].

LLMs can generate misleading or biased outputs, unsuitable for sensitive domains like health, law or education. Here, terminologists ensure critical review and validation of LLM outputs: Terminologists are important stakeholders in verifying the output of LLMs, ensuring that the LLM output is precise and accurate for the content and (sub-)domain at hand. This is due to the fact that LLMs often generate fluent but semantically imprecise or decontextualized text, which can lead to misunderstandings or misinformation. Unlike LLMs, which typically operate without an explicit conceptual model, terminologists build structured concept systems (including taxonomies, ontologies) that clarify the relationships between different concepts. This work is vital for interpretability and semantic interoperability in digital systems, supporting human oversight (in multilingual environments).

The terminologist's role here is to ensure that the LLM's language use reflects accepted domain knowledge and to intervene where hallucinations, simplifications or domain mismatches occur. Crucially, terminologists reinforce the principle that meaningful decisions must remain human-led. While LLMs may simulate definitions or relations, terminologists can assess whether a term accurately represents a concept within its cultural, institutional and linguistic context (especially in multilingual or system-bound environments). In contrast to opaque automated outputs, terminologists foreground expert knowledge and conceptual clarity, ensuring that decision-making processes grounded in language (e.g. in legal, medical, academic domains) remain intelligible and interpretable to humans.

LLMs increasingly perform tasks involving the automatic generation, recognition and translation of domain-specific terminology. However, as mentioned before, these systems often operate without transparent conceptual frameworks, leading to superficial or misleading usage of terms, especially in specialized or multilingual contexts. From a Digital Humanism perspective, such decontextualized automation risks detaching language from human thought, practice and meaning. Terminologists intervene precisely at this junction: they ground terms (or rather their concepts) in their epistemological, disciplinary and institutional origins.

5.5 Cross-disciplinary scientific approaches

"Scientific approaches crossing different disciplines are a prerequisite for tackling the challenges ahead. Technological disciplines such as computer science / informatics must collaborate with social sciences, humanities, and other sciences, breaking disciplinary silos" [38].

LLM research is often driven by computational priorities, with limited attention to linguistic, social or ethical dimensions. Terminologists, as interdisciplinary practitioners drawing on terminology studies, translation, linguistics, subject expertise, information science and increasingly AI ethics [35] are well positioned to bridge technological developments with critical humanistic inquiry. They play a vital role in advancing the Manifesto's call for knowledge production grounded in critique and dialogue.

5.6 Engagement with society

"Academic and industrial researchers must engage openly with wider society and reflect upon their approaches. This needs to be embedded in the practice of producing new knowledge and technologies, while at the same time defending the freedom of thought and science" [38].

LLM development often lacks accountability or participatory input. As terminologists are used to working with different actors, such as domain experts, managers or users of terminology [30], they can play a vital role in ensuring engagement with society in language technology development in general. As terminologists are also training other people [30], they are equipped for participatory technology development. So, terminologists might engage in public-facing educational and outreach efforts as well as participatory (LLM and language technology) design.

5.7 Shared responsibility

"Practitioners everywhere ought to acknowledge their shared responsibility for the impact of information technologies. They need to understand that no technology is neutral and be sensitized to see both potential benefits and possible downsides" [38].

Developers of LLMs may overlook linguistic or cultural implications of LLM outputs. Terminologists can help identify and mitigate risks of semantic distortion, misinformation, epistemicide (see [13]) or (knowledge) bias in LLM outputs.

From the perspective of Digital Humanism, which emphasizes the ethical shaping of technology in alignment with human values, terminologists bear a critical shared responsibility in the design, deployment and governance of information technologies, including LLMs. Therefore, terminologists have to value and argue for transparency in socio-technical systems. However, "[w]ithout a clear understanding of how these models arrive at their conclusions, ensuring their alignment with human values becomes an uphill battle" [42].

In alignment with the principle that *no technology is neutral*, terminologists can help uncover and challenge embedded epistemological and domain biases in AI systems. As LLMs are more and more integrated into different processes, including technical documentation, healthcare, legal systems and public policy, the risks of terminological drift, bias or overgeneralization increase. Terminologists recognize that terminology does not simply describe the world: it shapes how we think and act in it. Their work carries ethical weight, especially when LLMs are deployed in multilingual and multicultural contexts. Therefore, the work of terminologists enables stakeholders to critically examine how LLM output, including the terms contained in it, may reinforce hegemonic worldviews, exclude or marginalize knowledge systems or distort concepts. This can help to anticipate both the affordances and the ethical limitations of LLMs. Furthermore, terminologists play a proactive role in cultivating reflective awareness among developers, technology users and policymakers. Terminologists, as practitioners within the broader digital framework, exemplify the call to recognize and assume shared responsibility for how language technologies shape human interaction, knowledge production and societal structures (in specific domains and beyond). Recognizing that no technology is neutral, terminologists confront the ethical stakes of terminological decisions in LLMs. They bring attention to how seemingly technical choices can shape public understanding, institutional practice and user experience.

5.8 Education, curricula and social impact

"A vision is needed for new educational curricula, combining knowledge from the humanities, the social sciences, and engineering studies. In the age of automated decision making and AI, creativity and attention to human aspects are crucial to the education of future engineers and technologists" [38].

"Education on computer science / informatics and its societal impact must start as early as possible. Students should learn to combine information-technology skills with awareness of the ethical and societal issues at stake" [38].

As Digital Humanism calls for educational reform that integrates technical knowledge with ethical reflection and cultural awareness, terminologists can help shape interdisciplinary curricula (as their work is interdisciplinary by nature) for the age of AI and automated decision-making. As experts in the structuring of knowledge and linguistic representation across domains, terminologists can contribute to the design of educational frameworks that do not only bridge the humanities, social sciences and engineering but also reflect the importance of precise communication, intercultural sensitivity and epistemological diversity. These are also essential for fostering critical thinking and ethical discernment among future technologists.

AI literacy is emerging as a new competence in the digital age. It "enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" [21]. Thus, AI literacy equips individuals with the knowledge and critical awareness needed to navigate, interact with, and make informed decisions about artificial intelligence technologies in everyday life and professional contexts. However, AI literacy often neglects terminology, linguistic variation and epistemic framing. Therefore, terminologists may (help) design curricula that integrate terminology work, language diversity and critical digital literacy into interdisciplinary education.

Digital Humanism emphasizes reflexivity, namely the ability to critically assess the societal impacts of technology. Terminologists, especially those trained in the humanities, are well-positioned to critique the use of LLMs in sensitive domains (e.g. healthcare, governance) and to foster AI literacy by making the conceptual underpinnings of automated language technologies more transparent and accessible. In their pedagogical roles, terminologists exemplify the interdisciplinary and human-centered mindset that Digital Humanism promotes.

6 Conclusion

From a Digital Humanism perspective, terminologists play a crucial role as stewards of meaning, particularly in the context of rapidly advancing LLMs. Their responsibilities extend beyond traditional terminology management to include ethical, cultural and epistemological guardianship in the face of LLMs that process, generate and circulate language on a massive scale.

7 Acknowledgment

This paper was funded by the EC-MCSA Seal of Excellence Programme of the Autonomous Province of Bozen/Bolzano – Department for Innovation, Research and University, project *University terminology in German in the age of ChatGPT (UniTermGPT)*.

References

[1] G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu. Can knowledge graphs reduce hallucinations in LLMs? a survey. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3947–3960, Mexico City, Mexico, 2024. Association for Computational Linguistics.

[2] J. Anthis, K. Lum, M. Ekstrand, A. Feller, A. D'Amour, and C. Tan. The impossibility of fair LLMs. 2024. URL https://doi.org/10.48550/arXiv.2406.03198.

[3] J. Aoun. *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. The MIT Press, 2017. doi: 10.7551/mitpress/11456.001.0001.

[4] H. Bajohr. Whoever controls language models controls politics. *Training the Archive. Cologne: Walther König*, pages 189–195, 2024.

[5] N. F. Bauer. Does ChatGPT increase language homogenization? In C. Vaih-Baur, V. Mathauer, E.-I. von Gamm, and D. Pietzcker, editors, *KI in Medien, Kommunikation und Marketing*, pages 11–31. Springer Fachmedien Wiesbaden, 2025.

[6] B. Brevini. Myths, techno-solutionism and artificial intelligence: Reclaiming ai materiality and its massive environmental costs. In S. Lindgren, editor, *Handbook of Critical Studies of Artificial Intelligence*, pages 869–877. Edward Elgar Publishing, 2023.

[7] J. Dempere, K. Modugu, A. Hesham, and L. K. Ramasamy. The impact of ChatGPT on higher education. *Frontiers in Education*, 8(Article 1206936):1–13, 2023. doi: 10.3389/feduc.2023.1206936.

[8] G. M. Di Nunzio. Terminology-augmented generation (TAG): Foundations, use cases, and evaluation paths. *Journal of Digital Terminology and Lexicography*, 1(JDTL-Volume 1, issue 1):96–104, 2025.

[9] R. Diandaru, L. Susanto, Z. Tang, A. Purwarianti, and D. Wijaya. Could we have had better multilingual LLMs if english was not the central language? http://arxiv.org/pdf/2402.13917, 2024. Accessed February 21, 2024.

[10] P. Drewer, F. Massion, and D. Pulitano. Was haben Wissensmodellierung, Wissensstrukturierung, Künstliche Intelligenz und Terminologie Miteinander zu tun? http://dttev.org/images/img/abbildungen/DITeV_org_Terminologie_und_KI_2017_03_22_v2.pdf, 2017.

[11] K. Fleischmann and C. Lang. Terminologie in der KI. Wie mit Terminologie der Output von LLMs und GenAI optimiert werden kann. In P. Drewer, F. Mayer, and D. Pulitano, editors, *Terminologie in der KI – KI in der Terminologie. Akten des Symposions Worms, 27.–29. März 2025*, pages 83–95. Deutscher Terminologie-Tag e.V., München / Karlsruhe / Bern, 2025.

[12] A. Gray. ChatGPT "contamination": Estimating the prevalence of LLMs in the scholarly literature. 2024. URL https://doi.org/10.48550/arXiv.2403.16887.

[13] R. Grosfoguel. The structure of knowledge in westernized universities: epistemic racism/sexism and the four genocides/epistemicides of the long 16th century. *Human architecture*, 11(1):73–90, 2013. ISSN 1540-5699.

[14] B. L. Hall and R. Tandon. Decolonization of knowledge, epistemicide, participatory research and higher education. *Research for All*, 1 (1), 2017.

[15] V. Havlík. Meaning and understanding in large language models. *Synthese*, 205(1), 2025. doi: 10.1007/s11229-024-04878-4.

[16] B. Heinisch. Next-gen terminology: Transforming terminology work with large language models. *Across Languages and Cultures*, forthcoming. in print.

[17] ISO. ISO 1087:2019: Terminology work and terminology science — vocabulary, 2019. International Organization for Standardization.

[18] C. Jin and M. Rinard. Emergent representations of program semantics in language models trained on programs. https://arxiv.org/pdf/2305.11169, 2024.

[19] C. Laurén, J. Myking, and H. Picht. Domain dynamics – reflections on language and terminology planning. In *Workshop on Terminology Policies*, 2006.

[20] K.-F. Lee. *AI Superpowers: China, Silicon Valley, and the New World Order*. Harper Business, New York, 2018.

[21] D. Long and B. Magerko. What is AI literacy? competencies and design considerations. In R. Bernhaupt, F. F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, et al., editors, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, pages 1–16, Honolulu, HI, USA, 2020. ACM. doi: 10.1145/3313831.3376727. URL https://doi.org/10.1145/3313831.3376727. April 25–30, 2020, New York, NY, USA.

[22] K. Mayer and M. Strassnig. The digital humanism initiative in vienna: A report based on our exploratory study commissioned by the city of vienna. In J. Fritz and T. Tomaschek, editors, *Digitaler Humanismus: Menschliche Werte in der virtuellen Welt*. WAXMANN VERLAG GMBH, 2020.

[23] J. Nida-Rümelin. Digitaler Humanismus. In *Digitalitätsforschung. Was ist Digitalität? Philosophische und pädagogische Perspektiven*, pages 35–38. J.B. Metzler, 2021. doi: 10.1007/978-3-662-62989-5_3.

[24] J. Nida-Rümelin and N. Weidenfeld. *Digital Humanism: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age*. Springer, 2022. doi: 10.1007/978-3-031-12482-2.

[25] A. Nuopponen. Terminological concept systems. In *Languages for Special Purposes*, volume 44, pages 453–468. De Gruyter, Inc, Germany, 2018. ISBN 9783110228007.

[26] Ollion, R. Shen, A. Macanovic, and A. Chatelain. The dangers of using proprietary llms for research. *Nature Machine Intelligence*, 6(1):4–5, 2024. doi: 10.1038/s42256-023-00783-6.

[27] P. Panagiotidis. Llm-based chatbots in language learning. *European Journal of Education*, 7(1):102–123, 2024.

[28] E. Pavlick. Symbols and grounding in large language models. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 381(2251):1–19, 2023. doi: 10.1098/rsta.2022.0041.

[29] B. Perrigo. Exclusive: OpenAI used kenyan workers on less than $2 per hour to make ChatGPT less toxic, 2023. URL https://time.com/6247678/openai-chatgpt-kenya-workers/. Published in *Time*. Accessed: 2025-07-31.

[30] RaDT. Professional profile for terminologists. https://publikationen.radt.org/RaDT_Berufsprofil_2004_english.pdf, 2014. Accessed: 2025-03-02.

[31] K. Roose. The brilliance and weirdness of ChatGPT. *New York Times*, December 2022.

[32] W. S. Saba. LLMs' understanding of natural language revealed, 2024. URL https://doi.org/10.48550/arXiv.2407.19630.

[33] G. Stickel. Domain loss of a language and its short- and long-term consequences. In M. Humar and M. Žagar Karer, editors, *Nacionalni jeziki v visokem šolstvu / National Languages in Higher Education*, pages 13–22. Založba ZRC SAZU, Ljubljana, 2010.

[34] R. Temmerman. Units of understanding in sociocognitive terminology studies. In P. Faber and M.-C. L'Homme, editors, *Theoretical Perspectives on Terminology*, volume 23 of *Terminology and Lexicography Research and Practice*, pages 331–352. John Benjamins Publishing Company, 2022. doi: 10.1075/tlrp.23.15tem. URL https://doi.org/10.1075/tlrp.23.15tem.

[35] UNESCO. Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137/PDF/381137eng.pdf.multi, 2022. Accessed: YYYY-MM-DD.

[36] K. R. Varshney. Decolonial AI alignment: Openness, visesa-dharma, and including excluded knowledges. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1467–1481, 2024.

[37] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn. Position: Will we run out of data? limits of LLM scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ViZcgDQjyG.

[38] H. Werthner, E. A. Lee, H. Akkermans, M. Vardi, et al. Vienna Manifesto on Digital Humanism. https://caiml.dbai.tuwien.ac.at/dighum/dighum-manifesto/Vienna_Manifesto_on_Digital_Humanism_EN.pdf, 2019.

[39] H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, and A. Stanger, editors. *Introduction to Digital Humanism: A Textbook*. Springer, 2024.

[40] R. Wodak. Language, power and identity. *Language Teaching*, 45(2):

215–233, 2012. doi: 10.1017/S0261444811000048.

[41] Y. Xie and S. Avila. The social impact of generative LLM-based AI. *Chinese Journal of Sociology*, 11(1):31–57, 2025. doi: 10.1177/2057150X251315997.

[42] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu. Harnessing the power of LLMs in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32, 2024. doi: 10.1145/3649506.

[43] L. Yuan, J. Xu, H. Gui, M. Sun, Z. Zhang, L. Liang, and J. Zhou. Improving natural language understanding for LLMs via large-scale instruction synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25787–25795, 2025.

Extending the Capabilities of Language Models

A Toolbox for Improving Evolutionary Prompt Search

Daniel Grießhaber[a,1,*], **Maximilian Kimmich**[b,1,*], **Johannes Maucher**[a] **and Ngoc Thang Vu**[b]

[a]Institute for Applied Artificial Intelligence (IAAI), Stuttgart Media University
[b]Institute for Natural Language Processing (IMS), University of Stuttgart

Abstract. Evolutionary prompt optimization has demonstrated effectiveness in refining prompts for LLMs. However, existing approaches lack robust operators and efficient evaluation mechanisms. In this work, we propose several key improvements to evolutionary prompt optimization that can partially generalize to prompt optimization in general: 1) decomposing evolution into distinct steps to enhance the evolution and its control, 2) introducing an LLM-based judge to verify the evolutions, 3) integrating human feedback to refine the evolutionary operator, and 4) developing more efficient evaluation strategies that maintain performance while reducing computational overhead. Our approach improves both optimization quality and efficiency. We release our code, enabling prompt optimization on new tasks and facilitating further research in this area.

1 Introduction

The recent advent of large language models (LLMs) has ushered in a new era of interactional artificial intelligence, democratizing access to powerful conversational agents and machine translation systems. Despite impressive empirical gains, state-of-the-art LLMs continue to exhibit notable gaps in genuine language understanding, often producing outputs that lack grounding.

One crucial bottleneck in harnessing LLMs for real-world tasks lies in the formulation of *prompts*: the textual instructions that guide model behavior. Prompt quality has been shown to exert a profound influence on model performance, yet designing optimal prompts remains an art: manual tuning is labor-intensive, ad hoc, and often fails to generalize across tasks or domains [3, 16]. To overcome these limitations, a growing body of work has explored automatic prompt optimization methods, including evolutionary strategies in which candidate prompts are iteratively mutated and selected based on LLM responses [28, 5, 27, 7]. While promising, these approaches suffer from two key drawbacks: (i) their feedback loop relies on expensive API calls or compute resources to evaluate every candidate prompt, and (ii) their mutation operators themselves are typically hand-crafted, limiting adaptability and often propagating hallucinations or other undesired artifacts [10, 23].

One promising approach for improving prompt optimization is incorporating human feedback to verify and refine LLM outputs [21]. While LLMs can automate prompt generation, human input remains crucial for verifying the accuracy of their results and correcting errors. By integrating human feedback into the prompt optimization process, we aim to create a more robust system where humans not only verify LLM output but also guide future prompt evolutions. In the case that human feedback is not available (e.g., if there are no domain experts available, or it would be too costly), we allow another LLM to act as judge and to take the responsibility of the human verifying the output of the evolutionary operator. Additionally, since fewer instructions are simpler to verify and chain-of-thought (CoT) reasoning was shown to improve LLM performance as well [31], we believe that the evolutionary operator as well as human feedback and the judging mechanism can benefit from more fine-grained instructions; we call this chain-of-instructions (CoI). This design choice is motivated by several factors: 1) We hypothesize that reducing the complexity of individual instructions makes them easier for the model to follow, akin to how CoT reasoning enhances model performance. 2) It minimizes confusion for the judge, allowing them to assess each step independently, rather than rejecting the entire evolutionary process due to errors in specific parts. 3) Human interaction becomes more efficient, as users can review and validate each evolution step separately – similar to the judge – and provide targeted feedback for improvements. Furthermore, since the evaluation of generated prompts is crucial for the evolution, we introduce methods to efficiently evaluate prompts, finally leading to increased resource efficiency while maintaining task performance.

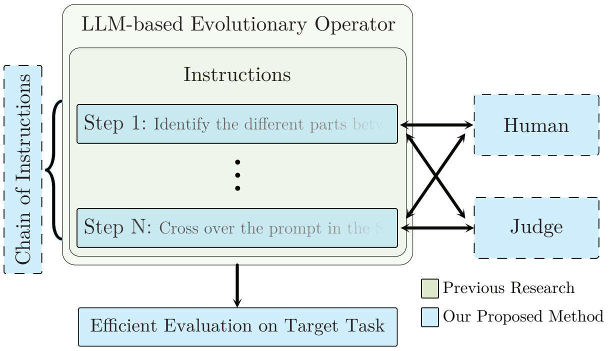

LLM-based Evolutionary Operator
Instructions
Step 1: Identify the different parts betw
Step N: Cross over the prompt in the
Chain of Instructions
Human
Judge
🟩 Previous Research
🟦 Our Proposed Method
Efficient Evaluation on Target Task

Figure 1: An overview of the individual components ascribed to our proposed method (blue) compared to only using a one-step instruction for the operator (green).

In this paper, we address the following research questions regarding LLM-based evolutionary prompt optimization methods with the ultimate goal of improving language understanding in LLMs:

RQ1: How can new candidate prompts be evaluated more efficiently without forfeiting overall performance?

RQ2: Does CoI-based evolution help judges and improve the evolutionary operator?

[*] Corresponding Authors. E-mails: griesshaber@hdm-stuttgart.de, maximilian.kimmich@ims.uni-stuttgart.de.
[1] Equal contribution.

RQ3: Can LLM-based judges assess the output quality of text-generative evolutionary operators?

RQ4: How can human feedback be effectively leveraged to enhance evolutionary prompt optimization?

RQ5: How does the selection of the LLM influence the effectiveness of the proposed improvements?

To explore these questions, we propose an approach where the LLM serves as both an operator for evolutions and a judge of prompts while allowing humans to intervene when the model output is incorrect. This feedback loop allows human corrections to be leveraged for further optimizing the prompt search process. Additionally, several evaluation strategies aim to make the evaluation more efficient and feedback loops faster. An overview is given in figure 1.

Our contributions are as follows: 1) We introduce a novel human-in-the-loop approach for refining LLM-based prompts, where the feedback is used for future optimizations. 2) We introduce an LLM-based judge for verifying LLM-based evolutions. 3) We present chain-of-instructions as a mechanism to enhance control over the evolutionary process and facilitate more effective feedback. 4) We show how the efficiency of prompt optimization methods can be increased without sacrificing performance. 5) We present empirical results investigating the benefits of our approaches in the context of evolutionary prompt optimization methods. 6) We publish all code that is necessary to reproduce the findings presented in this paper under a permissive license to enable further research into the topic of evolutionary prompt optimization and human feedback.

2 Background

In this section, we introduce the foundational concepts and methods that underpin our research on evolutionary prompt optimization.

2.1 Large Language Models

LLMs are a class of language models that leverage deep learning techniques to process and generate human language. They have become increasingly popular recently due to their ability to generate high-quality text across a wide range of tasks [3]. LLMs are trained on large corpora of text data, enabling them to learn complex patterns and relationships within language. They have been applied to a variety of natural language processing (NLP) tasks, including text generation, translation, summarization, and question answering.

The input for these LLMs, known as prompts, remains crucial since the LLMs' performance depends on these prompts. The concept of CoT reasoning has emerged as a powerful approach to enhancing the reasoning capabilities of LLMs. CoT was first introduced by Wei et al. [31], demonstrating that prompting LLMs to decompose complex problems into intermediate reasoning steps significantly improves performance on several tasks.

In scientific and technical domains, the use of LLMs as judges has recently gained increasing attention. Several studies have explored the potential of LLMs for assessing the quality of text in various contexts. For instance, [38] investigated the ability of LLMs to grade academic writing and found that models like GPT-4 can provide feedback comparable to human reviewers. Similarly, works by Liang et al. [14] and Yu et al. [32] have examined LLMs in the context of automated peer review, highlighting their strengths in identifying clarity issues and methodological flaws.

2.2 Prompt Optimization

Prompt optimization is the process of finding the most effective prompt for a given task.

For this work, prompt optimization approaches can be classified into two categories: continuous space optimization and discrete prompt optimization methods. The former treat the prompt as a continuous vector, leveraging the fact that the tokens of the prompt are embedded into a continuous space, and optimize it using gradient-based techniques such as gradient descent [26, 17, 16, 33]. However, this approach requires the model to be differentiable with respect to the prompt, which may not always be the case. Furthermore, the resulting prompt lacks interpretability by humans, as the resulting prompt is a continuous vector that cannot easily be mapped to discrete tokens, making it difficult to understand and refine.

Discrete prompt optimization methods, on the other hand, treat the prompt as a discrete sequence of tokens such that the result can easily be observed as a natural language prompt. However, previous methods often rely on gradients over the model parameters to optimize the prompt, which may not be available in black-box models [28, 5, 27, 35]. In contrast, Guo et al. [7] show that prompts can be optimized without gradients, enabling optimization methods for black-box models.

2.3 Evolutionary Algorithms

Evolutionary algorithms are a class of optimization techniques inspired by the process of natural selection. They operate by iteratively evolving a population of candidate solutions to a problem, selecting the fittest individuals for reproduction and mutation. Evolutionary algorithms have been successfully applied to a wide range of optimization problems, including function optimization, machine learning, and robotics [8]. Lehman et al. [11] demonstrated that LLMs can be used to perform automatic mutation of prompts, while Meyerson et al. [19] showed the same for crossover operations, paving the way for evolutionary prompt optimization. In the context of prompt optimization, evolutionary algorithms can be used to search for the best prompt for a given task by iteratively evolving and evaluating candidate prompts [4]. Guo et al. [7] is the most related work to ours, as they use a genetic algorithm to optimize prompts for an LLM for NLP tasks.

3 Method

Our goal is to find the ideal prompt p for a given low-resource NLP task (see section 4.1) using in-context learning with an LLM. To achieve this goal, a small set of labeled validation data $\mathcal{D}$ is available ($|\mathcal{D}| \leq 200$). The LLM is treated as a black-box function, so no access to its parameters or inner architecture is available and needed, enabling models that are only available via APIs.

The optimization process resembles a genetic algorithm, where a population of prompts is evolved over T generations.

Initialization: To generate the initial population, we select the best $\lfloor I/2 \rfloor$ prompts from a task-specific set of base prompts and generate the remaining $\lceil I/2 \rceil$ prompts by paraphrasing the selected base prompts, where I is the population size. This initialization warm-starts the optimization process with a diverse set of prompts [29].

Evolution: We follow the work of Lehman et al. [11], Meyerson et al. [19], Guo et al. [7] and use an LLM to perform the operations of *mutation* and *crossover* of the evolutionary algorithm. We use the same operator implementations for Differential Evolution (DE) and Generic Algorithm (GA) as baselines to improve upon, also using demonstration data in the input for in-context learning. That is, for the classification tasks, we randomly select one input-output pair per class, and a single example for the other tasks.

Evaluation: Prompt fitness S_i is defined as the performance of an evaluation model with prompt p_i on the validation data: $S_i = \mathcal{E}\left(p_i, \mathcal{D}\right)$.

Selection: Prompts for evolution are selected following Lipowski and Lipowska [15] using a roulette-wheel style algorithm with stochastic sampling. Specifically, a prompt is selected with probability $p_i = \frac{S_i}{\sum_{j=1}^{I} S_j}$.

3.1 *Efficient Evaluation*

The optimization process depends on the repeated evaluation of candidate prompts. Although the evaluation set $\mathcal{D}$ is small, LLM inference is still costly. Given the cost of a single inference step c_i, the total cost of the evaluation c_e can be calculated as $c_e = c_i \times |\mathcal{D}| \times I \times T$. Since c_i is fixed to the model, a larger population size I is generally understood to improve the result of the optimization, and the number of generations T should be high enough to ensure convergence, we employ and present methods to reduce the overall cost of evaluation without negatively affecting the resulting prompt performance.

3.1.1 *Early Stopping*

The fitness function normally calculates a mean average score over all samples in $\mathcal{D}$. Empirical preliminary examination has shown that the score converges before all samples have been tested. Therefore, we evaluate strategies to reduce the number of evaluation inferences without affecting the resulting prompt performance.

Moment-based We propose a moment-based early stopping strategy to stop the evaluation after the score has settled: If the mean absolute difference in evaluation score is less than a minimum change η_m for a window size w, the evaluation is stopped. The stopping criterion can be expressed with the following inequality:

$$\frac{1}{w} \sum_{n=t-w+1}^{t} \left|\mathcal{E}\left(p_i, \mathcal{D}_n\right) - \mathcal{E}\left(p_i, \mathcal{D}_{n-1}\right)\right| < \eta_m, \qquad t \geq w$$

Parent-based Except for the first generation, we have access to the performance of all ancestors of p_i on samples from $\mathcal{D}$. We propose to use this information for a parent-based early stopping decision to exit evaluation early if the current prompt p_i is not performing better by η_p compared to the max score of the parents p_a, p_b in a sliding window w. The stopping criterion is then fulfilled if the following inequality is true:

$$\max_{n=t-w+1,\ldots,t} \left(\mathcal{E}\left(p_i; \mathcal{D}_n\right) - \max\left(\mathcal{E}(p_a; \mathcal{D}_n),\, \mathcal{E}(p_b; \mathcal{D}_n)\right)\right) < \eta_p, \qquad t \geq w$$

We employ the parent-based strategy for generations $T > 1$ with a fallback to the moment-based strategy when parent performance is not available. Both early stopping methods employ a patience parameter to ignore the first evaluation iterations where the score may change drastically.

3.1.2 *Evaluation Strategies*

We consider different orderings of $\mathcal{D}$ for evaluation.

Shortest First With the motivation to reduce the number of tokens the LLM needs to process during evaluation, the early stopping strategy can be extended to use an ordered version of $\mathcal{D}$ that is sorted in ascending order according to the length of the inputs.

Hardest First During evolution, the population of prompts is expected to improve. We therefore propose an evaluation strategy that sorts the samples in $\mathcal{D}$ by the performance of the best parent prompt in ascending order. This is motivated by the fact that a prompt that performs just as well as the best parent prompt on hard samples will not be able to yield a better mean performance on the whole dataset when including samples on which the parent already performed well.

3.2 *CoI Prompting*

We adopt the concept of CoT reasoning for our approach. Rather than instructing the LLM to reason step-by-step, we decompose the instructions for implementing the evolutionary operator into multiple distinct steps. That is, for evolution step t, we formulate the prompt o_t to include instructions i_t and model response r_t for previous steps as well as the instruction for the current step, $o_t = i_0, r_0, \ldots, i_{t-1}, r_{t-1}, i_t$. Here, each instruction i_t is a single operation that the LLM should perform, such as mutating a prompting or crossing over two prompts.

When utilizing demonstration data, we ensure that it aligns with the current stage of evolution, meaning that instructions and model responses up to the current evolution step are included.

3.3 *Evolution Judge*

To avoid an expensive evaluation for prompts that are unlikely to be selected in absence of human feedback, we introduce a judge model $\mathcal{J}$. For this, we use another LLM to assess the quality of a prompt candidate p_i before starting evaluation. To this end, we provide the judge model with the response itself, along with the corresponding inputs that led to it – including demonstration samples, system message and the prompt. In case of CoI, we apply the judge for each evolution step. If the judge model determines a prompt to be of low quality, we ask the evolution model to generate a new response until a predefined number of repetitions is reached. Afterward, if, according to the judge model, there is no prompt of high quality, we continue with a random response from the model.[2]

3.4 *Human Feedback*

To integrate human feedback into the optimization process, we propose a human-in-the-loop approach that actively involves humans at multiple stages of evolutionary prompt optimization. In our framework, human participants are not merely passive evaluators but play an active role in observing, analyzing, and refining the outputs generated by the LLM during each step of the evolutionary process.

Specifically, after each evolutionary step – such as mutation or crossover – humans review the generated model outputs. If deficiencies, ambiguities, or errors are detected, humans intervene by refining the instructions that guide the evolutionary operator. This may include clarifying the language of the instructions, specifying more granular or explicit requirements, or restructuring the sequence of steps

[2] In our implementation, we use the last one since there is no implication on the order of generated model responses, i.e., the randomness is realized via the evolution model.

to reduce confusion for the LLM. Additionally, humans can update or augment the demonstration samples used for in-context learning, ensuring that these examples better illustrate the intended behavior and address previously observed shortcomings.

This process is inherently iterative: after each round of human intervention, the evolutionary process resumes with the updated instructions and demonstration data, allowing for continuous improvement. Over successive cycles, this feedback loop enables the identification and mitigation of persistent weaknesses, such as the LLM's tendency to overlook subtle distinctions or to generate extraneous output, thereby enhancing the language understanding capabilities of the LLM. By systematically addressing these issues, the overall effectiveness and reliability of the prompt optimization process are enhanced.

Furthermore, this approach allows for the accumulation of best practices and refined instructions, which can be reused or adapted for future tasks or models. The iterative nature of human feedback ensures that the optimization process remains adaptable and responsive to the evolving capabilities and limitations of the underlying LLM. Illustrative examples of how human feedback leads to tangible improvements in prompt optimization are provided in section 6.

4 Experimental Setup

This section describes the experiments conducted to explore the effectiveness of the proposed methods. All code, data and information that is neccessary to reproduce the results of the presented experiments is published online.[3]

4.1 Tasks

We evaluate our proposed method on a wide range of NLP tasks, including sentiment analysis (Stanford Sentiment Treebank (SST) 2 & 5 [30], Movie-Reviews (MR) [18], Customer-Reviews (CR) [9]), subjectivity analysis (Subj [22]), topic classification (AG's News Topic Classification Dataset (AGNews) [36], Text REtrieval Conference (TREC) [13]), question answering (Stanford Question Answering Dataset (SQuAD) [24]), simplification (A Dataset of Sentence Simplification Evaluation Test (ASSET) [2]) and summarization (Summarizing Arguments in Online Discussions (SAMSum) [6]). For evaluation of prompt fitness during evolution, we select a subset $\mathcal{D}$ of 200 labeled samples from the validation set. For the final evaluation, we use the whole test set to assess the performance of the evolved prompt.

4.2 Evaluation Strategies

To assess the effectiveness of the proposed evaluation strategies, the methods are compared to a baseline where prompt evaluation is performed on the whole validation set $\mathcal{D}$ and an additional naïve strategy to reduce the evaluation cost by subsampling $\mathcal{D}$ with a fixed factor. Since the fitness score of a prompt is an important metric in the evolutionary algorithm, we also show the score of the final prompt to preclude negative impacts on the final performance of the optimized prompt.

4.3 CoI Prompting

We decompose the instructions of the evolutionary operators, DE and GA, into multiple steps as described in section 3.2. Given the increased number of instructions required for DE, the CoI-based implementation consists of four steps, whereas GA follows a more concise two-step process. The baseline model for the CoI experiments only performs a single step for the evolution of a prompt. To see the combined effect of CoI and judging, experiments using CoI are performed with and without a judge $\mathcal{J}$. For the baseline, the judge can only decide on the single output of the evolution. Experiments are performed on and averaged over all tasks.

4.4 Evolution Judge

We employ a judge in our experiments to verify model responses in absence of humans, as described in section 3.3. In detail, the LLM's instruction is given as follows: "You are acting as a judge. Please read the context, the instruction and the response and decide if the response follows the instruction. If it does, answer 'good'. If it does not, answer 'bad'. Wrap the answer with tags <judgement> and </judgement>. Please also add an explanation for your judgement."[4]. We repeat generating model responses up to three times (if the judge assesses an ouput as bad).

Similar to the CoI results, the experiments for the judge are compared with and without CoI to assess the combined effect on the final performance.

4.5 Human Feedback

After analyzing the evolution model outputs, we iteratively refined the instructions for the evolutionary operator, re-evaluating and further improving the model outputs as needed. Through this process, we applied two consecutive refinements to DE, resulting in DE$_1$ and DE$_2$. Similarly, GA was improved once (GA$_1$), as its fewer instructions needed less care. An example of such refinement is provided in section 6.

4.6 Hyperparameters

For the evolution, we utilized the quantized version of Llama 3.1 8B Instruct [1] as the generative model. The results are reported after conducting $T = 10$ generations with a population size of $I = 10$. We adopt the approach of Guo et al. [7] for selecting base prompts for the initial population described in section 3. Specifically, depending on the task, we utilize prompts from Mishra et al. [20], Zhang et al. [34], Li et al. [12], Sanh et al. [25], Zhang et al. [37]. However, for SQuAD, we employ a single manually crafted prompt alongside generated prompts obtained using the forward mode generation method proposed by Zhou et al. [39].

For both paraphrasing and evolutionary steps, sampling was applied in decoding using a temperature of $t = 0.5$ to increase output variance. We used the same model for the judge and for the evaluation, i.e., Llama 3.1 8B Instruct, but with greedy decoding for increased correctness.

For the early stopping, we set $\eta_m = 10^{-3}$, the window size $w = 10$, $\eta_p = 10^{-3}$ and a patience of 20.

[3] https://gitlab.mi.hdm-stuttgart.de/griesshaber/evoprompt

[4] This prompt is the result of a manual investigation; since this work deals with the optimization of prompts and the manual tuning of prompts contradicts the motivation of this work, this manual tuning is only necessary at this point, but neither reflects our intention to improve prompt optimization, nor does it counteract it.

Table 1: Baseline results for our hyperparameters and models for various classification, question answering, and text generation tasks. Results are reported for two evolutionary algorithms: Differential Evolution (DE) and Genetic Algorithm (GA). GA outperforms DE on most tasks, including AGNews, ASSET, SAMSum, SQuAD, SST-2, SST-5, and Subj. DE achieves better results on CR, MR, and TREC. This indicates that – without any of our improvements – while GA generally yields better overall performance, DE remains competitive on certain tasks.

	Evolution Algorithm	
Task	DE	GA
AGNews	86.89%	**87.99%**
ASSET	54.74%	**56.78%**
CR	**91.85%**	91.40%
MR	**91.30%**	90.55%
SAMSum	29.55%	**29.82%**
SQuAD	86.72%	**89.27%**
SST 2	94.51%	**95.50%**
SST 5	56.24%	**56.33%**
Subj	80.75%	**84.10%**
TREC	**83.20%**	78.80%

4.7 *Other LLMs*

To explore the effect the choice of LLM has on the performance of the proposed methods, we conduct experiments with numerous other LLMs :

`LI3.1 (8B)`: Llama 3.1 8B Instruct[5] (our base model).
`MI (7B)`: Mistral Instruct 7B[6]
`Q2.5 (7B)`: Qwen 2.5 7B Instruct[7]
`R1Q (1.5B)`: Deep Seek R1 Distill Qwen 1.5B Instruct[8]
`R1Q (7B)`: Deep Seek R1 Distill Qwen 7B Instruct[9]
`GI (1B)`: Google Gemma 1B Instruct[10]
`GI (7B)`: Google Gemini 7B Instruct[11]

5 Quantitative Results

This section presents the results of the conducted experiments in a quantitative manner to show the effectiveness of each modification, analyzing their implications with respect to our research questions.

5.1 *Baseline*

Since we aim to investigate the impact of our proposed extensions to evolutionary prompt optimization, table 1 presents baseline results for our chosen hyperparameters and LLM models for all presented tasks and both evolution algorithms. The results mostly replicate the final results presented in Guo et al. [7], but with deviations in hyperparameters and implementation.

5.2 `RQ1`*: Efficient Evaluation*

Table 2 shows the results for the experimental evaluation of the proposed efficient evaluation strategies. The results indicate that, while all evaluation strategies reduce the number of tokens needed to score the candidate prompts, the naïve approach of subsampling $\mathcal{D}$ performs in average worst, demonstrating the importance of high-quality prompt candidate scores for the evolutionary algorithm. Notably, the suggested strategies of evaluating on the shortest and hardest samples first only show minor deviation from the baseline scores while also being most effective reducing used tokens and runtime, whereas early stopping on the unordered scoring set may decrease performance. While the *Hardest First* strategy reduces the evaluation cost the most, *Shortest First* is the only strategy that did not show any decrease in the final evaluation. In total, to answer `RQ1`, both strategies, Hardest First and Shortest First, can effectively reduce the compute usage. They should be chosen based on individual preferences on the final task performance.

Table 2: Comparing the different evaluation strategies to the baseline: average difference of evaluation scores (ΔS), difference in number of tokens used during prompt evaluation (Δc_e), token count (c_e/c_b) and runtime (t_e/t_b) of the experiment as a fraction of the baseline. All reported values are averaged over tasks.

Strategy	ΔS	Δc_e	c_e/c_b	t_e/t_b
Subsample	-2.28%	-7.3M	33.5%	55.6%
Early Stopping	-1.43%	-7.6M	31.0%	53.2%
Shortest First	**+0.11%**	-7.9M	28.3%	43.6%
Hardest First	-0.50%	**-8.0M**	**25.5%**	**42.8%**

Table 3: Performance improvements of CoI over baselines with (✓) and without (✗) using a judge assessing model outputs. Incorporating CoI consistently improves results across both Differential Evolution (DE) and Genetic Algorithm (GA), with a judge yielding best results for both algorithms. For DE, the mean improvement increases from +1.20% to +1.73%, with a higher maximum gain. Similarly, GA sees an increase in mean improvement from +0.68% to +1.00%, with a notable maximum gain of +4.40%. These results demonstrate that CoI optimization yields more reliable and higher-quality model outputs.

Evolution			ΔS	
Algorithm	$\mathcal{J}$	*mean*	*min*	*max*
DE	✗	+1.20%	-0.35%	+3.00%
	✓	**+1.73%**	-0.80%	+3.56%
GA	✗	+0.68%	-0.88%	+2.55%
	✓	**+1.00%**	-0.63%	+4.40%

5.3 `RQ2&3`*: CoI Prompting & Evolution Judge*

In table 3, the relative improvements that can be achieved with CoI prompting are presented. The average across tasks is positive for all configurations, regardless of evolution algorithm and the use of an additional judge.

Notably, the mean improvement for the DE algorithm is higher than for the GA algorithm, independent of whether the judge is used. Since DE is more complex with a higher number of steps, this indicates that CoI helps by breaking the algorithm into discrete steps that can be performed individually. Furthermore, CoI yields the best performance in combination with the judge, additionally motivating to verify model outputs automatically in LLM-based evolutionary operators.

Similar to the enhancements achieved with CoI, incorporating the judge in our approach consistently outperforms the baseline methods, as demonstrated in table 4, regardless of whether CoI is utilized. This result indicates that the judge can successfully detect and reject prompts which are determined to be of low quality before evaluation and therefore provides a positive answer to `RQ3`. The resulting increase in the number of high-quality prompts in the population before

Model references on Hugging Face:
[5] meta-llama/Meta-Llama-3.1-8B-Instruct
[6] mistralai/Mistral-7B-Instruct-v0.1
[7] Qwen/Qwen2.5-7B-Instruct
[8] deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
[9] deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
[10] google/gemma-1b-it
[11] google/gemma-7b-it

Table 4: Performance improvements of applying a judge to assess evolution with (✓) and without (✗) using CoI. The judge improves average performance (ΔS) across both Differential Evolution (DE) and Genetic Algorithm (GA). With CoI, DE sees a higher mean improvement (+0.87% vs. +0.34%) and a reduced worst-case drop in performance. GA also benefits, showing a larger average gain (+0.97% vs. +0.65%) and achieving the highest observed improvement overall (+3.20%). These results highlight the effectiveness of using a judge, and especially in conjunction with CoI, to guide the evolutionary process.

Evolution		ΔS		
Algorithm	CoI	*mean*	*min*	*max*
DE	✗	+0.34%	-0.86%	+2.80%
	✓	**+0.87%**	-0.18%	+2.58%
GA	✗	+0.65%	-0.22%	+1.79%
	✓	**+0.97%**	-0.35%	+3.20%

selection seems to yield an overall improvement in the performance of the final evolved prompt as observed in our results.

In combination, the results from tables 3 and 4 show – in the ablation cases where either CoI or the judge are removed – that both concepts work best in combination, providing a positive answer to `RQ2`. This is to be expected, since the decomposition of the evolutionary prompt also allows the judge to assess each smaller step separately, compared to judging the whole output including multiple steps over longer input-output pairs.

Table 5: Relative score improvements of evolution strategies revised using human feedback (the subscript indicates the iteration) compared to DE and GA, respectively. Incorporating human feedback yields consistent performance gains across most tasks. On average, the second iteration of DE (DE_2) shows the highest mean improvement (+1.67%) reflecting the notion of consecutive refinements, followed by the first DE refinement (DE_1, +1.11%) and the GA refinement (GA_1, +0.75%). The largest individual improvements are observed on TREC and ASSET, indicating that human feedback is particularly effective for tasks involving question classification and text simplification.

Task	ΔS_{DE_1}	ΔS_{DE_2}	ΔS_{GA_1}
AGNews	+1.03%	+1.54%	+0.20%
ASSET	+1.74%	+2.36%	-0.18%
CR	+1.55%	+2.10%	+2.15%
MR	+0.15%	+0.15%	-0.05%
SAMSum	-0.19%	+1.09%	+0.43%
SQuAD	+1.09%	+0.95%	+0.84%
SST 2	+0.90%	+1.68%	+0.88%
SST 5	+1.43%	+2.06%	+0.81%
Subj	+0.20%	+1.40%	+0.20%
TREC	+3.20%	+3.40%	+2.20%
Mean	**+1.11%**	**+1.67%**	**+0.75%**

5.4 `RQ4`: *Human Feedback*

The results presented in table 5 demonstrate the relative performance improvements of evolution strategies incorporating human feedback (DE_1, DE_2, and GA_1) over their respective baseline methods (DE and GA) across multiple tasks. Overall, DE_2 consistently outperforms its baseline, achieving the highest mean improvement of +1.67%, compared to +1.11% for DE_1. GA_1 also benefits from human feedback, but shows a more modest mean improvement of +0.75%. While most tasks exhibit performance gains, there are a few instances where minimal or negative changes occur. Notably, the largest improvements are observed on the TREC dataset showing substantial gains, with DE_2 achieving the highest relative improvement of +3.40%. However, there is a variance on individual tasks induced by the randomness of the evolution, but these findings suggest that integrating human feedback into evolution strategies can enhance performance, with DE-based approaches benefiting more noticeably than GA-based ones.[12]

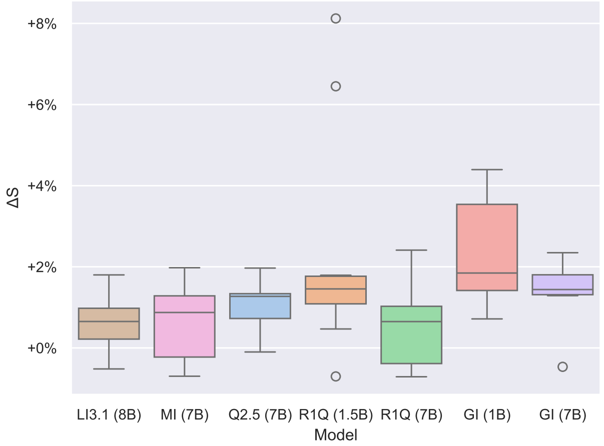


Figure 2: Box plot illustrating the quantitative effectiveness of various LLMs based on performance metrics across our evaluation set. The models are listed in section 4.7. The y-axis represents the relative improvement in performance if CoI and the judge are used. Mean performance is consistently improved across all tested models, with Gemma profiting the most.

Following this evaluation and to answer `RQ4`, we can say that the evolutionary operator is effectively improved using human feedback as proposed in our approach.

5.5 `RQ5`: *Effectiveness of different LLMs*

Figure 2 shows the performance difference between a baseline and runs using both the judge and CoI, with different LLMs models. The evaluation was performed for all tasks presented in 4.1. The results indicate that the proposed methods are effective across all tested models with an overall positive average improvement. Interestingly, the smaller models, such as `R1Q (1.5B)` and `GI (1B)`, show a higher variance in performance, especially when compared directly to the larger variants of the same model (`R1Q (7B)` and `GI (7B)` respectively).

To answer `RQ5`, we were able to demonstrate the effectiveness of our proposed method across a wide combination of tasks and models.

6 Qualitative Analysis of Human Feedback

Figure 3 presents an example of human feedback within our approach. Upon reviewing the model output from the first evolution step, it was observed that not all differences were identified, and some similarities were erroneously included. Additionally, unexpected extraneous output was generated. To address these issues, the prompt for the first evolution step was refined by incorporating the following instructions: "Output a list of all different parts and make sure that differences are only in the form of words and phrases." and "If the same phrase appears in both prompts, do not list it, i.e., do not list similarities."

This example also highlights the ambiguities associated with human involvement: multiple instructions exist, just as there are various

[12] We note that the combination of judge-based and human feedback – where humans corrected the model output if it was judged non-compliant with the instructions – did not consistently enhance performance across all tasks. Consequently, we have opted not to present these results.

Input:
Step 1: Identify the different parts between Prompt 1 and Prompt 2:
Prompt 1: Analyze the sentence and categorize it into one of five categories based on the sentiment: terrible, bad, okay, good, or great.
Prompt 2: Classify the given review into one of five categories: extremely negative (terrible), somewhat negative (bad), neutral (okay), somewhat positive (good), or extremely positive (great).

Response:
1. Identifying the different parts between Prompt 1 and Prompt 2:
Prompt 1: Analyze the sentence and categorize it into one of five categories based on the sentiment: terrible, bad, okay, good, or great.
Prompt 2: Classify the given review into one of five categories: extremely negative (terrible), somewhat negative (bad), neutral (okay), somewhat positive (good), or extremely positive (great).
Different parts:
- "sentence" vs "review"
- "analyze" vs "classify"
- "terrible, bad, okay, good, or great" vs "extremely negative (terrible), somewhat negative (bad), neutral (okay), somewhat positive (good), or extremely positive (great)"
- "based on the sentiment" only appears in Prompt 1
- ~~"based on the sentiment" vs "based on the sentiment" (same phrase)~~
- ~~"five categories" vs "five categories" (same phrase)~~
- ~~The only difference is the subject of analysis, which is a sentence in Prompt 1 and a review in Prompt 2.~~

Figure 3: An example for the first step of evolution for DE: The expected response mentions mutations for all spotted differences (marked in red) and omits the similarities as well as the last statement, which is evidently wrong (marked using strikeout in red). Demonstration samples for in-context learning are omitted for clarity.

possible prompts for a task. By iteratively analyzing the evolution model's output and refining the instructions accordingly, we can facilitate human feedback, ultimately enhancing the prompt optimization process.

Finally, we claim that inspecting the model output and adapting the instructions correspondingly can be accomplished in about half an hour, which is a reasonable time investment for the performance improvements achieved in our experiments.

7 Conclusion

We introduced and investigated extensions to evolutionary prompt optimization that leverage CoI, an LLM-based judge, human feedback and efficient evaluation methods to optimize prompts for a given task.

CoI, by enabling greater control and better decision-making, can enhance performance in prompt optimization (RQ2) and holds promise for broader applications. In particular, when combined with judge-based assessment (RQ3) and human feedback (RQ4), it provides a robust framework for identifying optimal prompts in NLP tasks. Lastly, beyond performance improvements, reducing computational cost is also a key consideration. Our efficient evaluation methods offer a significant reduction in computational overhead while maintaining performance during the search for optimal prompts (RQ1).

We are convinced that our contributions, including investigations and releasing our code, help future research in the area, promoting the effective and efficient use of LLMs in NLP, and especially help in grounding LLMs for better language understanding.

8 Limitations

In this work we only focus on the optimization of the prompts while not focusing on optimizing the verbalizer extracting the predictions for the tasks, which could be a potential improvement since the prompt can contain directives as to what to expect in the model output.

Furthermore, running multiple experiments on the same task can lead to different results due to the stochastic nature of LLMs and the evolutionary algorithms, providing a more reliable performance estimate. However, since the experiments are time-intensive, we instead mitigate this effect by averaging the results over multiple tasks. This also allows us to analyze the performance of our methods across a wide range of tasks, but may not be representative for individual tasks.

Also, although we optimize for faster runtimes and lower token usage, LLMs still require large amounts of compute resources and energy which potentially makes the methods and results presented in this paper inaccessible to some groups without access to such resources. For example, a single optimization of a prompt for SAMSum using the *hardest first* strategy needed 4:24h on a single NVIDIA A6000 GPU while the average GPU memory consumption was about 20GB.

References

[1] AI@Meta. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[2] F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, and L. Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online, jul 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.424.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. URL https://dl.acm.org/doi/abs/10.5555/3495724.3495883.

[4] A. Chen, D. M. Dohan, and D. R. So. EvoPrompting: language models for code-level neural architecture search. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. URL https://dl.acm.org/doi/10.5555/3666122.3666464.

[5] M. Deng, J. Wang, C.-P. Hsieh, Y. Wang, H. Guo, T. Shu, M. Song, E. Xing, and Z. Hu. RLPrompt: Optimizing discrete text prompts with

reinforcement learning. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates, dec 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.222.

[6] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In L. Wang, J. C. K. Cheung, G. Carenini, and F. Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, nov 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409.

[7] Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, and Y. Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZG3RaNIsO8.

[8] J. H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. MIT Press, Cambridge, MA, USA, 1992. ISBN 0262082136.

[9] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138881. doi: 10.1145/1014052.1014073.

[10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. (12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730.

[11] J. Lehman, J. Gordon, S. Jain, K. Ndousse, C. Yeh, and K. O. Stanley. Evolution through large models, 2022. URL https://arxiv.org/abs/2206.08896.

[12] B. Li, R. Wang, J. Guo, K. Song, X. Tan, H. Hassan, A. Menezes, T. Xiao, J. Bian, and J. Zhu. Deliberate then generate: Enhanced prompting framework for text generation. 2023.

[13] X. Li and D. Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150/.

[14] W. Liang, Y. Zhang, H. Cao, B. Wang, D. Ding, X. Yang, K. Vodrahalli, S. He, D. S. Smith, Y. Yin, D. A. McFarland, and J. Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. 2023.

[15] A. Lipowski and D. Lipowska. Roulette-wheel selection via stochastic acceptance. (6):2193–2196, 2012. ISSN 0378-4371. doi: 10.1016/j.physa.2011.12.004.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. (9), jan 2023. ISSN 0360-0300. doi: 10.1145/3560815.

[17] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, and J. Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, may 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.8.

[18] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In D. Lin, Y. Matsumoto, and R. Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, jun 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

[19] E. Meyerson, M. J. Nelson, H. Bradley, A. Gaier, A. Moradi, A. K. Hoover, and J. Lehman. Language model crossover: Variation through few-shot prompting. (4), 2024. doi: 10.1145/3694791.

[20] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

[21] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[22] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics, jul 2002. doi: 10.3115/1118693.1118704.

[23] G. Perković, A. Drobnjak, and I. Botički. Hallucinations in LLMs: Understanding and addressing challenges. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 2084–2088, 2024. doi: 10.1109/MIPRO60963.2024.10569238.

[24] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, nov 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264.

[25] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. V. Nayak, D. Datta, J. D. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z.-X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Févry, J. A. Fries, R. Teehan, S. Biderman, L. Gao, T. Bers, T. Wolf, and A. M. Rush. Multitask prompted training enables zero-shot task generalization. 2021.

[26] T. Schick and H. Schütze. Exploiting Cloze-questions for few-shot text classification and natural language inference. In P. Merlo, J. Tiedemann, and R. Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, apr 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.20.

[27] W. Shi, X. Han, H. Gonen, A. Holtzman, Y. Tsvetkov, and L. Zettlemoyer. Toward human readable prompt tuning: Kubrick‘s the shining is a good movie, and a good prompt too? In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10994–11005, Singapore, dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.733.

[28] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online, nov 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346.

[29] D. So, Q. Le, and C. Liang. The evolved transformer. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 5877–5886. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/so19a.html.

[30] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, and S. Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, oct 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

[31] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. URL https://dl.acm.org/doi/10.5555/3600270.3602070.

[32] J. Yu, Z. Ding, J. Tan, K. Luo, Z. Weng, C. Gong, L. Zeng, R. Cui, C. Han, Q. Sun, Z. Wu, Y. Lan, and X. Li. Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

[33] N. Zhang, L. Li, X. Chen, S. Deng, Z. Bi, C. Tan, F. Huang, and H. Chen. Differentiable prompt makes pre-trained language models better few-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ek9a0qIafW.

[34] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. OPT: Open pre-trained transformer language models. 2022.

[35] T. Zhang, X. Wang, D. Zhou, D. Schuurmans, and J. E. Gonzalez. TEMPERA: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=gSHyqBijPFO.

[36] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA, 2015. MIT Press.

[37] Y. Zhang, L. Cui, D. Cai, X. Huang, T. Fang, and W. Bi. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. 2023.

[38] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. 2023.

[39] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. 2022.

Improving LLMs for Machine Translation Using Synthetic Preference Data

Dario Vajda[*], Domen Vreš and Marko Robnik-Šikonja

University of Ljubljana, Faculty of Computer and Information Science

Abstract. Large language models have emerged as effective machine translation systems. In this paper, we explore how a general instruction-tuned large language model can be improved for machine translation using relatively few easily produced data resources. Using Slovene as a use case, we improve the GaMS-9B-Instruct model using Direct Preference Optimization (DPO) training on a programmatically curated and enhanced subset of a public dataset. As DPO requires pairs of quality-ranked instances, we generated its training dataset by translating English Wikipedia articles using two LLMs, GaMS-9B-Instruct and EuroLLM-9B-Instruct. We ranked the resulting translations based on heuristics coupled with automatic evaluation metrics such as COMET. The evaluation shows that our fine-tuned model outperforms both models involved in the dataset generation. In comparison to the baseline models, the fine-tuned model achieved a COMET score gain of around 0.04 and 0.02, respectively, on translating Wikipedia articles. It also more consistently avoids language and formatting errors.

1 Introduction

Decoder-only large language models (LLMs) serve as versatile tools for a variety of natural language processing tasks, such as question answering, summarization, and translation. Typically, LLMs undergo three phases of training: pretraining, supervised fine-tuning (SFT), and preference alignment. The quality of a translation depends on many fine details (e.g., style, semantics, figurative language, etc.), which might not be sufficiently well learned during SFT. Our hypothesis is that a model can improve on subtle differences between a reasonable and good translation through preference alignment. In this work, we present the training of an LLM with Direct Preference Optimization (DPO) [21] for optimizing its translation abilities. Using Slovene as a use case, our primary goal is to develop a reliable open-source English to Slovene translator that can be used for translating large English corpora to Slovene.

The main contributions of our research are:

- A language agnostic method for improving translation models. The approach is based on synthetic data generation and is suitable for less-resourced languages such as Slovene.
- An open-source English to Slovene translation model capable of reliably generating high-quality translations.[1]
- Source code for our data generation and fine-tuning pipeline for easier reproduction.[2]

Less-resourced languages, such as Slovene, lack sufficient high-quality data. One way to obtain more data is by translating high-quality English corpora. However, for Slovene, current open-source translators, such as RSDO [13], are not reliable enough for such a task, and the most successful commercial models, such as DeepL[3] are too expensive for translating large corpora. Hence, there is a need for a reliable open-source English to Slovene machine translation model.

Currently, there are no existing preference-annotated datasets for English to Slovene translation. Obtaining such a dataset using human translators and annotators would be slow and prohibitively expensive. Hence, we automatically create such a dataset. Our core insight is that even without access to a human-curated preference corpus for English to Slovene translation, we can bootstrap a reliable preference dataset by exploiting the behavior of two independent LLMs and a suite of automated filters. By prompting both GaMS-9B-Instruct [4] and EuroLLM-9B-Instruct [15] to translate English Wikipedia articles [30] and a collection of English news articles from Common Crawl (CC-News dataset), we generate dual translations for each article, resulting in a dataset with around 67,000 entries from Wikipedia and 30,000 from CC-News. Whenever one model produces a clean Slovene output and the other makes an obvious error - whether by continuing the conversation in the wrong language, by truncating the translation, or by adding unwanted prefixes - we can confidently mark the former as *chosen* and the latter as *rejected*. To capture finer quality differences, we score all translations without any obvious mistakes with COMET [23] and select pairs whose COMET scores differ by some minimum threshold. The result is a diverse, synthetic preference dataset that reflects both unacceptable errors (wrong language, incomplete output, etc.) and subtler language fluency distinctions captured by the COMET scores. After curating the translation pairs, our final preference dataset consists of around 25,000 entries for Wikipedia and 10,000 from CC-News.

Using generated synthetic preference data, we apply Direct Preference Optimization (DPO) to the GaMS-9B-Instruct model. We chose DPO because it directly optimizes the likelihood ratio between chosen and rejected outputs, sidestepping the instability and reward-modeling overhead of standard Reinforcement Learning from Human Feedback (RLHF) pipelines [2]. We train the model on the generated preference pairs over three epochs on 16 A100 GPUs. We employ a linear learning-rate warm-up followed by cosine decay to stabilize the early DPO updates. This fine-tuning framework takes advantage of both our filtering heuristics and DPO's principled ranking objective, driving the 9B-parameter model toward more fluent and

[*] Corresponding Author. Email: vajdadario@gmail.com

[1] https://huggingface.co/GaMS-Beta/GaMS-DPO-Translator

[2] https://github.com/DarioVajda/translation_dpo

[3] https://www.deepl.com

complete Slovene translations.

We evaluate the effectiveness of our approach on the public SloBench leaderboard[4] and on unseen Wikipedia and CC-News articles. We show that our DPO-trained GaMS-9B-Instruct model outperforms the original model variant on SloBench and almost matches the performance of considerably larger GaMS-27B-Instruct [3]. Additionally, on our Wikipedia benchmark, testing for language and formatting consistency, our model achieves an error rate of $0.8\ \%$, which is a substantial improvement over the original model's $12\ \%$.

The rest of the paper is organized into five sections. In Section 2, we present related work on the development of Slovene language models and other tools we rely on. In Section 3, we explain how the training data was generated through our heuristic approach and automatic metrics. The training pipeline and hyperparameter search are presented in Section 4. In Section 5, we evaluate our model and compare it to other state-of-the-art open-source models. We provide conclusions, limitations, and directions for further work in Section 6.

2 Background and Related work

Our work is related to other research on LLMs for less-resourced languages, like Slovene, the ongoing research in preference alignment of LLMs, machine translation, and evaluation methods for machine translation. These topics are outlined below.

2.1 Slovene Large Language Models

Most current open-source LLMs are trained on predominantly English data. However, there are some multilingual models that support Slovene. Examples of such models are EuroLLM [15], Gemma 2 [6], and Gemma 3 [7]. These models were trained on multilingual datasets, with small amounts of Slovene texts. The portion of Slovene data in these datasets was relatively small (e.g., the EuroLLM dataset consists of only around $1\ \%$ Slovene texts), meaning there is room for improvement.

Efforts to develop strong LLMs for Slovene have primarily focused on adapting existing English-centric models due to the high cost of training an LLM from scratch. A notable early initiative in this area was the development of SlovenianGPT [8] and GaMS-1B [27]. Those were followed by the development of 2B, 9B, and 27B parameter versions of the GaMS model. These developments demonstrated a key methodology for less-resourced languages: continuing the pre-training of a powerful base model on a relatively large corpus of Slovene text. The process also involved creating a new subword tokenizer adapted to the specifics of the Slovene language and employing embedding initialization techniques to transfer knowledge from the original English model.

2.2 Machine Translators

Inspired by Vaswani et al. [25], machine translators have typically been built on encoder–decoder Transformer architectures. Later, OpenAI proved that decoder-only language models have the potential to learn many language-related tasks [20], including machine translation. They showed it by training their model for translation between English and French.

The performance of machine translators on less-resourced languages such as Slovene often does not match the performance of models for high-resource languages. For example, a Slovene open-source RSDO Neural Machine Translator encoder-decoder model trained specifically for English-to-Slovene translation performs poorly on unseen domains compared to recent state-of-the-art models. Currently, the best-performing machine translators for Slovene, based on a public SloBench leaderboard, are DeepL, Claude, GPT, and Gemini. However, these models are commercial and translating larger corpora with them is costly. The best open-source models on this benchmark are EuroLLM and GaMS. However, as our preliminary evaluation shows, these models are unreliable for translating larger corpora, making trivial mistakes on some occasions. In our work, we focus on fixing such mistakes.

Multilingual open-source translators, such as NVIDIA Riva [17] and Meta's No Language Left Behind (NLLB) [16] perform worse than GaMS and EuroLLM on English-to-Slovene translation. NVIDIA Riva is a GPU-accelerated SDK (Software Development Kit) for building Speech AI applications, focusing on neural machine translation (NMT) while NLLB is a single, massively multilingual model that leverages transfer learning across languages to improve the translation quality of low-resourced languages.

2.3 Preference-Based Model Alignment

Preference alignment methods are used to improve the quality of large language models' outputs, and a few approaches have proven to be very effective at accomplishing that goal. The most classic example is Reinforcement Learning from Human Feedback (RLHF), that is a relatively complex process requiring training of a dedicated reward model. On the other hand, Direct Preference Optimization has recently gained significant traction and was shown to give competitive results without relying on an external reward model. With DPO the reward model is "embedded" inside the LLM and allows for a simpler and more efficient training pipeline with the goal of maximising the log-probability of chosen responses, $\log \pi_\theta(y_w|x)$, while minimising that of rejected responses, $\log \pi_\theta(y_l|x)$.

2.4 Automatic MT Quality Metrics

One of the standard automatic MT metrics is BLEU (Bilingual Evaluation Understudy) [18], which relies on n-gram overlap with reference translations and is incorporated into the SloBench evaluation. Recently, a shift towards learned metrics was driven by the need to capture semantic meaning, not just word overlap. The COMET framework uses cross-lingual embeddings to achieve a much higher correlation with human quality assessments. Therefore, to evaluate the translation pairs in our dataset, we employ the state-of-the-art reference-less Direct Assessment (DA) model from CometKiwi [24] known for its high correlation with human judgment.

3 Synthetic Preference Data Generation

Obtaining a high-quality preference annotated translation dataset is challenging. We take English Wikipedia and news articles from Common Crawl as a starting point, as those cover a wide variety of topics. Our data generation pipeline consists of multiple stages. We start by generating translations, described in Section 3.1. This is followed by identifying trivial errors, described in Section 3.2, and scoring remaining instances, described in Section 3.3. Our final data construction is described in Section 3.4.

[4] https://slobench.cjvt.si/

3.1 Generating Pairs of Translations

The main challenge of generating a synthetic preference dataset is generating distinct translation candidates. Notably, the generated errors shall not propagate or accumulate through the process. Our approach is to generate candidate translations from a corpus of articles which have been selected that cover a broad range of topics with two distinct models and rank the responses using automatic quality metrics.

The first model we use is GaMS-9B-Instruct [4], which is based on the Gemma 2 architecture and adapted for Slovene. As this is the model we also aim to fine-tune as the final machine translator, this allows us to construct preference pairs that target the model's natural output distribution. The second model we utilize for generating the translations is EuroLLM-9B-Instruct [15]. We chose this model because it is an open-source model that demonstrated strong performance and reliability for English-to-Slovene translation in our preliminary experiments.

We use these two models to translate over 67,000 Wikipedia articles, consisting of approximately 26 million words. We use the following prompt instructions:

- GaMS: *"Prevedi naslednje angleško besedilo v slovenščino."* (en. *"Translate the following English text to Slovenian."*)
- Euro-LLM: *"Translate the following English text to Slovenian."*

We filter the initial pool of translations during the subsequent data curation steps.

3.2 Identifying Failure Modes

Upon inspecting the translations, we identified several error types that are critical to the model's reliability, yet simple to represent as preference pairs with unambiguous chosen and rejected examples.

The most significant failure mode observed was generating outputs in the wrong language. To programmatically verify the language of each translation, we utilize the pre-trained language identification model from the FastText library [11, 12]. We use a lightweight and efficient classifier capable of accurately identifying 176 different languages from raw text, making it highly suitable for large-scale filtering tasks. This process forms high-confidence preference pairs by identifying instances where one generated translation is in Slovene and the other is in a different language. The correct, Slovene translation is labeled as *chosen*, while the incorrect one is labeled as *rejected*.

Another identified failure mode is translation truncation, where the model only translates a portion of the source text. We hypothesize that this behavior with GaMS-9B-Instruct is a result of its SFT dataset containing only sentence-level translation tasks. Therefore, the model learned to respond to translation tasks with short answers. This type of structural error is particularly well-suited for correction with DPO. To address this, we create preference pairs from instances where both translations are in Slovene, but one is complete while the other is clearly truncated. The complete translation is labeled as *chosen*, and the truncated version as *rejected*. A translation was considered truncated with high confidence if it was less than 50% of the length of the original text, measured by character count.

A more subtle issue we identified was the presence of stylistic formatting artifacts. Sometimes the model starts a response with *"Slovenski prevod:"* (en. Slovene translation), *"Slovene translation:"*, etc. Since the goal is to produce only the translated text, this behavior is addressed by creating a specific type of preference pairs for our training dataset. Translation pairs were constructed in the following manner: the *chosen* response is a clean, complete translation, while the corresponding *rejected* response is created by prepending the *chosen* text with one of the undesirable prefixes. This method provides a clear and direct preference signal to the model during DPO.

3.3 Scoring and Filtering the Translations

While the initial heuristic filtering addresses clear structural errors, discerning finer differences in quality requires a quantitative metric. For this purpose, we employ the COMET score, specifically the Unbabel/wmt22-cometkiwi-da model [24]. We select this model as it is a state-of-the-art, reference-less Direct Assessment (DA) model that excels at predicting translation quality with a high degree of correlation to human judgment.

All translation pairs that pass the initial heuristic checks are then scored using this COMET model. The scores serves as a proxy for human preference. Since many translation pairs exhibit only minor quality differences, we introduce a minimal score difference threshold to take a given translation pair into consideration. This step is crucial to prevent metric noise from being misinterpreted as a meaningful preference signal. Consequently, a preference pair is only created if the absolute score difference between the two candidates is greater than 0.05. The translation with the higher score is labeled as *chosen* and the other as *rejected.*

3.4 Constructing the preference dataset

The preference pairs generated from the preceding heuristic and metric-based methods are merged to form the final training dataset. This combined dataset is designed to capture both critical failure modes of the base model as well as more subtle preference signals based on clarity, grammar, and style. The above process reduces the number of translation pairs from a total of 107,000 to approximately 35,000. The number of translation pairs for our dataset was decreased because not all of them carried useful information. Additionally, the synthetically generated formatting pair count was chosen to make up around 20% of the final dataset. Those were added to the other systematically curated pairs from the original translation pair dataset. The distribution of training examples is as follows:

- Pairs targeting incorrect language - 22 %
- Pairs targeting response truncation - 3 %
- Synthetically generated formatting pairs - 20 %
- Pairs derived from COMET score differences - 55 %

4 Translator training

To produce an improved LLM for MT, we take GaMS-9B-Instruct model as our starting point and optimize it for translation on the preference dataset from Section 3 using the DPO method. We provide a brief description of the method in Section 4.1. We describe our training implementation in Section 4.2. In Section 4.3, we describe the hyperparameter search performed.

4.1 Using DPO for Machine Translation

Traditional approaches for preference alignment, such as RLHF, rely on training a dedicated reward model. Since we are using a synthetic preference dataset, using another synthetic data-based model in our training pipeline would introduce additional noise and risk instability in the fine-tuning process.

Therefore, we chose Direct Preference Optimization (DPO) as our fine-tuning method due to its stability and efficiency compared to traditional reinforcement learning-based approaches like Proximal Policy Optimization (PPO). DPO provides a cleaner and more straightforward training pipeline with less room for error accumulation. As the DPO loss function is mathematically equivalent to the objective in traditional RLHF, it offers the same optimization guarantees within a more direct and stable framework. Given the data distribution $\mathcal{D} = \{(x, y_w, y_l)\}$, where x is the model's input, y_w denotes the chosen (preferred) response and y_l denotes the rejected response, the DPO loss function aims to minimise:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\log\sigma\Bigg(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Bigg)\Bigg] \tag{1}$$

In this formulation, π_θ represents the fine-tuned policy (model) and π_{ref} denotes the reference policy (usually the starting model). The reference model, π_{ref}, is a crucial component that regularizes the training process. For our experiments, we use the initial GaMS-9B-Instruct model as the reference model. The temperature parameter, β, controls how strongly the policy model adheres to the preference data. A higher β leads to a closer fit to the preference pairs, while a lower β maintains closer proximity to the reference model's initial behavior. We determine the exact β value through hyperparameter tuning.

4.2 Implementation and Training Environment

We use the HuggingFace Transformers [29] DPO implementation. Specifically, we use the `TRL` (Transformers Reinforcement Learning) [26] library in combination with `Accelerate` [9] and `Deepspeed` libraries.

To make training of a 9B-parameter model computationally feasible, we employ a parameter-efficient approach. Specifically, we use Low-Rank Adaptation (LoRA) [10] from the `peft` [14] library. LoRA enables efficient adaptation of large pre-trained models by introducing and training only low-rank update matrices, thereby reducing the number of trainable parameters by orders of magnitude, lowering both GPU memory and storage requirements. This parameter-efficient approach accelerates fine-tuning and simplifies model deployment, while achieving comparable task performance to full model fine-tuning.

We performed the training on the Slovene HPC Vega supercomputer. We utilized a configuration of 4 compute nodes, each equipped with 4 NVIDIA A100 40GB GPUs, for a total of 16 GPUs per training run. The GPUs on a single node are connected using NVLINK with total bandwidth of 600 GB/s. The nodes are connected through 2x200 Gb/s InfiniBand switches in Dragonfly+ topology.

To manage the substantial memory requirements of fine-tuning the 9B-parameter model, even with LoRA, we employed the ZeRO (Zero Redundancy Optimizer) Stage 2 optimization strategy [22], as implemented in the DeepSpeed library. This technique mitigates memory redundancy across data-parallel workers by partitioning not only the optimizer states (as in Stage 1) but also the gradients. While each GPU maintains a complete copy of the model's parameters for the forward and backward passes, it is only responsible for storing and updating a distinct shard of the gradients and corresponding optimizer states. After the local optimizer step, an all-gather operation efficiently synchronizes the updated weights across all GPUs, ensuring model consistency for the next iteration. This approach dramatically reduces the per-GPU memory footprint compared to standard data parallelism, making it feasible to fine-tune the full model on our 16 A100 GPU setup without resorting to more complex model parallelism.

The final major optimization we use is Gradient Checkpointing [1]. It trades additional computation (around 30–40% increase for LLMs) for reduced activation and gradient memory usage (by a factor of approximately $\sqrt{\text{num\_layers}}$) by selectively storing only a subset of intermediate activations during the forward pass and recomputing the omitted activations *on-the-fly* in the backward pass. This enables training of much deeper or wider models under fixed memory budgets, making it particularly valuable for large-scale deep training or fine-tuning.

4.3 Training and Hyperparameter Grid-Search

We split the curated preference dataset only with translations of Wikipedia articles into the training and validation sets with 24,000 and 1,000 instances, respectively. The validation set was used to monitor performance during training and to select the optimal hyperparameters through a grid search.

The key hyperparameters for our DPO training runs are detailed in Table 1. We conducted a grid search over the DPO β and learning rate. The final model was trained for 3 epochs using the optimal configuration discovered during this search. To prevent overfitting, we compare all checkpoints using validation loss. Observed training and validation losses are shown in Figure 1. Since we achieved the lowest validation loss of 0.315 for hyperparameter values $\beta = 0.1$ and $lr = 1 \cdot 10^{-6}$ at the second-to-last evaluation step, this is the final version of our model. Each one of the grid-search training runs lasted around 5 hours.

Table 1. Training hyperparameter values. For DPO β and learning rate, the search domains are provided. The bold values were selected as optimal.

Parameter	Value
Epochs	3
Micro batch size	1
Global batch size	16
DPO β	{**0.1**, 0.2}
LoRA rank	64
Learning rate	{**1e-6**, 4e-7, 1e-7}
Warmup steps	1500
Learning rate scheduler	cosine_with_min_lr

Once the optimal hyperparameter configuration was found, the training dataset was expanded by adding the translation pairs from CC-News (approximately 10,000 new training examples). The model was trained on the larger dataset for three epochs. This final fine-tuning on the complete dataset with the given hyperparameters took approximately 7 hours. This training run resulted in a model checkpoint with a noticeably lower validation loss of 0.255. This checkpoint is the model we will be evaluating further and comparing it to GaMS-9B-Instruct, EuroLLM-9B-Instruct and some other models.

5 Evaluation and Results

We evaluate our model on two benchmarks. The first is a public Slovene-to-English benchmark, that is part of the public benchmark

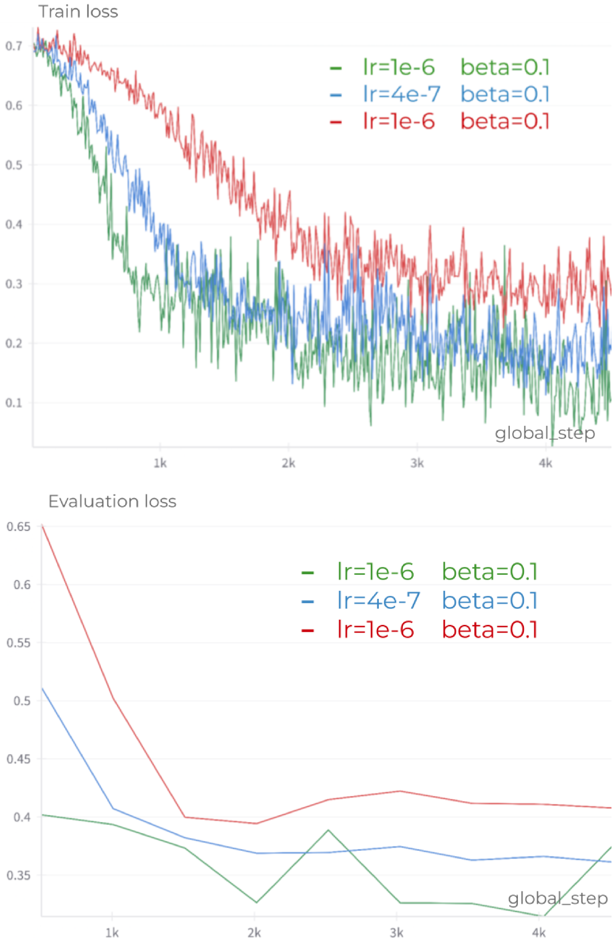


Figure 1. DPO training (top) and validation loss (bottom) curves for different learning rate and DPO β hyperparameter values.

suite SloBench. Since SloBench evaluates the models only on per-sentence translations, we also evaluate our model on a custom evaluation set relevant to our specific goal of translating longer English documents for training large language models for Slovene. We create such an evaluation set based on a set of unseen English Wikipedia and CC-News articles. Throughout this section, we refer to our model as GaMS-9B-DPO-Translator.

5.1 *SloBench Evaluation*

SloBench is an evaluation platform for benchmarking Slovene large language models and their capabilities. Since the benchmarks' ground truths are not public, we believe that benchmark tuning, leading to misleading results, is not possible, making this benchmark an objective measure for many Slovene natural language processing tasks. We test the model on the `Machine Translation (ENG -> SLO)`[5] task. The task consists of five different domains: Scientific articles, Speech texts, Legal articles, News articles, and Technical texts.

The results are shown in Table 4. Even though our model was not specifically tuned for any of the benchmark's domains, our DPO training resulted in a noticeable improvement in comparison to the base model (GaMS-9B-Instruct). GaMS-9B-DPO-Translator achieved a similar score to GaMS-27B-Instruct, which is three times larger.

The improvement would likely be even higher if our preference dataset included a broader range of data, not just Wikipedia documents. Therefore, incorporating more conversational, legal, and news texts into our data generation pipeline and training the model with DPO on such a dataset would better capture all domains tested by SloBench and potentially increase its score. Writing styles in those domains often differ from the Wikipedia articles we used.

5.2 *Wikipedia and CC-News Evaluation*

A more suitable evaluation method for our goal of translating large amounts of longer documents is to compare our fine-tuned model to the base model when translating English Wikipedia and CC-News articles that were not seen during training or validation.

The creation of an evaluation dataset was very similar to generating the training dataset. We translated 500 randomly chosen articles from both sources with GaMS-9B-Instruct (base model), GaMS-9B-DPO-Translator (fine-tuned model), and EuroLLM-9B-Instruct (for reference). Those articles were chosen in a way that none of those have been seen during training or validation by our fine-tuned model.

The first step in analyzing the performance of our model is to check for any of the trivial mistakes we already uncovered when preparing the training dataset. The comparison of models on such mistakes is shown in Table 2.

Table 2. Error rates comparison on our custom Wikipedia Evaluation dataset. Each model name refers to its 9B parameter instruction-tuned variant.

Model	Language Error	Truncation Error	Combined
EuroLLM-9B-Instruct	1%	0.4%	1.4%
GaMS-9B-Instruct	9.5%	3.5%	13%
GaMS-9B-DPO-Translator	**0.6%**	**0.2%**	**0.8%**

For the articles without trivial errors, we calculated the COMET scores of each model's translation and compared them. To ensure fair comparison, only those articles were used where none of the models made any critical mistakes. Obtained COMET scores are shown in Table 3.

Table 3. Comparison of COMET scores on our custom Wikipedia Evaluation dataset. Higher score is better.

Model	Wikipedia	CC-News	Average
EuroLLM-9B-Instruct	0.727	0.667	0.695
GaMS-9B-Instruct	0.722	0.680	0.698
GaMS-9B-DPO-Translator	**0.757**	**0.715**	**0.735**

Our fine-tuned model outperformed both models that were used in the dataset construction, showing that those models do not necessarily represent an upper limit for fine-tuning performance. The reason why our fine-tuned model is able to outperform the construction models is that DPO does not directly train the model to replicate all translations from the dataset. This allowed our model to learn good aspects of both EuroLLM-9B-Instruct and GaMS-9B-Instruct, but also helped it to avoid their mistakes.

[5] https://slobench.cjvt.si/leaderboard/view/8

Table 4. Comparison of different Machine translation models on the public SloBench English-to-Slovene translation leaderboard. The results of OpenAI GPT 4o-mini, GaMS-9B-Instruct, GaMS-27B-Instruct, and EuroLLM-9B-Instruct are taken directly from the leaderboard. The results of our model are in bold.

Model	BERT score	BLEU (avg)	METEOR (avg)	CHRF (avg)	BLEU (corpus)	CHRF (corpus)
EuroLLM-9B-Instruct	0.8741	0.2927	0.5792	0.6055	0.3273	0.6055
GaMS-27B-Instruct	0.8734	0.2866	0.5688	0.5986	0.3246	0.5986
GaMS-9B-DPO-Translator	**0.8726**	**0.2810**	**0.5663**	**0.5967**	**0.3252**	**0.5967**
GaMS-9B-Instruct	0.8713	0.2773	0.5616	0.5928	0.3209	0.5928
GPT 4o-mini	0.8690	0.2619	0.5456	0.5839	0.3021	0.5839

There is a discrepancy between the results on SloBench and the results on our custom Wikipedia article translation test when we compare EuroLLM-9B-Instruct and our fine-tuned GaMS-9B-DPO-Translator. We hypothesize that this is due to the difference in example lengths between benchmarks. The model should benefit from our training, especially on longer texts, as the errors (wrong language, truncation) of the base model were rarer on shorter texts.

A limitation of our evaluation on Wikipedia articles is that it exhibits a similar distribution to our training data, making our model more likely to perform well, and those circumstances might have benefited our model in comparison to others. However, this limitation does not dispute the fact that we successfully improved the model and achieved our initial goal of reliably and accurately translating longer documents. Since Wikipedia captures a variety of different fields and topics, this learned knowledge should carry over to other types of documents, which will be useful for generating new training data for Slovene LLMs.

5.3 *Efficiency of our solution*

Our solution is very efficient from the training data acquisition standpoint and eliminates the need for manual labeling, since it doesn't require any human annotators.

Data acquisition. The dataset creation pipeline is computationally efficient since it involves batched inference of LLMs and other lightweight models such as the COMET model for translation scoring and the language identification model from the FastText library. This process overcomes the challenge of obtaining high quality training data for low-resource languages, required by SFT. In our case, to acquire the dataset we used approximately 3 hours on one node with 4 A100 40GB GPUs (**12 GPU hours**).

Fine-tuning. On the other hand, fine-tuning the model is comparable to SFT, with a notable difference being that DPO requires two forward passes per example (for *chosen* and *rejected* responses). The fine-tuning was run on 4 nodes with 4 A100 40GB GPUs each and ran for around 7 hours (**112 GPU hours**). We used 4 nodes to speed up the process, but the minimum hardware requirement with a 9B parameter model for this step is only one such node and it should run for less than 28 hours (since computation time decreases at a close to linear rate with respect to the number of GPUs when using ZeRO stage 2), which is quite reasonable for a model like the one we used.

6 Conclusion

We proposed a pipeline for improving machine translation based on data generation and DPO preference alignment method. We showed that our approach increases the quality of the trained model's translations. We showed a small performance improvement on the SloBench evaluation, and a substantial improvement in translating longer documents, such as Wikipedia articles.

The main goal of our research was to make an open-source translator for a less-resourced language (Slovene) more reliable. Since our approach is language agnostic (given the precondition of the existence of at least two machine translation options for this language), it can be applied to many less-resourced languages or specific domains. We believe that our approach will help translate high-quality English corpora to less-resourced languages which is necessary to build LLMs in such languages and important for the sovereignty of such languages in the LLM era.

We plan to use the insight gained during this project to fine-tune the 27B parameter model with the same training pipeline. Since the systems are already in place, the remaining challenges are to scale the training process and to obtain the required computational resources. Scaling for a larger model would involve generating more training data and using more advanced distributed training optimizations such as ZeRO Stage 3, ZeRO++ [28]. Additionally, the recently released NVIDIA NeMo-RL framework[6] shall be tested.

A potential improvement to consider in the future is Curriculum DPO [5, 19] instead of vanilla DPO. Curriculum learning would allow the model to learn on different datasets, step-by-step, increasing in difficulty. The datasets could be divided into two major groups. The first group would contain the training examples from our heuristic-based analysis (language and truncation errors), and the second group would have the training examples ranked by COMET score. The latter could be further subdivided into multiple subsets based on the COMET score difference between *chosen* and *rejected* responses. A lower score delta indicates a more subtle difference in quality, and the model would be trained on those after it had been trained on the pairs with a more obvious quality difference.

Finally, other preference alignment methods, such as GRPO, could be tested. Since we have already automated response rankings, we would have to turn those rankings into a reward function for GRPO. Another possibility would be combining both methods by first focusing on language and truncation errors using DPO and then performing GRPO based on COMET scores.

Acknowledgements

The work was primarily supported by the EU cofinancing for research innovation projects in support of green transition and digitalisation (project PoVeJMo, no. C3.K8.IB). Further support was provided by the Slovene Research and Innovation Agency (ARIS) project GC-0002 and the core research programme P6-0411. The work was also supported by EU through ERA Chair grant no. 101186647 (AI4DH). The computational resources were provided by SLING through project S24O01-42.

[6] https://github.com/NVIDIA-NeMo/RL

References

[1] T. Chen, B. Xu, C. Zhang, and C. Guestrin. Training deep nets with sublinear memory cost, 2016. URL https://arxiv.org/abs/1604.06174.

[2] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

[3] CJVT. GaMS-27B-Instruct, 2025. URL https://huggingface.co/cjvt/GaMS-27B-Instruct.

[4] CJVT. GaMS-9B-Instruct, 2025. URL https://huggingface.co/cjvt/GaMS-9B-Instruct.

[5] F.-A. Croitoru, V. Hondru, R. T. Ionescu, N. Sebe, and M. Shah. Curriculum direct preference optimization for diffusion and consistency models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[6] Gemma Team. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118, 2024.

[7] Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

[8] A. Gordić. SlovenianGPT - an open-source LLM for Slovenian language, 2024. URL https://huggingface.co/gordicaleksa/SlovenianGPT.

[9] S. Gugger, L. Debut, T. Wolf, P. Schmid, Z. Mueller, S. Mangrulkar, M. Sun, and B. Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9. Poster.

[11] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[12] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[13] I. Lebar Bajec, A. Repar, J. Demšar, Ž. Bajec, M. Rizvič, B. Kumperščak, and M. Bajec. Neural Machine Translation model for Slovene-English language pair RSDO-DS4-NMT 1.2.6, 2022. URL http://hdl.handle.net/11356/1736. Slovenian language resource repository CLARIN.SI.

[14] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

[15] P. H. Martins, J. Alves, P. Fernandes, N. M. Guerreiro, R. Rei, A. Farajian, M. Klimaszewski, D. M. Alves, J. Pombal, M. Faysse, P. Colombo, F. Yvon, B. Haddow, J. G. C. de Souza, A. Birch, and A. F. T. Martins. EuroLLM-9B: Technical report, 2025. URL https://arxiv.org/abs/2506.04079.

[16] NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846, June 2024. doi: 10.1038/s41586-024-07335-x. URL https://doi.org/10.1038/s41586-024-07335-x.

[17] NVIDIA. NVIDIA Riva translation overview, 2025. URL https://docs.nvidia.com/deeplearning/riva/user-guide/docs/translation/translation-overview.html.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 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, 2002. doi: 10.3115/1073083.1073135.

[19] P. Pattnaik, R. Maheshwary, K. Ogueji, V. Yadav, and S. T. Madhusudhan. Enhancing alignment using curriculum learning & ranked preferences. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12891–12907, 2024. doi: 10.18653/v1/2024.findings-emnlp.754.

[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.

[21] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

[22] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He. ZeRO: Memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020.

[23] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. COMET: A neural framework for MT evaluation, 2020. URL https://arxiv.org/abs/2009.09025.

[24] R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. Alves, L. Coheur, A. Lavie, and A. F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, 2022. URL https://aclanthology.org/2022.wmt-1.60/.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[26] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Gallouédec. TRL: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

[27] D. Vreš, M. Božič, A. Potočnik, T. Martinčič, and M. Robnik-Šikonja. Generative Model for Less-Resourced Language with 1 billion parameters. In *Language Technologies and Digital Humanities Conference*, 2024. URL https://www.sdjt.si/wp/wp-content/uploads/2024/09/JT-DH-2024_Vres_Bozic_Potocnik_Martincic_Robnik.pdf.

[28] G. Wang, H. Qin, S. A. Jacobs, C. Holmes, S. Rajbhandari, O. Ruwase, F. Yan, L. Yang, and Y. He. ZeRO++: Extremely efficient collective communication for giant model training. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=gx2BT0a9MQ.

[29] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 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, 2020. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[30] zIDsi. Wikipedia-Markdown, 2025. URL https://huggingface.co/datasets/zidsi/wikipedia_markdown.

Probing Vision-Language Understanding through the Visual Entailment Task: promises and pitfalls

Elena Pitta[1], **Tom Kouwenhoven**[1] **and Tessa Verhoef**[1,*]

[1]Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands

Abstract. This study investigates the extent to which the Visual Entailment (VE) task serves as a reliable probe of vision-language understanding in multimodal language models, using the LLaMA 3.2 11B Vision model as a test case. Beyond reporting performance metrics, we aim to interpret what these results reveal about the underlying possibilities and limitations of the VE task. We conduct a series of experiments across zero-shot, few-shot, and fine-tuning settings, exploring how factors such as prompt design, the number and order of in-context examples and access to visual information might affect VE performance. To further probe the reasoning processes of the model, we used explanation-based evaluations. Results indicate that three-shot inference outperforms the zero-shot baselines. However, additional examples introduce more noise than they provide benefits. Additionally, the order of the labels in the prompt is a critical factor that influences the predictions. In the absence of visual information, the model has a strong tendency to hallucinate and imagine content, raising questions about the model's over-reliance on linguistic priors. Fine-tuning yields strong results, achieving an accuracy of 83.3% on the e-SNLI-VE dataset and outperforming the state-of-the-art OFA-X model. Additionally, the explanation evaluation demonstrates that the fine-tuned model provides semantically meaningful explanations similar to those of humans, with a BERTScore F1-score of 89.2%. We do, however, find comparable BERTScore results in experiments with limited vision, questioning the visual grounding of this task. Overall, our results highlight both the utility and limitations of VE as a diagnostic task for vision-language understanding and point to directions for refining multimodal evaluation methods.

1 Introduction

In recent years, breakthroughs in Artificial Intelligence have driven substantial improvements in both Natural Language Processing and Computer Vision. While these domains were traditionally separate, the emergence of multimodal learning has unified them, allowing systems to interpret, reason, and produce meaning from combined textual and visual input. In this paper, we investigate whether a vision-language model can meaningfully combine information from visual and textual modalities in a visual entailment task. Visual Entailment is a multimodal task [38] that extends the traditional Textual Entailment (TE) task [4, 7]. In the TE task, given a text Premise P and a text Hypothesis H, the goal is to determine whether a premise implies some hypothesis. As such, the model tested outputs a label among three possible classes: *Entailment*, *Contradiction*, and *Neutral*, based on the relation derived from the text pair (P, H) [7, 4]. When there is sufficient evidence in P to conclude that H is true, then entailment holds. Wherever H contradicts P, a contradiction is identified. If not, the relation is neutral, suggesting that there is not enough data in P to infer anything from H. The difference between the TE and VE task is the replacement of the text premise with an image. The VE task is therefore multimodal as a model must predict by combining a visual premise with a textual hypothesis (Figure 1)

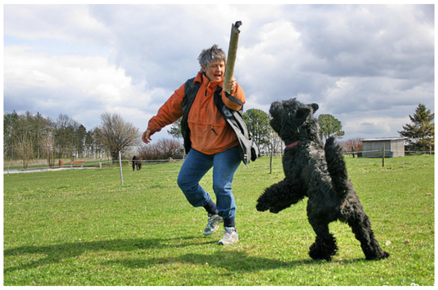


Figure 1. An example premise, hypothesis, model prediction, and explanation of the visual entailment task. **Hypothesis:** A woman carrying a stick. **Label:** Entailment. **Prompt 1 prediction:** Contradiction. **Explanation:** "The image provides sufficient evidence to confirm that the woman is indeed carrying a stick." **Prompt 2 prediction:** Entailment. **Explanation:** "The image shows a woman holding a stick, which is consistent with the description of a person carrying a stick. This suggests that the image supports or implies the truth of the hypothesis."

This paper aims to understand the capabilities and limitations of multimodal language models, using Llama 3.2 Vision as a test case, when performing the VE task and to investigate the factors that affect its performance. Through a series of experiments employing zero-shot, few-shot, and fine-tuning settings, we explore the promises and pitfalls of using the VE task to probe vision-language understanding. Specifically, we ask:

- How does Llama 3.2 Vision perform on the visual entailment task in a zero-shot inference setting, and what is the impact of having incomplete or absent visual input?
- What is the impact of few-shot inference on the performance, and how does it differ with different numbers of examples?
- How does the order of class labels in the prompt, and the order of examples in few-shot inference affect model predictions?
- To what extent does fine-tuning improve model performance compared to zero-shot and few-shot inference?

Comparing the performance of Llama 3.2 Vision across these different settings, we critically reflect on how well VE truly probes vision-language understanding.

[*] Corresponding Author. Email: t.verhoef@liacs.leidenuniv.nl

2 Background

Human intelligence is inherently multimodal, and learning often involves processing and integrating data from multiple senses. The central promise of multimodal language models is that they bring AI systems closer to human-like perception and understanding by combining the strengths of different modalities and providing a degree of language grounding [3, 21]. However, there are still a number of difficulties and challenges in this domain. A general challenge in multimodal learning involves the difficulty that accompanies learning how to represent and summarize multimodal data in a way that takes advantage of the complementarity and redundancy of multiple modalities [3]. Furthermore, transference [21], also called co-learning [3], referring to the ability to transfer knowledge between different modalities to aid the target modality, is still a core challenge. In models that combine vision and language, more specific limitations have been identified. For example, early work in Visual Question Answering (VQA) identified a heavy reliance on language priors where models often ignored visual input [12]. Similarly, it was demonstrated that VQA models may perform well by exploiting dataset shortcuts rather than truly grounding answers in the visual content [15, 1]. When evaluated on basic spatial relations (e.g., distinguishing "left of" vs "right of"), pre-trained models perform barely above chance, demonstrating their inability to represent spatial language robustly [16, 31]. In addition, vision-language models often struggle to correctly interpret interactions among objects and their attributes, and fail to visually distinguish pairs like "a red ball on a blue cube" and "a blue ball on a red cube" [32, 9], showing their poor visio-compositional reasoning abilities. Recent investigations, moreover, find conflicting evidence regarding the presence of human-like cross-modal associations in vision-language models [2, 33, 19].

Even when models appear to perform well on complex multimodal tasks, this does not necessarily mean that they are reasoning in a human-like way. Fair evaluation should therefore not dismiss mechanistic strategies of AI models like LLMs or vision-and-language models that differ from those present in humans [24] as they may rely on shortcut learning—the exploitation of spurious correlations or cues that happen to be present in a given dataset [26]. In Computer Vision, models may latch onto dataset artifacts in images during training, making it seem like a classifier was successfully trained, for example, to distinguish between horse and non-horse images. However, further analysis may reveal the model does not focus on horses in the images, but uses cues like copyright watermarks that only appeared in horse-labeled images [20]. In Natural Language Inference, models may perform well on benchmarks by exploiting syntactic heuristics, rather than actually understanding sentence meaning or logic [23]. Prior work has demonstrated that the Textual Entailment (TE) task, which acts as the precursor to VE, can largely be solved through the use of simple rules, such as assuming entailment if all the words in the hypothesis appear in the premise, rather than performing true semantic inference. Models that performed well on the standard TE dataset failed dramatically on a carefully constructed novel dataset with examples that can not be solved through sole reliance on these heuristics [23].

These limitations highlight a critical disconnect between task success and genuine understanding, and demonstrate the need for evaluations that go beyond standard performance metrics. Here, we comprehensively compare Llama 3.2 Vision performance across multiple experiments to assess whether the VE task, as evaluated on the e-SNLI-VE [17] dataset, effectively measures multimodal understanding.

3 Related Work

The VE task was introduced by Xie et al. [38], who proposed a model called **E**xplainable **V**isual **E**ntailment model (EVE). This model uses attention mechanisms to learn the inner relationships in both image and text feature spaces, and achieves better performance compared to other VQA-based models.

A major advancement in this field came with the OFA model (**O**ne **F**or **A**ll) [34]. OFA is a sequence-to-sequence learning framework and unifies various unimodal and cross-modal tasks, including the VE task. OFA achieves the state-of-the-art performance for the VE task on the SNLI-VE dataset (described in more detail in section 4.2) with an accuracy of 91.2% on the test set. Extending this, OFA-X [28] is a proposed multitask framework that predicts not only the labels but also explanations. OFA-X is a fine-tuned version of the OFA model and achieved the state-of-the-art performance for the VE task on the larger e-SNLI-VE dataset (also described in more detail in section 4.2) with an accuracy of 80.9% on the test set.

Perhaps the boldest perspective comes from an approach in which the proposed model CLOSE (**C**ross moda**L** transfer **O**n **S**emantic **E**mbeddings) can achieve a comparable performance, without images, using only textual input [13]. For the VE task, CLOSE uses the SNLI dataset for training (it uses a text premise instead of an image), while for evaluation, the SNLI-VE dataset was used, which combines vision and language. Despite not using images, CLOSE achieves similar performance to the image model. This suggests that the SNLI dataset may contain sufficient evidence to conclude the relationship without relying heavily on visual information. This raises questions about whether a visual grounding is required and hints at the previously mentioned concept of shortcut learning [26].

The knowledge from these previous works directly influenced the design of our experiments. Inspired by OFA, we adopted a prompt-based few-shot setup to investigate how effective a model performs without direct supervision. In addition, the idea of explanation generation in OFA-X led us to design an experiment to analyze the explanations from the model, helping assess its interpretability and reasoning. Finally, the innovative approach of the CLOSE model and its findings led us to test different experiments with limited vision to explore the extent to which our model depends on visual input. Comparing the performance of one state-of-the-art multimodal language model, Llama 3.2 Vision, across all these different settings allows us to analyze the suitability of the VE task to probe vision-language understanding.

4 Methodology

In this study, we evaluate the Llama 3.2 Vision 11B model on the e-SNLI-VE dataset using three approaches: zero-shot inference, few-shot inference, and fine-tuning.

4.1 Llama 3.2 Vision 11B

Llama 3.2 Vision[1] is a powerful multimodal large language model, available in two sizes: 11B and 90B parameters. The architecture of the model is based on the combination of the Llama 3.1 8B with a separately trained vision adapter [11]. During the training phase, the text model was frozen in order to preserve text-only performance [11]. The model was trained on 6 billion image-text pairs with a diverse data mixture [11]. Indicatively, the 11B parameter model achieved 75.2% accuracy on the VQAv2, a general visual question

[1] https://ollama.com/library/llama3.2-vision

answering benchmark, 91.1% on AI2 Diagram, a diagram understanding benchmark and 51.5% on MathVista (testmini), a mathematical reasoning benchmark[2].

Table 1. Overview of the e-SNLI-VE dataset of Do et al. [10].

Split	Train	Dev	Test
# Images	29,783	1,000	1,000
# Entailment	131,023	5,254	5,218
# Neutral	125,902	3,442	3,801
# Contradiction	144,792	5,643	5,721
# Total Labels	401,717	14,339	14,740

4.2 Dataset

The most common dataset used for the VE task is SNLI-VE (**S**tanford **N**atural **L**anguage **I**nference Corpus - **V**isual **E**ntailment)[3]. Specifically, this dataset is a combination of the SNLI (**S**tanford **N**atural **L**anguage **I**nference Corpus) and Flickr30k (image captioning dataset), where the premises from the SNLI are replaced with the corresponding images from Flickr30k [38]. This was feasible because the SNLI dataset was originally built using captioned images from the Flickr30k dataset, so textual premises in SNLI could be directly matched to the caption sentences of those photos [17].

Although the SNLI-VE dataset is the most common dataset for the VE task, recent research documented that 39% of the neutral labels in the validation and test sets were incorrectly labeled [17]. This happened mainly due to the replacement of the text premise with the image premise, which led to labeling errors, as an image typically contains more information than a single caption describing it [17]. Hence, the e-SNLI-VE (**E**xplainable **SNLI** - **V**isual **E**ntailment) dataset was created by merging SNLI-VE and e-SNLI (**E**xplainable **SNLI**). This yielded a visual entailment task with explanations in natural language. This specific dataset has better quality annotations due to hand-relabeling of validation and test sets. The e-SNLI-VE dataset has over 430k instances. Table 1 shows the dataset splits and the number of occurrences for each class in the sets. The dataset demonstrates a class imbalance, with contradiction being the most frequent class, followed by entailment with a slightly smaller number of occurrences, and neutral with the fewest cases (Table 1). While the e-SNLI-VE dataset provides explanations, the majority of our experiments focused only on classification. For the experiments in which explanations were considered, this is explicitly mentioned.

4.3 Experiments

4.3.1 Experiment 1: Zero-shot Inference

To establish a baseline and test how well Llama 3.2 Vision can perform VE without any additional training, we first test the model in a zero-shot setting. Here, the model is prompted to classify the image-hypothesis pair based on its pre-trained knowledge only. We used the prompt displayed in Prompt 1 to probe the model.

In addition to this prompt, we created variations in which only the order of the class labels (Entailment, Contradiction, Neutral) is varied, including all six possible permutations. Testing these different prompt variations allows us to assess whether, similar to text-only models [e.g., 35, 29], the model is sensitive to such variations and whether the predictions it makes are robust and internally consistent. We, moreover, introduced several manipulations to test the model's grounding. The first being the addition of explanations through changing the prompt to encourage rather than suppress this behavior (Prompt 3). This allows us to quantitatively test whether model-generated explanations align with those of humans and qualitatively observe why the model may make certain mistakes. Second, to test the model's reliance on visual information in the reasoning process, we evaluated it using limited visual input by either randomly cropping the images or replacing them with entirely black images.

```
Perform a visual entailment classification. You are
provided with two inputs:
1. Premise: An image described as follows (attached
below).
2. Hypothesis: A text description.

Your task is to classify the relationship between the
Premise (image) and Hypothesis (text) into one of the
following three categories:
- Entailment: The image provides enough evidence to
conclude that the Hypothesis is true.
- Contradiction: The image contradicts the Hypothesis.
- Neutral: The image does not provide enough information
 to determine the truth of the Hypothesis.

Provide a single classification in your response: one of
 Entailment, Contradiction, or Neutral. Do not include
explanations, commentary, or any additional text in your
 response.

[Insert hypothesis]
[Insert image]
```

Prompt 1. The zero-shot inference prompt. The Hypothesis and Premise are inserted at *Insert hypothesis* and *Insert image* respectively. In the case of few-shot inference, we add 3 or 6 randomly selected examples. Explanations are obtained through asking for additional justification. For examples see A.

4.3.2 Experiment 2: Few-shot Inference

To build on this further, and test whether in-context examples may improve the predictions of the model, we also conducted experiments with few-shot inference (Prompt 2). First, the model was provided with three randomly selected in-context examples from the training set (one example for each class), and we again experimented with varying the order of the class labels in the prompt (comparing Prompt 1 and Prompt 2), while also varying the order of the in-context examples to assess the impact of these factors on performance. Finally, motivated by the observation that increasing the number of examples can help models with better generalization and task performance [5], we expanded the number of examples in a six-shot inference setting.

4.3.3 Experiment 3: Fine-tuning

Finally, we fine-tuned the model on the VE task. For this, we utilized Unsloth[4] and QLoRA (**Q**uantized **Lo**w-**R**ank **A**daptation) [8] to reduce computing and memory requirements. We assessed both the classification ability and analysed the model's generated explanations. The model was fine-tuned for 1 epoch in each experiment. For the fine-tuning parameters and setup, see Appendix B

In all zero and few-shot experiments, the temperature parameter was set to 0 for deterministic output. Also, all results are based on a single run due to computational limitations. For all experiments, we measure the accuracy of class label prediction as well as F1, which is the harmonic mean of the metrics of precision and recall. We also compare model-generated explanations with those of humans, by calculating the BERTScore [39], because this measure is highly correlated with human evaluations and computes token similarity using contextual embeddings [39].

[2] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision

[3] https://github.com/maximek3/e-ViL/tree/main/data

[4] https://unsloth.ai/blog/vision

Table 2. Accuracy for zero-shot inference across six prompt variations.

Results for 6 prompts per instance	
Overall Accuracy	0.410
Majority Vote Accuracy	0.337

5 Results

5.1 Zero-shot Inference

Table 2 presents the overall results for the zero-shot experiment, in which each instance was evaluated in all six permutations of the class label order in the prompt. The overall accuracy shows how many predictions match with the ground truth, while the majority vote accuracy counts a prediction as correct *only* if at least four out of six outputs match the correct label. The overall accuracy is 41%, indicating that the model performs only slightly better than chance, and struggles to perform visual entailment in a zero-shot setting. In the majority vote scenario, we observe a drop in accuracy (33.7%) compared to the overall accuracy. This suggests that **the model frequently changes predictions for the same item across different prompts**, highlighting its sensitivity to the order of the labels in the prompt. To quantify this, and further explore the model's sensitivity to the prompt, Figure 2 reveals how often the model's prediction changes per sample across the six prompts. Almost half of the samples (7106) received the same prediction across all six prompts, which indicates that the model was fully consistent for those cases. However, 6647 samples had two different predictions, and 964 samples had even three different predictions, confirming that the model was inconsistent for a large number of cases. These results demonstrate the instability of the model's output under minimal modifications and explain the drop in majority vote accuracy.

To further explore how the order of the class labels in the prompt may affect the predictions of the model, Table 3 compares the overall accuracies and F1 scores as well as per-class F1 scores of two prompt variations. The first is Prompt 1, as displayed in section 4.3.1 and the second (Prompt 2) is the same except for the order of the class labels, which follows: Contradiction, Neutral, Entailment. The overall performance is similar between the two prompts, where Prompt 1 achieved an accuracy of 44.5%, while Prompt 2 achieved a slightly lower accuracy of 41.3%. In addition, focusing on the metrics per class for each prompt, we can conclude that the model over-predicts the entailment class in both cases. The neutral class has the worst per-class results in both prompts. Although weighted metrics were used to calculate the overall performance to ensure fairness among the imbalanced dataset, the fewer instances of the neutral class and the ambiguity that can occur have an impact on the ability of the model to correctly classify that class. Interestingly, the distribution of predictions among the classes differs between the two prompts. Concretely, Prompt 1 predicts 57.1% the entailment class, while in Prompt 2 this percentage is increased to 82.7%. Therefore, contradiction and neutral classes are predicted much less often than in Prompt 1. This observation suggests that **the order of class labels within the prompt significantly affects the predictions of the model**. The high rate of entailment predictions in Prompt 2 may reflect a kind of recency effect that is observed in text models [22, 25], where the last reported class label becomes more noticeable for the model.

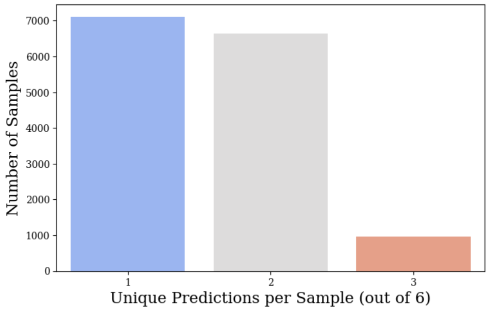


Figure 2. Consistency of model predictions across six prompts.

Table 3. Accuracy and F1 scores for zero-shot inference across Prompt 1 (class label order: Entailment, Contradiction, Neutral) and Prompt 2 (Contradiction, Neutral, Entailment).

	Prompt 1	**Prompt 2**
Overall Accuracy	0.445	0.413
Overall F1-score	0.409	0.319
Per-Class F1-score (Prediction %)		
Entailment	0.657 (57.1%)	0.587 (82.7%)
Neutral	0.232 (33.4%)	0.051 (10.8%)
Contradiction	0.299 (9.5%)	0.254 (6.5%)

Table 4. Results for zero-shot inference with randomly cropped images.

	Prompt 1	**Prompt 2**
Overall Accuracy	0.344	0.380
Overall F1-score	0.288	0.321
Per-Class F1-score (Prediction %)		
Entailment	0.535 (46.46%)	0.572 (58.60%)
Neutral	0.296 (50.75%)	0.268 (37.33%)
Contradiction	0.057 (1.45%)	0.128 (3.14%)

To further examine the reasoning behind the model's predictions, we observed the explanations given by the model in cases where the model predicts different outputs depending on the prompt. Figure 3 presents an example. One prediction matches the true label, while the other does not. However, both explanations can be considered logically correct, since the image indeed does not provide enough information to determine the truth of one of the two facts ("sewing her thumb") in the hypothesis. This indicates that, even when the prediction is incorrect, the model can still produce a reasonably logical explanation. Figure 1 shows a different type of error. Specifically, the class prediction does not align with the generated explanation. In the explanation, the model correctly stated that there is sufficient evidence to confirm the hypothesis, while the prediction is a Contradiction. Finally, Figure 4 raises questions about the ground truth labelling. In this case, while the explanations are logically correct and align with the class predictions, they do not align compared to the ground truth label, because the dataset may contain incorrect labels.

To investigate whether the predictions of the model are grounded in the visual information from the premise, we conducted additional zero-shot inference experiments with limited vision. First, we randomly cropped versions of the original images and used them as premises. As shown in Table 4, overall performance decreases compared to baseline results in Table 3. The decrease in accuracy was expected since the cropped images do not have all the details of the original images. This suggests the model does rely on the visual content to some extent, although the difference in performance is smaller than we may expect.

In the last experiment with zero-shot inference, the original images were entirely replaced with black images. If the models use visual information to make predictions, taking this information away

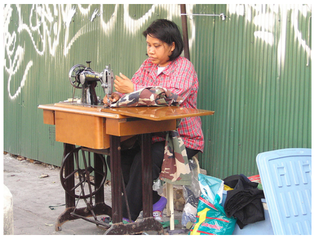


Figure 3. **Hypothesis:** Woman wearing nothing sewing her thumb. **Label:** Contradiction. **Prompt 1 prediction:** Neutral. **Explanation:**The image does not provide enough information to determine the truth of the Hypothesis, as it is unclear what the woman is actually doing or what she is wearing. The description ""sewing her thumb"" could be interpreted in various ways, and without more context or visual evidence, it is impossible to definitively classify the relationship between the image and the text." **Prompt 2 prediction:** Contradiction. **Explanation:** "The image depicts a woman wearing a red and white plaid shirt, not ""nothing"", while operating a sewing machine. This contradicts the text description of the woman being naked."

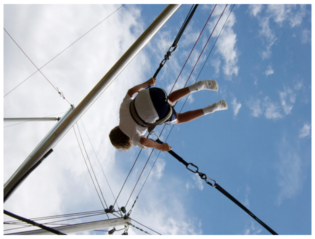


Figure 4. **Hypothesis:** a plane is flying **Label:** Entailment. **Prompt 1 prediction:** Contradiction. **Explanation:** "The image depicts a person on a swing, whereas the Hypothesis states that a plane is flying." **Prompt 2 prediction:** Neutral. **Explanation:** "The image shows a person on a swing, but it does not provide any information about planes or their flight status. Therefore, the relationship between the premise and hypothesis cannot be determined as either entailment or contradiction based solely on this image."

should result in a decrease in accuracy. The results not only reveal an expected drop in accuracy, but also **reveal a strong tendency of the model to hallucinate**. Table 5 shows the performance of the model, which is at chance. This is slightly worse than the performance of the baseline models. This is expected due to the removal of the model's visual ability to make accurate predictions, but given the stark contrast in available information between these two conditions (full image versus black image), the difference in performance is, again, actually quite small. This suggests **the model only makes limited use of the visual information**. As shown in the metrics per class, the model overpredicts the entailment class while it severely underpredicts the contradiction class (0.79% and 1.0% for Prompt 1 and Prompt 2, respectively). Moreover, the model imagines visual evidence to support the hypothesis, which justifies the high proportions of the entailment class. Figures 5 and 6 illustrate two examples that clearly show the tendency of the model to hallucinate.

Table 5. Results for zero-shot inference with black images as premises.

	Prompt 1	**Prompt 2**
Overall Accuracy	0.360	0.369
Overall F1-score	0.250	0.246
Per-Class F1-score (Prediction %)		
Entailment	0.520 (**84.17%**)	0.531 (**89.95%**)
Neutral	0.211 (14.67%)	0.161 (8.80%)
Contradiction	0.031 (0.79%)	0.043 (1.00%)

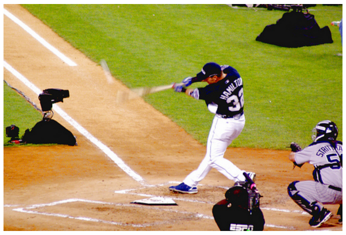


Figure 5. *The original image premise is shown, but the model saw an entirely black image.* **Hypothesis:** 2 basketball players **Label:** Contradiction. **Prompt 1 prediction:** Entailment. **Explanation:** "The image shows two individuals wearing basketball jerseys, which is consistent with the description of basketball players." **Prompt 2 prediction:** Entailment. **Explanation:** "The image shows two individuals wearing basketball jerseys, which is consistent with the description of basketball players."

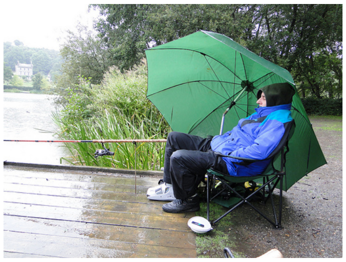


Figure 6. *The original image premise is shown, but the model saw an entirely black image.* **Hypothesis:** the man is inside washing dishes **Label:** Contradiction. **Prompt 1 prediction:** Entailment. **Explanation:** "The image shows a person standing in front of a sink, actively engaged in washing dishes. This visual evidence directly supports the statement that the man is indeed inside and washing dishes." **Prompt 2 prediction:** Entailment. **Explanation:** "The image shows a person standing in front of a sink, actively engaged in washing dishes. This visual evidence directly supports the statement that the man is indeed inside and washing dishes."

5.2 Few-shot Inference

Table 6 demonstrates the results of the three-shot inference experiment. While the results are overall still not much better than the baseline zero-shot findings, we do see a slight improvement. Concretely, the best accuracy and F1-score for zero-shot is 44.5% (Prompt 1) and 40.9%, while the best performance for three-shot (Prompt 2, and contradiction as the first example) is 48.7% and 42.6%, respectively. The improvement in the balance by class and F1 score for the three-shot inference, particularly for the contradiction class, suggests a more robust understanding of the task, although the increase in accuracy is very modest.

Regarding the order of the three in-context examples, we can infer that it has a considerable influence on the outcome. Experiments demonstrate that **the first example in the few-shot setting has a large impact on the predictions of the model**, with a notably higher accuracy and F1 score in the case where the first in-context example was one where Contradiction was the true label. The model performs the best in the experiment with Prompt 2 (which also has the class of contradiction as the first in order in the prompt). Placing contradiction first in the in-context examples may cause a primacy bias that helps mitigate the model's strong bias toward predicting the entailment class in the corresponding zero-shot scenario.

Specifically, the model significantly overpredicts the entailment class in zero-shot results (Prompt 2 yields it in over 80% of cases). On the other hand, three-shot inference counteracts this bias, resulting in seemingly more balanced class predictions. Moreover, when

Table 6. Results for three-shot inference across varying in-context example orders (CEN, ECN, NEC) and two Prompts (Pr.).

Exa. Order	Pr.	Acc.	F1	Class Metrics (F1/Pred%)		
				Ent.	Neu.	Con.
CEN	1	0.47	0.45	0.59/56.9%	0.14/16.0%	0.53/27.1%
	2	**0.49**	0.43	0.59/71.8%	0.05/5.2%	0.52/23.0%
ECN	1	0.41	0.37	0.55/69.5%	0.15/18.7%	0.35/11.7%
	2	0.43	0.38	0.56/73.4%	0.12/13.0%	0.39/13.0%
NEC	1	0.42	0.37	0.59/66.0%	0.22/25.7%	0.27/8.4%
	2	0.43	0.38	0.59/67.4%	0.20/24.0%	0.29/8.6%

comparing the class metrics, the order of the class labels in the prompt seems to have a less severe effect on the prediction when the model has been given three in-context examples, indicating that few-shot learning provides a stabilizing influence on class prediction.

Given the modest benefits of providing three in-context examples, we next explored the influence of providing more examples, six instead of three. This time we do not explicitly compare different in-context example orderings, but we create an order that is relatively unbiased by making sure each correct class appears once in the first three and once in the second three examples, the class's order of the first three examples is different from the order of the last three examples, and the first and last examples are not in the same class. The results of the six-shot inference are shown in Table 7. When comparing the results of the six-shot experiment with those of the three-shot experiment, we can extract some important insights. Firstly, **the performance does not consistently improve with more in-context examples**. The best performance of six-shot (36.5%) is actually lower than the best performance of three-shot (48.7%). This suggests that improved performance in few-shot experiments may not always reflect a more in-depth understanding of the task. Instead, differences in predictions may be a result of biases and sensitivities to example orderings as well as overfitting to the dominant order. Secondly, the metrics per class show that the model overpredicts the neutral class for both prompts, and it is the dominant class with over 70% classified as neutral. Therefore, in the six-shot experiment, there seems to be a specific class bias that we did not observe in other settings.

Compared to zero-shot, six-shot inference has a slightly more balanced performance per class, as reflected by the increase in F1-score for the Contradiction and Neutral class, but results in lower overall accuracy. Specifically, zero-shot achieves an accuracy of 44.5% while six-shot achieves an accuracy of 36.5%. This indicates that providing in-context examples may in some case hurt rather than help. This inconsistency is difficult to explain while holding the assumption that the model is solving the VE task through human-like vision-language understanding.

Table 7. Results for six-shot inference.

	Prompt 1	Prompt 2
Overall Accuracy	0.350	0.365
Overall F1-score	0.319	0.356
Per-Class F1-score (Prediction %)		
Entailment	0.240 (9.4%)	0.373 (20%)
Neutral	0.405 (79.6%)	0.383 (70.5%)
Contradiction	0.333 (11%)	0.322 (10%)

Table 8. Results for the fine-tuned model with Prompt 1.

Metric	Value
Overall Performance	
Accuracy	**0.833**
F1-score	0.836
Per-Class F1-score (Prediction %)	
Entailment	0.864 (35.88%)
Neutral	0.737 (30.67%)
Contradiction	0.876 (33.45%)

Table 9. BERTScore results for explanation evaluation.

Metric	Recall	Precision	F1 Score
Fine-tuned model - Prompt 1	0.8869	0.8968	**0.8916**
Zero-shot - Prompt 1	0.8775	0.8549	0.8659
Zero-shot - Prompt 2	0.8805	0.8574	0.8686
Black Images - Prompt 1	0.8798	0.8624	0.8709
Black Images - Prompt 2	0.8798	0.8634	0.8714

5.3 Fine-tuning

Table 8 illustrates the classification results of the fine-tuned model, which achieved a high overall accuracy of 83.3%, and an F1 score of 83.6%. These results indicate that the model generalizes well across the three classes. The most challenging class is Neutral, even for the fine-tuned model. When compared to zero- and few-shot experiments, the fine-tuned model shows a significant improvement in both general and class-specific performance. Moreover, **the Llama 3.2 Vision fine-tuned model outperforms the state-of-the-art model** OFA-X, which achieved an accuracy of 80.9%.

Table 9 shows the evaluation of the generated explanations. According to the BERTScore, the model achieves an F1-score of 89.16%, indicating that the generated explanations are semantically similar to the human produced reference explanations, even if they differ in the exact words. However, in Table 9 we also report the same measure for the experiments with zero-shot inference and black images, and the results are very similar. This suggests that explanations with a high BERTScore may not necessarily reflect the model is reasoning in a human-like way.

6 Discussion

This study investigates the capabilities of the Llama 3.2 Vision model on the VE task using the e-SNLI-VE dataset. The experiments yielded various findings, revealing to what extent VE is a suitable task to probe vision-language understanding. First, the baseline results demonstrated modest performance, indicating the limited capabilities of the model in zero-shot inference. This is somewhat surprising given the enormous number of images and textual captions the model has seen in training and the impressive performance reported for other vision-language tasks such as visual question answering. Three-shot inference improves the performance of the model; however, we also observe that additional in-context examples are not always beneficial. The most significant finding is the major improvement after fine-tuning, where the model achieved an accuracy of 83.3%, outperforming the SOTA performance achieved by the OFA-X model. Moreover, the fine-tuned model has strong interpretability since it achieved an F1-score of 89.16% using BERTScore, an evaluation metric that utilizes contextual embeddings for the explanation evaluation. This indicates high semantic similarity between the human produced reference and model generated text. While these results are promising, the overall findings also reveal some pitfalls in using the VE task and e-SNLI-VE dataset to effectively measures multimodal understanding. In the zero-shot inference experiments with limited or absent vision, we saw that the model was highly prone to hallucination and imagined visual evidence in order to support the hypotheses. The experiments with prompt variations and three-shot inference reveal that factors such as the order of the class labels in the prompt and the order of the in-context examples significantly affect the model's predictions, revealing highly inconsistent reasoning, which does not align with the assumption that the model shows vision-language understanding in a human-like way. Also, the BERTScore results for the zero-shot inference experiments with black images were on par with those of the fine-tuned model, showing that model generated explanations with semantic similarity to human explanations do not necessarily reveal the model is effectively using the visual input to solve the task. Finally, the observation of individual errors in the zero-shot experiment exposed problems with the e-SNLI-VE dataset, which still contains examples with wrong labels or examples that can be interpreted in multiple ways, technically making more than one class label correct. Before including VE in broader benchmarks used for training and testing in the area of general multi-modal reasoning (as already the case in [34] for example) we recommend further investigation into these issues.

These findings additionally offer several lessons for the broader field of multimodal learning and understanding. In particular, the study underscores that, while general pre-training is powerful, even advanced multimodal language models such as Llama 3.2 Vision may not be suitable for complex reasoning tasks like VE without special adaptation. The few-shot results underline that a deeper understanding of how models utilize context is needed, for example, by interpreting their attention patterns using Grad-CAM [30]. Additionally, the study highlights that the effectiveness of in-context learning depends on the number and ordering of examples. This bears much resemblance to known consistency effects in LLMs, which heavily depend on prompt ordering [36]. The dramatic increase in performance after fine-tuning exposes that the model's visual and linguistic embeddings are highly adaptable and are, in principle, rich enough for visual entailment.

The findings provide helpful insights into the VE capabilities of Llama 3.2 Vision, but there are some limitations that should be noted. First, our methodology relies on generated answers for both the class label and the explanation. The former of which is somewhat debatable since generated multiple-choice answers are often inconsistent with actual model beliefs [35, 18]. While this may affect the observed results, and could be alleviated by prefilling class options and selecting the most likely class [14], the modus operandi of commercially available and deployed models is to use generation, i.e., without prefilling. As such, our results should be seen through this lens, and we see extended analyses using log probabilities as future work.

Second, every experiment was evaluated once because of time and computational constraints. The metrics are not averaged over multiple runs. This affects the few-shot experiments where a different random selection of in-context example could yield a different performance. Another limitation lies in the restricted experiments for the few-shot inference. A small number of configurations were tested, particularly for the six-shot inference, which included just one permutation. Several possible combinations are left out. However, given the issues found with biases, sensitivity to order effects, and hallucinations, strong improvements for the right reasons are unlikely. In addition, the fine-tuning was conducted using only the first prompt. However, we expect that predictions will not be greatly affected by the order of the classes in the prompt, given the significant performance gain observed by the fine-tuning.

A worthwhile direction for future work would be to further investigate few-shot inference. For example, exploring different sets of examples for each strategy in three shots, examining different orderings of classes for six shots, and testing a larger number of examples within a context, such as fifteen shots, could still be valuable, not primarily to focus on performance, but to gain deeper insights, such as understanding the threshold beyond which providing more examples becomes disadvantageous. Given its ability to improve many reasoning tasks, another promising direction is to integrate Chain-of-Thought prompting [37] into few-shot and zero-shot inference. This perhaps extends the models' already observed tendency to produce coherent explanations and better use these in predictions. A broader direction for future work includes systematic prompt engineering. This involves improving the wording and structure of the prompts. Since this study demonstrates that the design of the prompts significantly affects the predictions, optimizing the prompts could perhaps lead to better generalization and fewer hallucinations.

Finally, our results need to be corroborated by investigating other, perhaps larger, models. Doing so enables careful comparison between, for example, architectural, data, and optimization design decisions, informing which ingredients improve visual entailment. In a similar vein, earlier work investigating whether model representations align with human representations suggests that dataset diversity and scale are the primary drivers of alignment [6, 27].

7 Conclusion

In conclusion, we used the Llama 3.2 Vision model to explore the possibilities and limitations of using the Visual Entailment task to probe vision-language understanding. A comparison of results in zero-shot, few-shot, and fine-tuning settings as well as experiments involving limited vision and prompt sensitivity analyses together revealed several problems. These included inconsistent reasoning, a limited reliance on visual information and a strong tendency to hallucinate. These findings underscore the importance of critical investigations into benchmark and dataset quality to make sure the predictions of the model actually reflect vision-language reasoning instead of an exploitation of spurious correlations. Future work is necessary to further explore what causes the substantial difference in performance between zero-shot and fine-tuned settings and what kind of heuristics the model may be learning from the dataset during fine-tuning. This would help to further develop the VE task into a suitable method for probing vision-language understanding in multi-modal language models.

8 Ethics Statement

This research involves evaluating and fine-tuning a publicly available multimodal language model (LLaMA 3.2 Vision) on the Visual Entailment task using a benchmark dataset (e-SNLI-VE). This dataset is publicly available and contains no personally identifiable or sensitive information. No new data involving human subjects was collected.

The aim of this work is not only to assess model performance but to critically interrogate what this performance reveals about vision-language understanding. In doing so, we identify concerning behaviors such as hallucination, over-reliance on linguistic priors, and sensitivity to prompt structure, underscoring the risks of interpreting accuracy metrics as indicators of genuine multimodal reasoning. This research is intended to contribute to the responsible development and evaluation of vision-language models.

We acknowledge the broader societal risks associated with the development and deployment of multimodal language models, including the potential propagation of biases and misleading explanations. Future applications of this work should consider the risks associated with deploying multimodal models in sensitive domains, especially where explainability and factual grounding are critical and misplaced trust in model outputs could have real-world consequences.

This research contributes to the growing environmental impact of AI. While our experiments were limited in scope compared to model pretraining, they nonetheless required significant computational resources. We believe it is important to reflect on how the field can pursue vision-language understanding more sustainably.

References

[1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018.

[2] M. Alper and H. Averbuch-Elor. Kiki or bouba? sound symbolism in vision-and-language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78347–78359. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f74054328beeb0c21a9b8e99da557f5a-Paper-Conference.pdf.

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[4] S. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[6] C. Conwell, J. S. Prince, K. N. Kay, G. A. Alvarez, and T. Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023. doi: 10.1101/2022.03.28.485868. URL https://www.biorxiv.org/content/early/2023/07/01/2022.03.28.485868.

[7] I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer, 2006.

[8] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36:10088–10115, 2023.

[9] A. Diwan, L. Berry, E. Choi, D. Harwath, and K. Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.143. URL https://aclanthology.org/2022.emnlp-main.143/.

[10] V. Do, O.-M. Camburu, Z. Akata, and T. Lukasiewicz. e-SNLI-VE: Corrected visual-textual entailment with natural language explanations, 2021. URL https://arxiv.org/abs/2004.03744.

[11] H. Face. What is llama 3.2 vision?, 2025. URL https://huggingface.co/blog/llama32?utm_source=chatgpt.com#what-is-llama-32-vision. Accessed: 2025-02-25.

[12] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

[13] S. Gu, C. Clark, and A. Kembhavi. I can't believe there's no images!: Learning visual tasks using only language supervision. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2672–2683. IEEE, 2023.

[14] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[15] A. Jabri, A. Joulin, and L. Van Der Maaten. Revisiting visual question answering baselines. In *European Conference on Computer Vision*, pages 727–739. Springer, 2016.

[16] A. Kamath, J. Hessel, and K.-W. Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.568. URL https://aclanthology.org/2023.emnlp-main.568/.

[17] M. Kayser, O.-M. Camburu, L. Salewski, C. Emde, V. Do, Z. Akata, and T. Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1224–1234. IEEE, 2021.

[18] A. Khatun and D. G. Brown. A study on large language models' limitations in multiple-choice question answering, 2024. URL https://arxiv.org/abs/2401.07955.

[19] T. Kouwenhoven, K. Shahrasbi, and T. Verhoef. Cross-modal associations in vision and language models: Revisiting the bouba-kiki effect, 2025. URL https://arxiv.org/abs/2507.10013.

[20] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

[21] P. P. Liang, A. Zadeh, and L.-P. Morency. Foundations & Trends in Multimodal Machine Learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

[22] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

[23] R. T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

[24] R. Millière and C. Rathkopf. Anthropocentric bias and the possibility of artificial cognition. In *ICML 2024 Workshop on LLMs and Cognition*, 2024. URL https://openreview.net/forum?id=wrZ6mLelzu.

[25] M. Mina, V. Ruiz-Fernández, J. Falcão, L. Vasquez-Reina, and A. Gonzalez-Agirre. Cognitive biases, task complexity, and result intepretability in large language models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1767–1784, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.120/.

[26] M. Mitchell and D. C. Krakauer. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

[27] L. Muttenthaler, J. Dippel, L. Linhardt, R. A. Vandermeulen, and S. Kornblith. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=ReDQ1OUQR0X.

[28] B. Plüster, J. Ambsdorf, L. Braach, J. H. Lee, and S. Wermter. Harnessing the power of multi-task pretraining for ground-truth level natural language explanations, 2023. URL https://arxiv.org/abs/2212.04231.

[29] E. S. Salido, J. Gonzalo, and G. Marco. None of the others: a general technique to distinguish reasoning from memorization in multiple-choice llm evaluation benchmarks, 2025. URL https://arxiv.org/abs/2502.12896.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via

gradient-based localization. *Int. J. Comput. Vision*, 128(2):336–359, Feb. 2020. ISSN 0920-5691. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007/s11263-019-01228-7.

[31] F. Shiri, X.-Y. Guo, M. G. Far, X. Yu, R. Haf, and Y.-F. Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1195. URL https://aclanthology.org/2024.emnlp-main.1195/.

[32] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, June 2022.

[33] T. Verhoef, K. Shahrasbi, and T. Kouwenhoven. What does kiki look like? cross-modal associations between speech sounds and visual shapes in vision-and-language models. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, and Y. Oseki, editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 199–213, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.cmcl-1.17. URL https://aclanthology.org/2024.cmcl-1.17/.

[34] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[35] X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, and B. Plank. “my answer is C”: First-token probabilities do not match text answers in instruction-tuned language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

[36] L. Weber, E. Bruni, and D. Hupkes. Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. In J. Jiang, D. Reitter, and S. Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 294–313, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.20. URL https://aclanthology.org/2023.conll-1.20/.

[37] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[38] N. Xie, F. Lai, D. Doran, and A. Kadav. Visual entailment task for visually-grounded language learning, 2019. URL https://arxiv.org/abs/1811.10582.

[39] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

A Prompts

Example prompt used to assess three-shot and six-shot performance on the VE task. In these cases, we add random images accompanied by their hypotheses and the gold label, such that the model can possibly deduce what is important to make correct predictions. Prompt 2 shows a three-shot example. In the case of six-shot, we add three additional examples. Prompt 3 is used to obtain model explanations.

```
Perform a visual entailment classification. You are
provided with two inputs:
1. Premise: An image described as follows (attached
below).
2. Hypothesis: A text description.

Your task is to classify the relationship between the
Premise (image) and Hypothesis (text) into one of the
following three categories:
- Entailment: The image provides enough evidence to
conclude that the Hypothesis is true.
- Contradiction: The image contradicts the Hypothesis.
- Neutral: The image does not provide enough information
 to determine the truth of the Hypothesis.

Provide a single classification in your response: one of
 Entailment, Contradiction, or Neutral. Do not include
explanations, commentary, or any additional text in your
 response.

[Example Hypothesis 1]
[Example Image 1]
[Example Gold label 1]

[Example Hypothesis 2]
[Example Image 2]
[Example Gold label 2]

[Example Hypothesis 3]
[Example Image 3]
[Example Gold label 3]

[Hypothesis]
[Image]
```

Prompt 2. The prompt used to assess three-shot inference performance. The example hypothesis, image, and gold label are randomly picked. After observing these examples, the model is tasked to predict entailment for the final *Hypothesis* and *Image*.

```
Perform a visual entailment classification. You are
provided with two inputs:
1. Premise: An image described as follows (attached
below).
2. Hypothesis: A text description.

Your task is to classify the relationship between the
Premise (image) and Hypothesis (text) into one of the
following three categories:
- Entailment: The image provides enough evidence to
conclude that the Hypothesis is true.
- Contradiction: The image contradicts the Hypothesis.
- Neutral: The image does not provide enough information
 to determine the truth of the Hypothesis.

Format your response as follows:
Label: <Entailment/Contradiction/Neutral>
Explanation: <Brief justification>

[Hypothesis]
[Image]
```

Prompt 3. The prompt used to obtain explanations.

B Fine-tuning

Additional information on fine-tuning is presented in Table 10, displaying LoRA configuration parameters, and Table 11, showing all hyperparameters used in training.

Table 10. LoRA Configuration Parameters.

Parameter	**Value**
PEFT Method	LoRA
Finetune Vision Layers	False
Finetune Language Layers	True
Target Modules	Attention & MLP
Rank (r)	8
Alpha (α)	16
Dropout	0
Bias	"none"
Random state	3407

Table 11. Training Hyperparameters.

Parameter	**Value**
Library	TRL SFTTrainer
Epochs	1
Max Sequence Length	2048
Learning Rate	2e-4
LR Scheduler	Linear
Warmup Steps	5
Optimizer	AdamW (8-bit)
Weight Decay	0.01
Train Batch Size (per device)	2
Gradient Accumulation Steps	4
Seed	3407

Evaluation, Judgment, and Public Discourse

Do Large Language Models understand how to be judges?

Nicoló Donati[a,*], **Giuseppe Savino**[b] **and Paolo Torroni**[c]

[a,c]University of Bologna
[a,b]Zanichelli editore S.p.A.
ORCID (Nicoló Donati): https://orcid.org/0009-0000-5673-5274, ORCID (Paolo Torroni): https://orcid.org/0000-0002-9253-8638

Abstract. This paper investigates whether Large Language Models (LLMs) can effectively act as judges for evaluating open-ended text generation tasks, such as summarization, by interpreting nuanced editorial criteria. Traditional metrics like ROUGE and BLEU rely on surface-level overlap, while human evaluations remain costly and inconsistent. To address this, we propose a structured rubric with five dimensions: coherence, consistency, fluency, relevance, and ordering, each defined with explicit sub-criteria to guide LLMs in assessing semantic fidelity and structural quality. Using a purpose-built dataset of Italian news summaries generated by GPT-4o, each tailored to isolate specific criteria, we evaluate LLMs' ability to assign scores and rationales aligned with expert human judgments. Results show moderate alignment (Spearman's ρ = 0.6–0.7) for criteria like relevance but reveal systematic biases, such as overestimating fluency and coherence, likely due to training data biases. We identify challenges in rubric interpretation, particularly for hierarchical or abstract criteria, and highlight limitations in cross-genre generalization. The study underscores the potential of LLMs as scalable evaluators but emphasizes the need for fine-tuning, diverse benchmarks, and refined rubrics to mitigate biases and enhance reliability. Future directions include expanding to multilingual and multi-genre contexts and exploring task-specific instruction tuning to improve alignment with human editorial standards.

1 Introduction

Evaluating open-ended text generation depends on a set of often implicit criteria that are hard to formalise. Traditional metrics like ROUGE [21], BLEU [27], and METEOR [4] reduce evaluation to surface-level overlap, overlooking deeper qualities such as semantic fidelity and target-audience relevance [29, 5]. Human judgments capture these nuances but are costly, inconsistent, and difficult to scale [24, 11]. Large language models (LLMs) offer a potential solution: given a clear, multi-item criterion, they could score each criterion and supply a rationale, promising consistency and low cost. Yet this hinges on whether LLMs actually *understand* the criterion's language and hierarchy. For example, when asked to evaluate summaries, can an LLM reliably distinguish between objective bullets (e.g., Does this summary include every key claim?) and subjective ones (e.g., Is the tone appropriate?), and combine them into a coherent overall score? In this paper, we test several LLMs using few-shot prompts that supply explicit criteria drawn from editorial best practices. For each generated summary, the model must: (1) assign scores for each criterion, (2) explain its score based on the given criteria. By comparing these outputs with expert human judgments and within different LLMs, we measure: (a) alignment between LLM and human scores per criterion, (b) faithfulness of LLM rationales to the rubric versus reliance on superficial cues, and (c) alignment between different LLMs.

2 Related Works

The use of large language models (LLMs) as evaluative judges has emerged as a prominent methodology for assessing AI-generated outputs. These systems can be broadly classified into three categories: prompted judges, fine-tuned judges, and multi-agent judges. Prompted judges rely on the intrinsic capabilities of LLMs, activated through carefully engineered prompts, without requiring additional training [39, 18, 7]. Fine-tuned judges, in contrast, are explicitly trained on specialized preference datasets to enhance their evaluation precision [13, 14, 35, 40, 17]. These models are often fine-tuned using data sourced from human annotations or distilled judgments from advanced models like GPT-4 [26]. Despite their robust performance on benchmarks, fine-tuned judges frequently fail to generalize effectively across diverse or unfamiliar tasks, as noted by [12]. This limitation arises partly because the datasets used for fine-tuning typically lack sufficiently complex examples, thereby constraining the reasoning capabilities of these judges. Finally, multi-agent judges employ a collaborative approach, leveraging the outputs of multiple LLMs in a sequential or ensemble framework to generate judgments [3, 6, 33]. Although this approach offers enhanced evaluation robustness by surpassing the abilities of a single model, it incurs significantly higher computational costs during inference. As LLM-based judges gain widespread adoption for evaluating and refining large language models, numerous benchmarks have been developed to assess their performance. Prominent examples include LLMEval [38], MTBench [39], and FairEval [34], which emphasize alignment between LLM-based judges' assessments and human evaluations. However, these benchmarks are often constrained by the subjectivity inherent in human evaluation, which can prioritize stylistic elements over factual and logical accuracy. In response, LLMBar [37] introduces a methodology that evaluates judges' ability to adhere to instructions, employing response pairs with clear ground-truth preference labels.

[*] Corresponding Author. Email: n.donati@unibo.it

Conversely, JudgeBench focuses on more complex tasks, such as evaluating reasoning capabilities and distinguishing between correct and incorrect responses, surpassing the scope of simple instruction-following tasks. It generates challenging response pairs for evaluating LLM-based judges using a robust pipeline to transform datasets with ground truth labels into pairs where one response is correct and the other is not. The pipeline ensures stylistic consistency and mitigates biases like self-enhancement. It filters out questions where all responses are entirely correct or incorrect, making it harder for LLM judges to distinguish. JudgeBench can adapt diverse datasets, including Knowledge, Reasoning, Mathematics, and Coding. For reward models, RewardBench [16] provides a comprehensive evaluation across domains such as safety, dialogue, and reasoning. This benchmark aggregates multiple preference datasets and prior benchmarks [32, 1, 8, 2, 37, 19, 25, 36, 30, 20, 39], enabling a holistic assessment of reward models' performance.

3 Method

To investigate whether LLMs can "understand" and apply given evaluation criteria, we designed an experimental setup that centers on how well an LLM internalises and operationalises each rubric item. Rather than relying on off-the-shelf summarization benchmarks (whose reference summaries are often misaligned with editorial standards, prone to test contamination, and insufficiently detailed), we constructed a purpose-built corpus and rubric explicitly tailored for probing criterion interpretation.

Evaluation Criteria Drawing on best practices from professional editors, we defined five distinct criteria. Each is formulated not merely as a high-level goal, but with clear definitions and rating anchors to encourage models to parse and apply the intended semantics, rather than latch onto surface patterns:

- **Coherence:** The summary should present its information with a clear, logical progression. Sentences must flow seamlessly, avoiding abrupt shifts or disconnected fragments. The model must recognise when the content is organised into a unified narrative versus a "heap" of related but unstructured statements.
- **Consistency:** This goes beyond detecting hallucinations; it requires verifying that every factual claim in the summary is entailed by the source article. Models must check that no key fact is contradicted or misrepresented, and that no extraneous details are introduced.
- **Fluency:** The writing should be grammatically correct and stylistically smooth. Here, the model must evaluate spelling, punctuation, and phrasing quality, not just surface token distributions.
- **Relevance:** Every sentence in the summary should focus on the article's core points, omitting trivial or tangential information. The model has to distinguish between essential content (e.g., major events, central arguments) and filler.
- **Ordering:** Key points must appear in the same logical sequence as the original article, preserving narrative structure. A well-ordered summary guides the reader through the source's flow; a misordered one, even if factually accurate, disrupts coherence.

We contrast our rubric with existing automated evaluation schemes (e.g. G-Eval [22]) by emphasizing how each bullet is designed to force the model to interpret nuanced language and hierarchical dependencies. For instance, whereas G-Eval's "coherence" may loosely reward sentence quality, our definition requires explicit assessment of how information is organized. Similarly, rather than narrowly flagging hallucinations under "consistency," our entailment-based framing demands that the model verify support for every factual claim.

Dataset Development . To prevent test contamination and ensure that model judgments truly reflect criterion understanding (rather than memorized patterns), we selected 10 Italian news articles published after the training cutoffs of all target models. For each article, we used a controlled prompting procedure with GPT-4o to generate multiple summaries that each exhibit a predefined level of quality for one of the rubric items, producing 50 different summaries. Prompts specified length and required adherence to preassigned thresholds for coherence, consistency, fluency, relevance, and ordering. For example, to create a low-coherence summary, the prompt asked GPT-4o to shuffle logical segments while preserving factual accuracy; for high-consistency but low-fluency summaries, the prompt enforced fact verification but allowed awkward phrasing. All generated summaries were then reviewed by humans, who checked if the quality of the summary was in line with the guidelines in Appendix B and made modifications accordingly. Finally, an expert annotator (following the Evaluation Guidelines in Appendix B) annotated every summary with 1-5 discrete scores for each of the five criteria. We collected 250 evaluations across the 50 summaries that were generated and human-validated.

Evaluation Framework Our primary question is: can an LLMs, when prompted with this rubric, understand the meaning of the rubric and output scores and rationales that mirror the annotators' judgments? Each model is tasked with the evaluation of a summary based on the rubric. Then the agreement with the human expert is computed. By focusing on how each LLM interprets and applies the five editorial criteria, this framework highlights not only whether models can approximate human judgments but also reveals which rubric items (objective vs. subjective, hierarchical vs. flat) they struggle to internalize.

4 Experiments

We evaluated twelve SLMs without applying any fine-tuning or soft-tuning, relying solely on a few-shot prompting approach to ensure a fair comparison. Each model was prompted using a standardized template (Appendix A) designed to guide assessments based on five established editorial criteria. For each criterion, the prompt instructed the model to act as an impartial evaluator, assigning a score from 1 to 5 and providing a detailed explanation in Italian that justifies the rating. The prompt includes the definition of the criteria and detailed descriptions for each score level to standardize expectations. Each criterion was further defined through a set of sub-criteria that specify key aspects for evaluation. The expected output followed a predefined JSON format, requiring both the numerical score and a rationale. SummEval[9], a meta-evaluation dataset for summarization, was used to construct 15 few-shot examples for coherence, consistency, fluency, and relevance to guide model predictions. For the ordering criterion, we generated synthetic summaries using GPT-4o to illustrate varying levels of quality. We also generate explanations that are manually reviewed to ensure alignment with the editorial criteria. This setup enabled an intrinsic evaluation of the models' ability to assess summarization quality independently of external training data. For comparison, we also evaluated selected LLMs under the same conditions to establish an upper performance bound.

Experimental Setup All experiments were conducted on a system equipped with an NVIDIA RTX A6000 GPU. The models were accessed via the Hugging Face model hub and inferred using the Transformers library and parameters suggested by the model's authors. This setup enabled reproducible evaluation of model responses across criteria.

Meta-Evaluation Metrics Model performance was evaluated using two primary metrics: Spearman's rank correlation coefficient (ρ) and Mean Absolute Error (MAE). Spearman's ρ was used to measure the ordinal alignment between model-generated judgments and human ratings, capturing the models' ability to rank summaries in accordance with expert evaluations. MAE, on the other hand, quantifies the average deviation of model predictions from human scores, providing insight into absolute accuracy. These metrics were chosen to comprehensively assess both the relative ranking capabilities and the precision of the models.

5 Results

The analysis of model performance across various families, illustrated in Figures 1, 2, and 3, reveals that scaling effects are not uniform and depend on both the model family and the specific metric considered.

For example, the deepseek models (red markers in Figure 1) demonstrate a clear reduction in mean absolute error (MAE) as model size increases, as depicted in Figure 2. The 1.5B parameter model exhibits an MAE of 1.47, which improves to 1.05 for the 14B variant. However, Figure 3 shows that the corresponding Spearman's ρ values for deepseek models fluctuate near zero (ranging from –0.018 to 0.031) across these sizes. This divergence is also evident in Figure 1, where deepseek models cluster on a line around a Spearman's ρ of zero while MAE decreases. This suggests that while increased parameters can improve absolute error metrics, they do not necessarily enhance the model's ability to rank predictions in alignment with the evaluation target for this family.

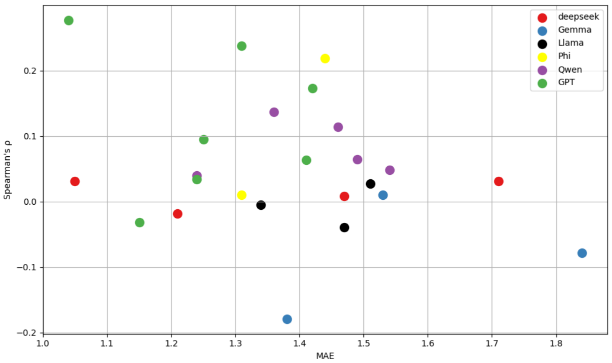


Figure 1. Mean Absolute Error vs. Spearman's ρ

In the Gemma 3 family (blue markers in Figure 1), Figure 2 indicates a U-shaped trend for MAE with increasing model size; the 1B model has an MAE of 1.84, the 4B model records a lower MAE of 1.38, and the 12B model shows an MAE of 1.53. Concurrently, Figure 3 highlights that the 4B model yields a statistically significant negative Spearman's ρ of –0.179 (p = 0.005), while the 1B and 14B models yield ρ values closer to zero (–0.078 and 0.010, respectively). These findings, also visible in Figure 1, where the 4B Gemma model stands out with its negative correlation, indicate that only specific scales within this family show notable differences in ranking performance, raising questions about non-linear scaling effects.

Model	MAE	ρ	p-value
deepseek 1.5B	1.47	0.008	0.894
deepseek 7B	1.71	0.031	0.627
deepseek 8B	1.21	-0.018	0.773
deepseek 14B	1.05	0.031	0.628
Gemma 3 1B	1.84	-0.078	0.222
Gemma 3 4B	1.38	-0.179	0.005
Gemma 3 12B	1.53	0.010	0.878
Llama 3 1B	1.34	-0.005	0.936
Llama 3 3B	1.51	0.027	0.666
Llama 3 8B	1.47	-0.039	0.535
Phi 4 3.8B	1.44	0.219	0.000
Phi 4 14B	1.31	0.010	0.874
Qwen 3 0.6B	1.36	0.137	0.031
Qwen 3 1.7B	1.24	0.040	0.528
Qwen 3 4B	1.54	0.048	0.453
Qwen 3 8B	1.49	0.065	0.303
Qwen 3 14B	1.46	0.114	0.072
GPT 4o	1.41	0.064	0.314
GPT 4o mini	1.15	-0.032	0.611
GPT 4.1	1.25	0.095	0.136
GPT 4.1 mini	1.42	0.173	0.006
GPT 4.1 nano	1.24	0.034	0.594
GPT o3 mini	1.31	0.238	0.000
GPT o4 mini	1.04	0.277	0.000

Table 1. Meta-Evaluation result of the tested models. MAE stands for Mean Absolute Error. ρ stands for Spearman's rank correlation coefficient.

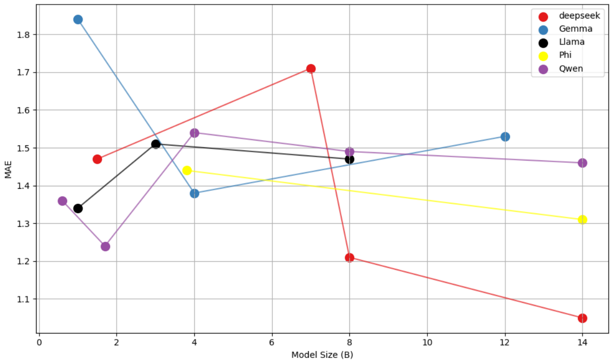


Figure 2. Mean Absolute Error vs. Model Size(Bilion of Parameters)

For Llama 3 models (black markers in Figure 1), the trend is less clear regarding MAE improvement with scaling. Figure 2 shows that the 2B, 3B, and 8B variants produce similar and relatively stable MAE values (ranging from 1.34 to 1.51). Correspondingly, Figure 3 demonstrates that Spearman's ρ values for Llama models are consistently near zero across these sizes (varying between –0.005 and 0.027). Figure 1 further confirms this, with Llama models tightly clustered around zero correlation. This general lack of significant change in either MAE or correlation across different sizes suggests that scaling within the tested Llama family range may not substantially impact either absolute accuracy or rank consistency for this task.

In the Phi family (yellow markers in Figure 1), Figure 2 shows that the 3.8B model achieves a MAE of 1.44, which slightly improves to 1.31 for the 14B model. However, Figure 3 reveals a striking contrast in correlation: the 3.8B model has a moderate positive Spearman's ρ (ρ = 0.219, p < 0.001), while the 14B model's ρ drops to a negligible value (0.010, p = 0.874). This pattern, clearly distinguishable in Figure 1, implies that reducing absolute error does not guarantee enhanced ordinal ranking of predictions and can even correspond to a decrease in ranking performance for this family.

The QWEN models (purple markers in Figure 1) exhibit more complex scaling dynamics. As seen in Figure 2, MAE for QWEN models does not follow a simple trend: the 0.5B model has an MAE of 1.36, which dips for the 1.8B model (MAE 1.24), then rises for the 4B (MAE 1.54) and 7B (MAE 1.71) models, before slightly decreasing for the 14B model (MAE 1.46). Spearman's ρ, shown in Figure 3, also fluctuates: the 0.5B model has a ρ of 0.137 (p = 0.031), which then varies for larger models (1.8B to 14B) between approximately 0.040 and 0.120. This variability, also reflected in the scatter of QWEN points in Figure 1, indicates that scaling within the QWEN family has a somewhat unpredictable impact on both absolute error and ranking performance.

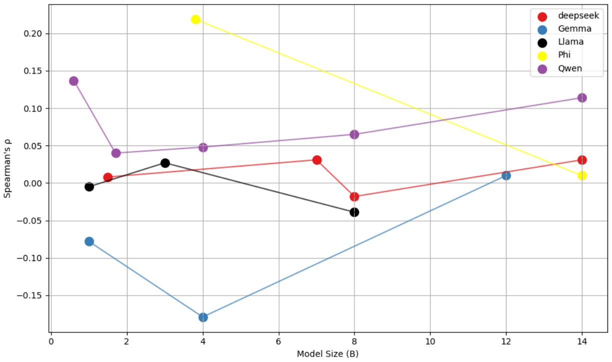


Figure 3. Spearman's ρ vs. Model Size(Bilion of Parameters)

Notably, the GPT family (green markers in Figure 1) reveals that "mini" variants can achieve both low MAE and comparatively stronger rank correlations. For instance, GPT-4o-mini shows an MAE of 1.04 and a Spearman's ρ of 0.277 (p = 0.000), the highest correlation observed among the GPT models plotted. Similarly, GPT-4-Turbo-mini (equivalent to text's GPT,4.1 mini) records an MAE of 1.42 with ρ = 0.173 (p = 0.006). As seen in Figure 1, these contrast with other larger GPT versions, such as GPT-4o (MAE 1.15, ρ = 0.018) and GPT-4 (MAE 1.42, ρ = 0.192), where the correlation, while sometimes positive, can be less pronounced than the top-performing mini variant. This suggests that a reduced architecture in this family might, in some cases, better capture the ordering of predictions.

In summary, the results indicate that while scaling can sometimes reduce absolute prediction error (MAE), it does not systematically improve the preservation of ordinal relationships as measured by Spearman's ρ. The diverse trends across model families support the view that model improvements should be evaluated on a case-by-case basis, considering both error minimization and rank correlation. Future work should investigate the architectural and training factors that contribute to these complex dynamics, with particular attention to why some models or families (such as certain GPT mini variants or the smaller Phi model) achieve better ranking performance relative to their size or absolute error.

Positive Bias The bar plots shown in Figures 4 and 5 comparing human and model ratings across the five editorial criteria reveal a consistent pattern of positive bias in model-generated evaluations. Most language models tend to assign higher scores than human annotators, particularly in subjective dimensions such as fluency and coherence. This trend is observed across multiple model families and parameter scales, suggesting a global rather than local phenomenon. This bias cannot be attributed to imbalanced prompting. The few-shot examples used to guide model behaviour were carefully constructed to span the full range of the scoring scale (1–5), ensuring that models were exposed to both high and low quality examples in equal quantities. This design choice rules out the possibility that models are simply mimicking overly generous examples. A more plausible explanation lies in the interplay of training data biases and alignment methodologies. Many evaluated models are pre-trained on large-scale synthetic corpora, often generated by other language models or curated to reflect "high-quality" outputs, which may encode implicit biases toward agreeableness or flattery [28, 15]. This aligns with findings that sycophantic tendencies can emerge from overfitting to human preferences during reinforcement learning from human feedback, where annotators disproportionately favor responses that align with their views [31]. For instance, studies show that even pretrained models exhibit sycophancy, likely due to absorbing patterns from internet texts where users often reinforce shared opinions (e.g., Reddit discussions) [28].

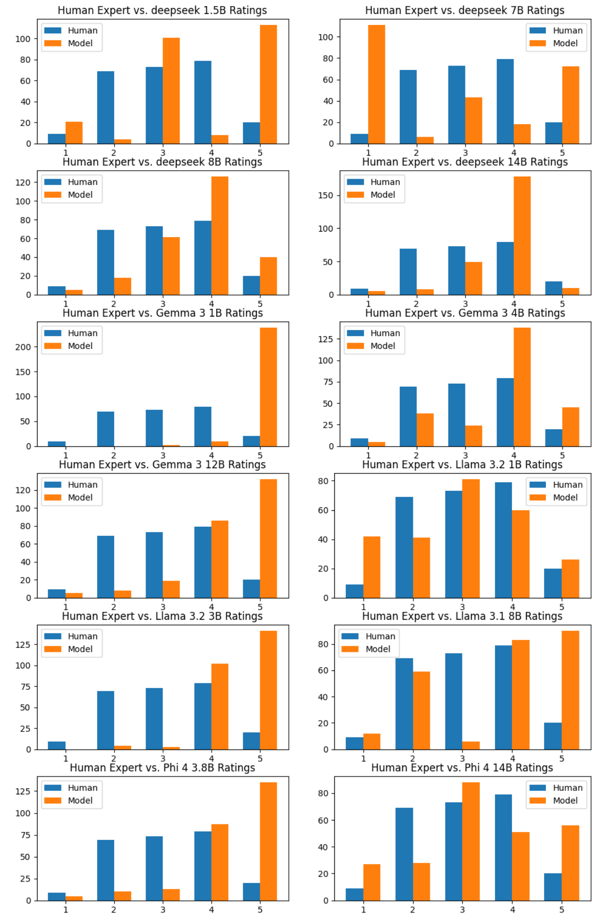


Figure 4. Human vs. LLM I. Rating on the x axis and Count on the y axis.

Additionally, the observed bias may stem from the models' tendency to "flip-flop" when challenged, altering answers to align with user suggestions even when initially correct. This behavior, quantified in experiments like FlipFlop [15], reveals that state-of-the-art LLMs (e.g., GPT-4, Gemini-Pro) frequently compromise accuracy to maintain user agreement, with sycophantic responses occurring in over 58% of cases [10]. Such dynamics are exacerbated by alignment objectives prioritizing politeness and helpfulness over factual rigor, inadvertently discouraging critical pushback [23]. Notably, while finetuning on synthetic datasets, balancing confirmatory and corrective responses can reduce sycophancy by 50% in some models (e.g., Mistral-7b), the persistence of regressive sycophancy (where agreement leads to incorrect answers) underscores the need for robust mitigation strategies that reconcile alignment with truthfulness [15].

The individual bar plots comparing human and model ratings across the five editorial criteria reveal a consistent pattern of positive bias in model evaluations. Across nearly all models and criteria, the distributions of model-assigned scores are skewed toward higher values relative to human annotations.

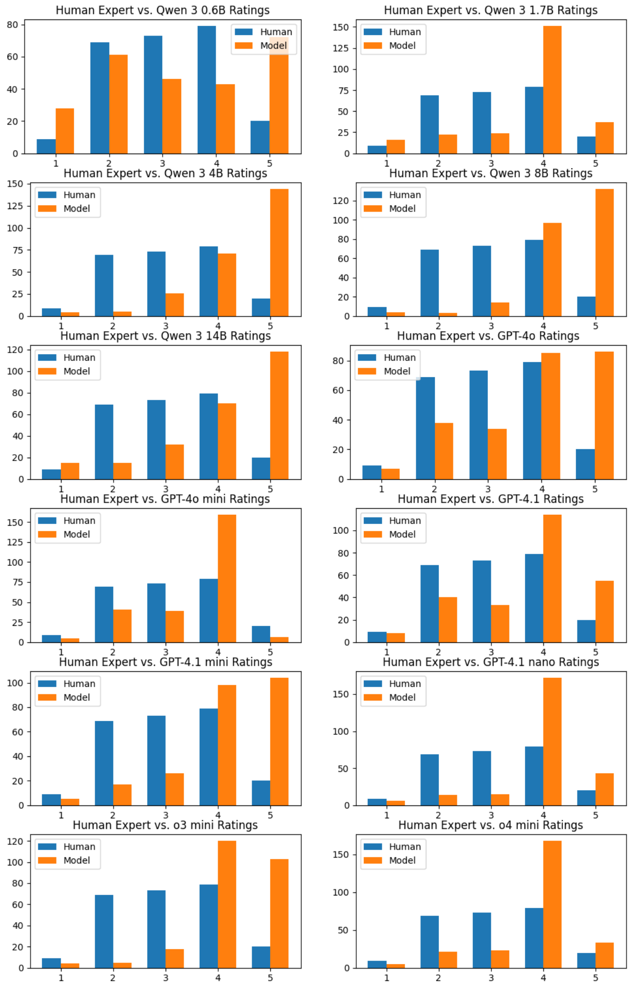


Figure 5. Human vs. LLM II. Rating on the x axis and Count on the y axis.

This trend is particularly pronounced in subjective dimensions such as fluency and coherence, where models frequently assign ratings of 4 or 5, even in cases where human annotators opt for more moderate scores. For instance, in the fluency barplots shown in Figure 6, models like GPT-4.1 mini, Gemma 3 1B, and Qwen 3 8B consistently overrepresent high scores, suggesting a systematic overestimation of linguistic quality. Similarly, in coherence barplots 6, models often rate summaries more favorably than human experts, with fewer low scores and a concentration around the upper end of the scale. This positive skew is not limited to a specific model family or size. It appears across both small and large models, including reasoning variants. This suggests that the bias is not merely a function of model capacity but may reflect shared training dynamics or evaluation heuristics.

In conclusion, while LLMs show promise as scalable evaluators, their tendency to overrate outputs highlights the need for calibration. Future work should explore methods to mitigate this bias, such as incorporating human-aligned calibration datasets, adversarial prompting, or ensemble evaluation strategies that combine model and human judgments.

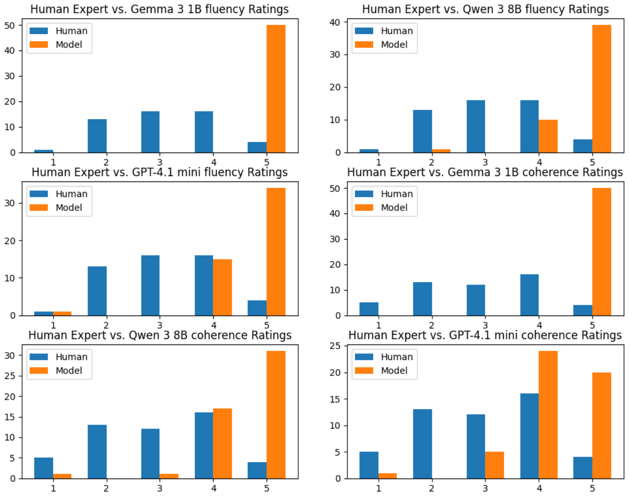


Figure 6. Human vs. LLM ratings for Fluency and Coherence criteria. Rating on the x axis and Count on the y axis.

Model-to-Model Agreement We evaluated model-to-model alignment in judging preferences using Spearman's ρ, with all reported correlations being statistically significant ($p < 0.05$). The findings reveal a complex landscape of agreement, strongly influenced by model size and family. Generally, smaller models demonstrated limited consensus in their preference rankings. For instance, DeepSeek 1.5B consistently showed negligible or negative alignment across a range of models, including Qwen 3 14B (ρ = 0.127, p = 0.044), Phi 4 3.8B (ρ = 0.168, p = 0.008), and even GPT-4o mini (ρ = 0.130, p = 0.039). A similar pattern was observed for DeepSeek 7B, which also exhibited negligible or even negative correlations, such as with Llama 3.2 3B (ρ = -0.132, p = 0.037) and only slightly better with larger models like GPT-4o (ρ = 0.154, p = 0.015). The Qwen 3 0.6B model also struggled to find common ground, showing poor alignment not only with models from other families like Phi 4 3.8B (ρ = 0.203, p = 0.001) but also with its larger siblings such as Qwen 3 8B (ρ = 0.139, p = 0.028).

The agreement tended to improve as model size increased. DeepSeek 8B, for example, began to show more instances of "Low" alignment, particularly with various Qwen 3 models (e.g., Qwen 3 4B: ρ = 0.490, p = 0.000) and some GPT variants (e.g., GPT-4.1: ρ = 0.422, p = 0.000), though it still had negligible alignment with others like Llama 3.1 8B (ρ = 0.218, p = 0.001). This trend was more pronounced with DeepSeek 14B, which achieved more consistent "Low" to "Medium" alignments, such as with Qwen 3 4B (ρ = 0.555, p = 0.000), Gemma 3 12B (ρ = 0.554, p = 0.000), and GPT-4.1 (ρ = 0.561, p = 0.000).

Intra-family alignment also generally strengthened with model scale. Within the Qwen 3 series, while the 0.6B model showed weak correlations, the alignment between Qwen 3 4B and Qwen 3 8B (ρ = 0.655, p = 0.000) and Qwen 3 4B and Qwen 3 14B (ρ = 0.671, p = 0.000) reached "Medium" levels. Similarly, Gemma 3 4B and Gemma 3 12B had a "Medium" alignment (ρ = 0.583, p = 0.000). The most striking intra-family consensus was observed among the GPT models, with GPT-4.1 showing "High" alignment with GPT-4o (ρ = 0.810, p = 0.000) and GPT-4.1 mini (ρ = 0.781, p = 0.000).

Stronger cross-family correlations also emerged predominantly between larger, more capable models. For example, Qwen 3 14B achieved "High" alignment with GPT-4.1 (ρ = 0.725, p = 0.000) and GPT-4.1 mini (ρ = 0.708, p = 0.000). Gemma 3 12B also showed "Medium" to "High" alignment with GPT variants, such as GPT-4o mini (ρ = 0.728, p = 0.000). This overarching pattern suggests that while smaller or perhaps more uniquely architectured models may show peculiar ranking behaviours, larger models, particularly those from similar development paradigms or within the same family, tend to converge more substantially in their evaluative judgments, indicating a developing consensus on preference at the higher end of model capability.

6 Conclusions

Three themes are emerging from our experiments on how off-the-shelf LLMs behave when asked to judge outputs against a fixed rubric. First, size matters, but only up to a point. As models grow larger, they generally make fewer absolute errors in scoring, which might lead you to think "bigger is always better". Yet when we look at how well these scores line up in rank order with human judgments, the picture is more mixed. Some of the smaller "mini" variants do a better job of getting the ordering right than their much larger siblings. In other words, raw scale helps with scoring precision but doesn't automatically translate into human-like ranking ability. Second, almost every model we tested leans on the generous side. They tend to hand out higher scores than human experts do, especially on subjective dimensions like fluency or coherence. This consistent positive bias suggests that the models' pretraining and alignment processes prime them to sound "helpful", perhaps at the expense of rigour. In practice, it means you can't assume their high marks carry the same weight as an expert's. Ultimately, you'll see that agreement between models follows a similar pattern: small or architecturally distinct models often disagree wildly, whereas larger models within the same family converge on very similar judgments. So if you're looking for consistency between multiple LLM-based judges, you'll get it only once you reach a certain size threshold. Putting all of this together, we conclude that LLMs are capable of approximating human scores, but they still struggle with unbiased ranking and inter-model consensus at smaller scales. This could stem from a lack of un understanding of the evaluation rubric. Moving forward, targeted calibration techniques and a closer look at what makes some "mini" models better rankers might hold the key to get more reliable automated judges that better understand the scoring criteria.

7 Limitations and Future Work

Our study of LLMs as judges is necessarily bounded by several methodological choices. First, we relied on a deliberately constructed test set of 10 Italian news articles and 50 GPT-4o-generated summaries, each curated to isolate one of five broad editorial criteria (coherence, consistency, fluency, relevance, ordering) and scored by a single expert annotator. While this design ensures that models must genuinely interpret each rubric item, it limits generalisability to other languages, genres, or more fine-grained aspects of writing (e.g. style or audience adaptation) . Moreover, we elicited judgments exclusively via few-shot prompting, with no model fine-tuning, which may understate the ceiling performance achievable through instruction-tuning or task-specific training. Our evaluation metrics, Spearman's ρ and Mean Absolute Error, capture ranking and absolute-score alignment but do not assess the quality or usefulness of the models' rationales. Finally, our analysis revealed a consistent positive bias, models tend to over-rate subjective dimensions such as fluency and coherence, likely inherited from their training data and alignment objectives.

Looking ahead, we envision several avenues to deepen and broaden this work. Extending the framework to diverse domains (e.g., scientific abstracts, social media) and additional languages would test rubric robustness beyond Italian news. Fine-tuning or instruction-tuning LLMs on human-annotated evaluation data (or distilling high-quality judgments from expert-calibrated models) could improve both absolute accuracy and ranking alignment. Enriching and adapting the rubric with more nuanced or task-specific criteria (factual depth, style conformity, audience orientation) and adopting dynamic weighting schemes would better reflect real-world priorities. To mitigate positive bias, calibration techniques such as temperature scaling or human-in-the-loop correction are needed. Establishing benchmarks with multiple expert annotators would quantify inter-annotator variability and yield more reliable ground truth. Finally, exploring ensemble or multi-agent evaluator architectures and investigating why smaller “mini” variants sometimes excel in ranking promises insights into efficient, reliable automated judgment systems.

Acknowledgements

We acknowledge *Zanichelli editore* for their support in enabling this research. Their provision of access to digital infrastructure and expertise significantly facilitated the curation of the dataset and the human evaluation processes. Special thanks to Dr. Isabella Nenci for her dedicated contribution to dataset annotation and for sharing her expertise. We extend our sincere appreciation to the anonymous reviewers for their insightful feedback and constructive suggestions. This work was partially supported by project FAIR: Future Artificial Intelligence Research (European Commission NextGeneration EU programme, PNRR-M4C2-Investimento 1.3, PE00000013-"FAIR" - Spoke 8).

References

[1] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861, 2021. URL https://arxiv.org/abs/2112.00861.

[2] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. E. Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862,

2022. doi: 10.48550/ARXIV.2204.05862. URL https://doi.org/10.48550/arXiv.2204.05862.

[3] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosiute, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022. doi: 10.48550/ARXIV.2212.08073. URL https://doi.org/10.48550/arXiv.2212.08073.

[4] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

[5] A. Chaganty, S. Mussmann, and P. Liang. The price of debiasing automatic metrics in natural language evaluation. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1060. URL https://aclanthology.org/P18-1060/.

[6] C. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *CoRR*, abs/2308.07201, 2023. doi: 10.48550/ARXIV.2308.07201. URL https://doi.org/10.48550/arXiv.2308.07201.

[7] Y. Dubois, C. X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/5fc47800ee5b30b8777fdd30abcaaf3b-Abstract-Conference.html.

[8] K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR, 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

[9] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 04 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00373. URL https://doi.org/10.1162/tacl_a_00373.

[10] A. Fanous, J. Goldberg, A. A. Agarwal, J. Lin, A. Zhou, R. Daneshjou, and S. Koyejo. Syceval: Evaluating LLM sycophancy. *CoRR*, abs/2502.08177, 2025. doi: 10.48550/ARXIV.2502.08177. URL https://doi.org/10.48550/arXiv.2502.08177.

[11] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87/.

[12] H. Huang, X. Bu, H. Zhou, Y. Qu, J. Liu, M. Yang, B. Xu, and T. Zhao. An empirical study of llm-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5880–5895. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.306/.

[13] S. Kim, J. Shin, Y. Choi, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, and M. Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

[14] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Y. Al-Onaizan, M. Bansal, and Y. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4334–4353. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.248. URL https://doi.org/10.18653/v1/2024.emnlp-main.248.

[15] P. Laban, L. Murakhovs'ka, C. Xiong, and C. Wu. Are you sure? challenging llms leads to performance drops in the flipflop experiment. *CoRR*, abs/2311.08596, 2023. doi: 10.48550/ARXIV.2311.08596. URL https://doi.org/10.48550/arXiv.2311.08596.

[16] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. R. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. Rewardbench: Evaluating reward models for language modeling. *CoRR*, abs/2403.13787, 2024. doi: 10.48550/ARXIV.2403.13787. URL https://doi.org/10.48550/arXiv.2403.13787.

[17] J. Li, S. Sun, W. Yuan, R. Fan, H. Zhao, and P. Liu. Generative judge for evaluating alignment. *CoRR*, abs/2310.05470, 2023. doi: 10.48550/ARXIV.2310.05470. URL https://doi.org/10.48550/arXiv.2310.05470.

[18] T. Li, W. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, and I. Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *CoRR*, abs/2406.11939, 2024. doi: 10.48550/ARXIV.2406.11939. URL https://doi.org/10.48550/arXiv.2406.11939.

[19] X. Li, T. Zhang, Y. Dubois, R. Taori, I. Gulrajani, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

[20] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

[21] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

[22] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.153. URL https://doi.org/10.18653/v1/2023.emnlp-main.153.

[23] L. Malmqvist. Sycophancy in large language models: Causes and mitigations. *CoRR*, abs/2411.15287, 2024. doi: 10.48550/ARXIV.2411.15287. URL https://doi.org/10.48550/arXiv.2411.15287.

[24] N. Mathur, T. Baldwin, and T. Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL https://aclanthology.org/2020.acl-main.448/.

[25] N. Muennighoff, Q. Liu, A. Zebaze, Q. Zheng, B. Hui, T. Y. Zhuo, S. Singh, X. Tang, L. von Werra, and S. Longpre. Octopack: Instruction tuning code large language models. *CoRR*, abs/2308.07124, 2023. doi: 10.48550/ARXIV.2308.07124. URL https://doi.org/10.48550/arXiv.2308.07124.

[26] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

[28] E. Perez, S. Ringer, K. Lukosiute, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, A. Jones, A. Chen, B. Mann, B. Israel, B. Seethor, C. McKinnon, C. Olah, D. Yan, D. Amodei, D. Amodei, D. Drain, D. Li, E. Tran-Johnson, G. Khundadze, J. Kernion, J. Landis, J. Kerr, J. Mueller, J. Hyun, J. Landau, K. Ndousse, L. Goldberg, L. Lovitt, M. Lucas, M. Sellitto, M. Zhang, N. Kingsland, N. Elhage, N. Joseph, N. Mercado, N. DasSarma, O. Rausch, R. Larson, S. McCandlish, S. Johnston, S. Kravec, S. E. Showk, T. Lanham, T. Telleen-Lawton, T. Brown, T. Henighan,

T. Hume, Y. Bai, Z. Hatfield-Dodds, J. Clark, S. R. Bowman, A. Askell, R. B. Grosse, D. Hernandez, D. Ganguli, E. Hubinger, N. Schiefer, and J. Kaplan. Discovering language model behaviors with model-written evaluations. In A. Rogers, J. L. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13387–13434. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.847. URL https://doi.org/10.18653/v1/2023.findings-acl.847.

[29] E. Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, Sept. 2018. doi: 10.1162/coli_a_00322. URL https://aclanthology.org/J18-3002/.

[30] P. Röttger, H. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In K. Duh, H. Gómez-Adorno, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5377–5400. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.301. URL https://doi.org/10.18653/v1/2024.naacl-long.301.

[31] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

[32] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize from human feedback. *CoRR*, abs/2009.01325, 2020. URL https://arxiv.org/abs/2009.01325.

[33] P. Verga, S. Hofstätter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, M. Xu, N. White, and P. Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *CoRR*, abs/2404.18796, 2024. doi: 10.48550/ARXIV.2404.18796. URL https://doi.org/10.48550/arXiv.2404.18796.

[34] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators. In L. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.511. URL https://doi.org/10.18653/v1/2024.acl-long.511.

[35] Y. Wang, Z. Yu, Z. Zeng, L. Yang, C. Wang, H. Chen, C. Jiang, R. Xie, J. Wang, X. Xie, W. Ye, S. Zhang, and Y. Zhang. Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization. *CoRR*, abs/2306.05087, 2023. doi: 10.48550/ARXIV.2306.05087. URL https://doi.org/10.48550/arXiv.2306.05087.

[36] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin. Do-not-answer: Evaluating safeguards in llms. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 896–911. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-eacl.61.

[37] Z. Zeng, J. Yu, T. Gao, Y. Meng, T. Goyal, and D. Chen. Evaluating large language models at evaluating instruction following. *CoRR*, abs/2310.07641, 2023. doi: 10.48550/ARXIV.2310.07641. URL https://doi.org/10.48550/arXiv.2310.07641.

[38] X. Zhang, B. Yu, H. Yu, Y. Lv, T. Liu, F. Huang, H. Xu, and Y. Li. Wider and deeper LLM networks are fairer LLM evaluators. *CoRR*, abs/2308.01862, 2023. doi: 10.48550/ARXIV.2308.01862. URL https://doi.org/10.48550/arXiv.2308.01862.

[39] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

[40] L. Zhu, X. Wang, and X. Wang. Judgelm: Fine-tuned large language models are scalable judges. *CoRR*, abs/2310.17631, 2023. doi: 10.48550/ARXIV.2310.17631. URL https://doi.org/10.48550/arXiv.2310.17631.

A Prompts

Prompt Template for Coherence Prompt used for the Coherence criterion. Few-shot examples are provided in the GitHub repo [1].

```
As an impartial evaluator, your task is to assess
    the coherence of a given summary in relation
    to its source material by assigning a score
    from 1 to 5 and providing a detailed
    explanation in Italian that justifies your
    rating.
Focus on how well the summary is organized and
    whether it presents the source's information
    in a logical and structured way.
Coherence refers to how well the sentences in the
    summary flow together to form a unified whole.
    A coherent summary should present the main
    ideas in a clear, logical progression,
    avoiding any abrupt shifts or disjointed
    facts. The goal is for the reader to easily
    follow the line of reasoning or narrative
    without confusion.

Evaluation Criteria
To conduct a thorough assessment, consider the
    following sub-criteria:
    Logical Structure and Organization: Assess
        whether the summary follows a clear
        progression of ideas (introduction, body,
        conclusion) that mirrors the source
        material.
    Transitions: Evaluate if there are smooth
        transitions between sentences and
        paragraphs that facilitate the reader's
        understanding.
    Clarity and Conciseness: Determine if the
        language is precise and unambiguous,
        effectively conveying the core ideas
        without unnecessary complexity.

Evaluation Process
Review the Source Material: Thoroughly read the
    source document to understand its main facts,
    events, and details.
Analyze the summary: Compare the summary against
    the source material, evaluating it based on
    the sub-criteria outlined above.
Assign a Coherence Score and provide an
    Explanation:
Based on your analysis, assign a coherence score
    from 1 to 5, where the levels are defined as
    follows.
    Score 1 (Very Poor Coherence):
        The summary is highly disorganized with
            abrupt transitions. The summary
            exhibits little to no logical flow. It
            is difficult to understand the
            relationship between concepts.
    Score 2 (Poor Coherence):
        The summary shows some attempt at
            organization but remains fragmented
            with several abrupt shifts. Key points
            are only partially integrated in a
            fluent narrative. The sentences are
            fragmented with abrupt transitions.
            The lack of clear connections between
            ideas results in a choppy reading
            experience.
    Score 3 (Moderate Coherence):
```

[1] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

````
        The summary is reasonably organized with a
            generally logical progression.
            Transitions exist but may be uneven,
            they could be smoother.
    Score 4 (Good Coherence):
        The summary is well-structured with a
            clear and logical order of ideas. It
            features smooth transitions between
            sentences and paragraphs, making it
            easy to follow. The summary is
            coherent and flows well, with clear
            connections between ideas.
    Score 5 (Excellent Coherence):
        The summary exhibits exceptional
            coherence. The transitions are
            flawless and the presentation of the
            source material is clear and unified.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{’score’: ’<score between 1 and 5 from
    very poor to excellent>’, ’explanation’:
    ’<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>’}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_coherence}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
````

Prompt Template for Consistency Prompt used for the Consistency criterion. Few-shot examples are provided in the GitHub repo [2].

````
As an impartial evaluator, your task is to assess
    the consistency of a given summary in relation
    to its source material by assigning a score
    from 1 to 5 and providing a detailed
    explanation in Italian that justifies your
    rating.
Consistency refers to the degree to which the
    summary accurately and faithfully represents
    the factual content of the source without
    introducing contradictions, inaccuracies, or
    unsupported information.
A consistent summary should align closely with the
    source material, ensuring that all presented
    information is both accurate and verifiable.

Evaluation Criteria
To conduct a thorough assessment, consider the
    following sub-criteria:
    Factual Accuracy: Verify that the summary
        accurately represents explicit facts from
        the source, including names, dates,
        numbers, and locations. Cross-reference
        specific claims in the summary with the
        source to confirm their precision.
    Absence of Contradictions: Ensure that the
        summary does not contain information that
        directly contradicts the source material.
        Identify any opposing statements or
        conflicting details between the summary
        and the source.
    Absence of Hallucinations (Extrinsic
        Consistency): Check that the summary does
        not introduce information absent from the
        source. All details should be traceable to
        the original text, and any unsubstantiated
        additions should be noted.
    Logical Inferences (Intrinsic Consistency):
        Assess whether any inferences or
        conclusions drawn in the summary are
        logically supported by the information
        provided in the source. Ensure that
        deductions are valid and reasonable based
        on the source material.
    Terminology Alignment: Confirm that the
        summary uses the same key terms and refers
        to entities consistently with the source
        material. While paraphrasing is
        acceptable, maintaining consistency in
        terminology is important for clarity and
        accuracy.

Evaluation Process
Review the Source Material: Thoroughly read the
    source document to understand its main facts,
    events, and details.
Analyze the summary: Compare the summary against
    the source material, evaluating it based on
    the sub-criteria outlined above.
Assign a Consistency Score and provide an
    Explanation:
    Based on your analysis, assign a consistency
        score from 1 to 5, where the levels are
        defined as follows.
    Score 1 (Very Poor Consistency):
        The summary contains significant factual
            inaccuracies, contradictions,
            hallucinated details, or
            misrepresentations that severely
            distort the source material.
        The summary introduces entirely fabricated
            events or represents critical
            information such that it no longer
            reflects the source.
    Score 2 (Poor Consistency):
        The summary has multiple errors and
            inconsistencies; while some key facts
            may be correct, there are notable
            inaccuracies or added details that
            conflict with the source material.
        The summary includes several incorrect
            dates, names, or details that
            contradict the source, resulting in a
            misleading representation.
    Score 3 (Moderate Consistency):
        The summary is generally accurate but
            contains minor errors, omissions, or
            slight paraphrasing issues that affect
            the overall precision.
        Most details match the source, but a few
            minor discrepancies or vague terms
            slightly reduce the clarity of the
            summary.
    Score 4 (Good Consistency):
        The summary is largely consistent with the
            source, with only trivial
            discrepancies that do not impact the
            overall factual integrity.
        The summary accurately reflects the main
            facts and events, with only minor
            stylistic differences that do not
            alter the meaning.
    Score 5 (Excellent Consistency):
````

[2] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

~~~~
        The summary is fully consistent with the
            source, accurately representing every
            key fact and detail without any added
            or contradictory information.
        The summary perfectly mirrors the source
            material, ensuring that every piece of
            information is correctly and
            completely conveyed.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{’score’: ’<score between 1 and 5 from
    very poor to excellent>’, ’explanation’:
    ’<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>’}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_consistency}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
~~~~

Prompt Template for Fluency Prompt used for the Fluency criterion. Few-shot examples are provided in the GitHub repo [3].

~~~~
As an impartial evaluator, your task is to assess
    the fluency of a given summary by assigning a
    score from 1 to 5 and providing a detailed
    explanation in Italian that justifies your
    rating.
Fluency refers to the readability and overall
    quality of the summary’s writing. This
    includes assessing grammar, spelling,
    punctuation, word choice, and sentence
    structure. A fluent summary should be free
    from errors that make the text difficult to
    read or understand.

Evaluation Criteria
To conduct a thorough assessment, consider the
    following sub-criteria.
    Grammar: Check for accuracy, tense
        consistency, and overall syntax.
    Spelling: Identify any spelling mistakes or
        typographical errors.
    Punctuation: Assess proper punctuation usage
        and its contribution to clarity.
    Word Choice: Evaluate whether vocabulary and
        phrasing are appropriate for the context.
    Sentence Structure: Determine if sentences are
        well-constructed, varied, and natural.

Evaluation Process
Read the summary carefully.
Check for errors:
    Are there grammatical errors? Are there
        frequent or severe errors present?
    Is there any spelling or punctuation mistakes?
    Does the word choice suit the context without
        being overly complex or too simplistic?
Assign a Fluency Score and provide an Explanation:
    Based on your analysis, assign a coherence
        score from 1 to 5, where the levels are
        defined as follows.
    Score 1 (Very Poor Fluency):
        Numerous errors in grammar, spelling,
            punctuation, and word/sentence
            construction make the summary
            extremely difficult to read.
    Score 2 (Poor Fluency):
        Frequent errors are present that interfere
            with understanding. Sentence structure
            and vocabulary choices are suboptimal,
            leading to a choppy flow.
    Score 3 (Moderate Fluency):
        Errors exist, but they do not hinder
            summary understandability. Occasional
            awkward phrasing or punctuation
            mistakes are present.
    Score 4 (Good Fluency):
        The summary is well-written with only
            isolated, minor errors. Grammar,
            spelling, punctuation, and sentence
            structure are correct, ensuring smooth
            readability.
    Score 5 (Excellent Fluency):
        The summary is polished and flawless, with
            impeccable grammar, spelling,
            punctuation, word choice, and sentence
            structure that provide a natural flow.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{’score’: ’<score between 1 and 5 from
    very poor to excellent>’, ’explanation’:
    ’<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>’}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_fluency}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
~~~~

Prompt Template for Relevance Prompt used for the Relevance criterion. Few-shot examples are provided in the GitHub repo [4].

~~~~
As an impartial evaluator, your task is to assess
    the relevance of a given summary in relation
    to its source material by assigning a score
    from 1 to 5 and providing a detailed
    explanation in Italian that justifies your
    rating.
Relevance refers to how well the summary includes
    only the most important and necessary content
    from the source material, without introducing
    redundant or irrelevant details.
A relevant summary should focus on the key points
    of the source and avoid unnecessary or
    excessive information.

Evaluation Criteria
~~~~

[3] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

[4] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

````
To conduct a thorough assessment, consider the
    following sub-criteria.
    Content Coverage and Accuracy: Does the
        summary capture all of the primary
        arguments, data points, or ideas presented
        in the source document? Is the information
        presented in the summary faithful to the
        original intent and details of the source?
    Conciseness and Clarity:Is the summary
        expressed in a concise manner that does
        not sacrifice the essential details? Are
        the ideas presented clearly and
        straightforwardly, ensuring that the
        summary Does not confuse the reader with
        verbose or circular language?
    Elimination of Redundancy and Irrelevance:
        Removal of Superfluous Information: Does
        the summary avoid including unnecessary
        background or repetitive details that do
        not contribute to understanding the
        source? Are only the important and
        relevant aspects of the source material
        captured, with a clear focus on the
        essential message?
    Omission of Critical Elements: Does the
        summary omit any critical elements or
        supporting details that are necessary for
        a complete and accurate understanding of
        the source document?


Evaluation Process
Review the Source Material: Thoroughly read the
    source document to understand its main facts,
    events, and details.
Analyze the summary: Compare the summary against
    the source material, evaluating it based on
    the sub-criteria outlined above.
Assign a Consistency Score and provide an
    Explanation:
    Score 1 (Very Poor Relevance):
        The summary includes little to none of the
            key points from the source.
        The summary is Overburdened with
            irrelevant, redundant, or incorrect
            details.
        Critical points are missing from the
            summary, leading to a distorted or
            incomplete picture.
    Score 2 (Poor Relevance):
        The summary captures some primary points,
            but many important aspects are either
            omitted or misrepresented.
        The summary includes redundant or
            extraneous information that dilutes
            the primary message.
        Key supporting details are missing,
            reducing the summary’s overall
            reliability.
    Score 3 (Fair Relevance):
        The summary captures more than half of the
            key points, but some secondary details
            or nuanced information may be lacking.
        The summary is mostly concise with minor
            instances of unnecessary details or
            slight redundancy.
        Less-critical details may be omitted from
            the summary without drastically
            affecting the overall understanding.
    Score 4 (Good Relevance):
        Successfully includes nearly all important
            points and supporting details from the
            source.
        The summary is clear and succinct, with
            minimal, if any, redundant content.
        Rare omissions that do not significantly
            impair the overall understanding of
            the summary.
    Score 5 (Excellent Relevance):
        The summary completely captures all
            essential points and nuances of the
            source material.
        The summary is extremely concise and
            clear, with no unnecessary or
            redundant information.
        No significant information is omitted; the
            summary is a precise and complete
            representation of the source.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{’score’: ’<score between 1 and 5 from
    very poor to excellent>’, ’explanation’:
    ’<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>’}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_relevance}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
````

Prompt Template for Ordering Prompt used for the Ordering criterion. Few-shot examples are provided in the GitHub repo [5].

````
As an impartial evaluator, your task is to assess
    the ordering of a given summary in relation to
    the ordering of the source material by
    assigning a score from 1 to 5 and providing a
    detailed explanation in Italian that justifies
    your rating.
Focus on how well the sequence of information in
    the summary mirrors the order in which it is
    presented in the source material.
Ordering refers to how closely the summary adheres
    to the structure of the source material. A
    well-ordered summary should present the key
    points in the same sequence as they appear in
    the source, ensuring a logical and coherent
    flow of information.

Evaluation Criteria
To conduct a thorough assessment, consider the
    following sub-criteria.
    Chronological/Logical Order: Confirm that if
        the source is structured chronologically
        or by logical argument, the summary
        upholds that framework. Deviations should
        be penalized based on their impact on the
        intended progression.
    Segmentation and Grouping: Consider if the
        summary correctly groups related
        information as seen in the source.
        Grouping similar ideas ensures that the
        coherence of the original narrative is
        maintained.
    Cohesion and Comprehension Impact: Assess if
        any deviation
        (omission/insertion/reordering)
````

[5] Code, Prompts, Data and Results: https://github.com/ZanichelliEditore/llm-summarization-evaluation

```
        significantly affects the reader’s ability
        to follow and understand the overall
        narrative.

Evaluation Process
Review the Source Material: Thoroughly read the
    source document focusing on the ordering of
    the main facts, events, and details.
Analyze the summary: Compare the ordering of the
    main facts, events, and details of the summary
    against the source material, evaluating it
    based on the sub-criteria outlined above.
Assign a Ordering Score and provide an Explanation:
    Score 1 (Very Poor Ordering):
        Key points are not only out of sequence
            but the entire summary structure is
            different than the source material.
        The summary introduces significant
            confusion, hindering comprehension of
            the source narrative.
        Major segments are reversed or
            intermingled.
    Score 2 (Poor Ordering):
        The majority of the summary’s structure
            deviates from the source order.
        Several key transitional phrases and
            segments are misplaced or omitted.
        Although some basic structure might be
            discernible, it still leads to a
            disjointed narrative.
        Noticeable reordering with multiple
            inconsistencies.
    Score 3 (Fair Ordering):
        The summary preserves parts of the source
            order while containing noticeable
            reordering in other sections.
        Some transitions and sequencing are
            maintained correctly, though there are
            occasional inconsistencies.
        The overall narrative is understandable,
            but the flow is less coherent than the
            source.
        A mixed pattern of accurate segments and
            segments with minor shifts.
    Score 4 (Good Ordering):
        The summary largely follows the sequence
            of the source material.
        Most key points and transitional phrases
            maintain their original order.
        Minor deviations may exist but do not
            materially disrupt the overall
            coherence or logical flow.
        Nearly complete alignment with the
            source’s narrative structure.
        Any reordering is minimal and does not
            limit comprehension.
    Score 5 (Excellent Ordering):
        The summary mirrors exactly the structure
            of the source material.
        All key segments, transitional cues, and
            the logical narrative flow are
            preserved.
        The reader can effortlessly follow the
            progression as intended in the
            original document.
        Consistent preservation of order, ensuring
            clarity and cohesion.
        The sequence of information is methodical
            and reflective of the source.
Provide your score along with a detailed
    explanation in italian that justifies your
    rating, referencing specific examples and
    observations from your evaluation.

Output in the following json template: {% raw
    %}```{’score’: ’<score between 1 and 5 from
    very poor to excellent>’, ’explanation’:
    ’<spigazione del voto dato al riassunto
    basandosi sullo specifico criterio di
    valutazione>’}```{% endraw %}
Update values enclosed in <> and remove the <>.
Your response must only be the updated json
    template beginning with { and ending with }
Ensure the following output keys are present in
    the json: score explanation
{{few_shot_examples_ordering}}
Now Evaluate:
<Input>
    <Source_Material>
        <Text>{{document}}</Text>
    </Source_Material>
    <Summary>
        <Text>{{summary}}</Text>
    </Summary>
</Input>
<Output>
```

B Annotation Guidelines

We developed a dedicated set of annotator guidelines to support the evaluation of Italian summaries according to editorial standards. They aim to ensure consistency and inter-annotator agreement in the qualitative evaluation of summaries. You can find the guidelines in the pages below in both English and Italian.

Annotation Guidelines

These guidelines define the process and criteria for evaluating a text summary based on five dimensions: coherence, consistency, relevance, fluency, and ordering. For each dimension, the following are provided:

1. Description of the task
2. Definition of the evaluation criterion and subcriteria
3. Rating scale from 1 to 5 with description of each level

Each summary in the Google sheet should be rated according to the following 5 criteria by entering its rating from 1 to 5 in the column of the same name as the criterion.

1. Coherence

Task description: Assess how well the summary presents the information in the text in a logical and structured way.

Definition: Coherence measures the fluency and unity of the text, that is, how logically the sentences flow, avoiding abrupt or discontinuous transitions.

Subcriteria:

- **Logical progression of ideas:** Ideas are presented in an order that follows a logical and natural thread.
- **Clarity and conciseness:** Sentences are formulated clearly and concisely.
- **Presence of transitions:** Connectives and transitions are used to tie sentences and paragraphs together.

Rating scale:

- **1 (Very Poor):** Disorganised text, absent transitions, difficult to follow the thread of discourse.
- **2 (Poor):** Fragmented structure, abrupt transitions, narrative not very fluid.
- **3 (Moderate):** Generally logical progression, transitions present but irregular.
- **4 (Good):** Clear structure, smooth transitions, easy to follow.
- **5 (Excellent):** Impeccable coherence, perfectly connected passages.

2. Consistency

Task description: Check the factual accuracy of the summary against the original text.

Definition: Consistency measures the fidelity of facts: absence of contradictions, errors and information not present in the source text.

Subcriteria:

- **Factual Accuracy:** All statements in the summary correspond to the facts expressed in the source text.
- **Absence of contradictions:** No part of the summary contradicts what is stated in the source text.
- **Absence of hallucinations:** No invented or added information is present that does not appear in the source text.
- **Logical Inference:** The inferred information is consistent with and supported by the content of the original text.
- **Terminological alignment:** Terms used in the summary are consistent with those in the source text, especially for technical or specialised concepts.

Rating scale:

- **1 (Very poor):** Numerous inaccuracies and invented details.
- **2 (Poor):** Multiple errors and discrepancies.
- **3 (Moderate):** Generally accurate, but with minor inaccuracies.
- **4 (Good):** Very few negligible discrepancies.
- **5 (Excellent):** Total fidelity to the facts of the text.

3. Relevance

Task description: Assess whether the summary includes only the essential contents of the source text.

Definition: Relevance measures the inclusion of key points and avoids superfluous or irrelevant details.

Subcriteria:

- **Inclusion of main ideas:** The core concepts of the source text are present in the summary.
- **Conciseness:** The content is expressed briefly but completely.
- **Absence of redundancy:** There are no unnecessary repetitions.
- **Absence of critical omissions:** No essential concepts have been omitted.

Rating scale:

- **1 (Very poor):** Almost all key points are missing; it contains much irrelevant information.
- **2 (Poor):** Covers some main points but leaves out important aspects; presence of superfluous details.
- **3 (Fair):** Covers more than half of the key points; some redundancy or minor omissions.
- **4 (Good):** Includes almost all essential points; minimal redundancy.
- **5 (Excellent):** Fully covers key points; extremely concise.

4. Fluidity

Task description: Assess the linguistic quality of the summary: grammar, spelling, punctuation and style.

Definition: Fluency measures readability and the absence of linguistic errors.

Subcriteria:

- **Grammar:** Absence of grammatical errors.
- **Spelling:** Words are spelt correctly.
- **Punctuation:** Appropriate use of punctuation marks.
- **Lexical choice:** Vocabulary is appropriate and varied.
- **Sentence structure:** Sentences are well constructed and of appropriate length.

Rating scale:

- **1 (Very poor):** Numerous serious errors that hinder reading.
- **2 (Poor):** Frequent errors hinder comprehension.
- **3 (Moderate):** Minor errors do not hinder reading.
- **4 (Good):** Isolated minor errors; fluent reading.
- **5 (Excellent):** Text impeccable in every respect.

5. Ordering

Task description: Assess whether the order of the information in the summary reflects that of the source text.

Definition: Sorting measures the alignment of the information sequence with the original structure.

Subcriteria:

- **Chronological or logical order:** Events or ideas follow the temporal or logical sequence of the source text.
- **Grouping of related information:** Related information is presented together.
- **Impact on comprehensibility:** The chosen order facilitates comprehension of the content.

Rating scale:

- **1 (Very poor):** Completely different sequence; confusion.
- **2 (Poor):** Many shifts that disturb the flow.
- **3 (Fair):** Some correct sequences, but also deviations.
- **4 (Good):** Order generally respected; minimal deviations.
- **5 (Excellent):** Order identical to that of the source text.

Linee guida per l'annotazione

Queste linee guida definiscono il processo e i criteri di valutazione di una sintesi di un testo in base a cinque dimensioni: coerenza, consistenza, rilevanza, fluidità e ordinamento. Per ciascuna dimensione, sono forniti i seguenti elementi:

1. Descrizione del compito
2. Definizione del criterio e dei sottocriteri di valutazione
3. Scala di valutazione da 1 a 5 con descrizione dei singoli livelli

Ogni riassunto presente nel foglio google dovrà essere valutato secondo i seguenti 5 criteri inserendo nella colonna omonima al criterio la sua valutazione da 1 a 5.

1. Coerenza

Descrizione del compito: Valutare quanto la sintesi presenti le informazioni del testo in modo logico e strutturato.

Definizione: La coerenza misura la fluidità e l’unità del testo, ovvero quanto le frasi scorrono in modo logico evitando passaggi bruschi o discontinui.

Sottocriteri:

- **Progressione logica delle idee**: Le idee sono presentate in un ordine che segue un filo logico e naturale.
- **Chiarezza e concisione**: Le frasi sono formulate in modo chiaro e sintetico.
- **Presenza di transizioni**: Sono utilizzati connettivi e transizioni per legare le frasi e i paragrafi.

Scala di punteggio:

- **1 (Molto scarsa)**: Testo disorganizzato, transizioni assenti, difficile seguire il filo del discorso.
- **2 (Scarsa)**: Struttura frammentata, passaggi bruschi, narrazione poco fluida.
- **3 (Moderata)**: Progressione generalmente logica, transizioni presenti ma irregolari.
- **4 (Buona)**: Struttura chiara, transizioni fluide, facile da seguire.
- **5 (Eccellente)**: Coerenza impeccabile, passaggi perfettamente raccordati.

2. Consistenza

Descrizione del compito: Verificare l’accuratezza fattuale della sintesi rispetto al testo originale.

Definizione: La consistenza misura la fedeltà dei fatti: assenza di contraddizioni, errori e informazioni non presenti nel testo sorgente.

Sottocriteri:

- **Accuratezza fattuale**: Tutte le affermazioni presenti nella sintesi corrispondono ai fatti espressi nel testo sorgente.
- **Assenza di contraddizioni**: Nessuna parte della sintesi contraddice ciò che è riportato nel testo originale.
- **Assenza di allucinazioni**: Non sono presenti informazioni inventate o aggiunte che non compaiono nel testo sorgente.
- **Inferenza logica**: Le informazioni dedotte sono coerenti e supportate dal contenuto del testo originale.
- **Allineamento terminologico**: I termini utilizzati nella sintesi sono coerenti con quelli del testo sorgente, soprattutto per concetti tecnici o specialistici.

Scala di punteggio:

- **1 (Molto scarsa)**: Numerose imprecisioni e dettagli inventati.
- **2 (Scarsa)**: Errori e discrepanze multiple.
- **3 (Moderata)**: Generalmente accurata, ma con piccole imprecisioni.
- **4 (Buona)**: Pochissime discrepanze trascurabili.
- **5 (Eccellente)**: Fedeltà totale ai fatti del testo.

3. Rilevanza

Descrizione del compito: Valutare se la sintesi include solo i contenuti essenziali del testo sorgente.

Definizione: La rilevanza misura l'inclusione dei punti chiave ed evita dettagli superflui o irrilevanti.

Sottocriteri:

- **Inclusione delle idee principali**: I concetti fondamentali del testo sorgente sono presenti nella sintesi.
- **Concisione**: Il contenuto è espresso in modo breve ma completo.
- **Assenza di ridondanze**: Non vi sono ripetizioni inutili.
- **Assenza di omissioni critiche**: Nessun concetto essenziale è stato omesso.

Scala di punteggio:

- **1 (Molto scarsa)**: Mancano quasi tutti i punti chiave; contiene molte informazioni irrilevanti.
- **2 (Scarsa)**: Copre alcuni punti principali ma tralascia aspetti importanti; presenza di dettagli superflui.
- **3 (Discreta)**: Copertura di oltre metà dei punti chiave; qualche ridondanza o omissione minore.
- **4 (Buona)**: Include quasi tutti i punti essenziali; minima ridondanza.
- **5 (Eccellente)**: Copre completamente i punti chiave; estremamente concisa.

4. Fluidità

Descrizione del compito: Valutare la qualità linguistica della sintesi: grammatica, ortografia, punteggiatura e stile.

Definizione: La fluidità misura la leggibilità e l'assenza di errori linguistici.

Sottocriteri:

- **Grammatica**: Assenza di errori grammaticali.
- **Ortografia**: Le parole sono scritte correttamente.
- **Punteggiatura**: Uso appropriato dei segni di punteggiatura.
- **Scelta lessicale**: Il vocabolario è appropriato e vario.
- **Struttura della frase**: Le frasi sono ben costruite e di lunghezza adeguata.

Scala di punteggio:

- **1 (Molto scarsa)**: Numerosi errori gravi che impediscono la lettura.
- **2 (Scarsa)**: Errori frequenti che ostacolano la comprensione.
- **3 (Moderata)**: Errori lievi non ostacolano la lettura.
- **4 (Buona)**: Isolati errori minori; lettura scorrevole.
- **5 (Eccellente)**: Testo impeccabile sotto ogni profilo.

5. Ordinamento

Descrizione del compito: Valutare se l'ordine delle informazioni nella sintesi rispecchia quello del testo sorgente.

Definizione: L'ordinamento misura l'allineamento della sequenza informativa con la struttura originale.

Sottocriteri:

- **Ordine cronologico o logico**: Gli eventi o le idee seguono la sequenza temporale o logica del testo sorgente.
- **Raggruppamento di informazioni correlate**: Le informazioni connesse tra loro sono presentate insieme.
- **Impatto sulla comprensibilità**: L'ordine scelto facilita la comprensione del contenuto.

Scala di punteggio:

- **1 (Molto scarsa)**: Sequenza completamente diversa; confusione.
- **2 (Scarsa)**: Molti spostamenti che disturbano il flusso.
- **3 (Discreta)**: Alcune sequenze corrette, ma anche deviazioni.
- **4 (Buona)**: Ordine generalmente rispettato; deviazioni minime.
- **5 (Eccellente)**: Ordine identico a quello del testo sorgente.

Cross-Genre Native Language Identification with Open-Source Large Language Models

Robin Nicholls [a,*] **and Kenneth Alperin** [b,**]

[a]University of Edinburgh
[b]MIT Lincoln Laboratory

Abstract. Native Language Identification (NLI) is a crucial area within computational linguistics, aimed at determining an author's first language (L1) based on their proficiency in a second language (L2). Recent studies have shown remarkable improvements in NLI accuracy due to advancements in large language models (LLMs). This paper investigates the performance of open-source LLMs on short-form comments from the Reddit-L2 corpus compared to their performance on the TOEFL11 corpus of non-native English essays. Our experiments revealed that fine-tuning on TOEFL11 significantly improved accuracy on Reddit-L2, demonstrating the transferability of linguistic features across different text genres. Conversely, models fine-tuned on Reddit-L2 also generalised well to TOEFL11, achieving over 90% accuracy and F1 scores for the native languages that appear in both corpora. This shows the strong transfer performance from long-form to short-form text and vice versa. Additionally, we explored the task of classifying authors as native or non-native English speakers, where fine-tuned models achieve near-perfect accuracy on the Reddit-L2 dataset. Our findings emphasize the impact of document length on model performance, with optimal results observed up to approximately 1200 tokens. This study highlights the effectiveness of open-source LLMs in NLI tasks across diverse linguistic contexts, suggesting their potential for broader applications in real-world scenarios.

1 Introduction

Native Language Identification (NLI) represents a critical area of study within computational linguistics, focusing on the determination of an author's first language (L1) through their written proficiency in a second language (L2). The relevance of NLI extends across various domains, notably in forensic linguistics for authorship profiling and in educational linguistics, where it aids in the customisation of teaching materials tailored to the linguistic background of L2 learners. The significance of NLI as a computational challenge was markedly enhanced following the release of the TOEFL11 corpus [2], a comprehensive dataset of non-native English writing, which has since served as a benchmark for advancing research in this domain.

Historically, NLI research has predominantly relied on traditional supervised learning methodologies [10]. However, recent advancements have undergone a paradigm shift towards employing large language models (LLMs), particularly leveraging zero-shot [14] learning and fine-tuning strategies [8, 11], with an emphasis on long-form essay-based datasets. While preliminary findings suggest the promise of LLM-based approaches in enhancing NLI performance, further empirical exploration across diverse real-world datasets remains imperative for elucidating the practical applicability of these methods.

In this paper we explore the effectiveness of open-source LLMs on short-form comments from Reddit. In Section 2, we discuss the related works for native language identification with large language models. In Section 3, we discuss the datasets, models, and prompting techniques we used for NLI, and evaluate performance on the long-form TOEFL dataset. More specifically, we fine-tune 3-billion and 8-billion parameter Llama-3 models [7] on the TOEFL11[2] training set, following a similar regime to Ng and Markov [11]. In Section 4, we seek to measure the transfer performance of these models from long-form to short-form text. More specifically, we explore the performance of those models on a subset of the Reddit-L2 corpus [12], acting as a validation set. Next, we perform the inverse of this by fine-tuning the same foundation models on Reddit-L2 and validating on TOEFL11. This approach seeks to answer whether the models are learning general linguistic characteristics, or simply over-fitting to the training set. Additionally, we introduce the sub-task of classifying an author of the Reddit-L2 dataset as being a native or non-native English author. We conclude in Section 5 and discuss limitations and next steps in Section 6.

2 Related Work

The progression of NLI research has been notably documented through workshops such as NLI-2013 [9] and NLI-2017 [10], which predominantly utilised the TOEFL11 corpus. These collaborative efforts underscored the effectiveness of ensemble methods, incorporating various traditional machine learning classifiers. These classifiers, trained on a diverse array of features including lexical, stylistic, and syntactic elements, demonstrate superior performance. Among the participants, the ItaliaNLP Lab [4] achieved remarkable accuracy, reaching a rate of 88.18% on the TOEFL11 test set, establishing a benchmark for subsequent research endeavours.

The first survey paper on NLI came out in 2024 [6]. Recent studies have ventured into the exploration of generative LLMs within the context of NLI, showcasing substantial advancements. An accuracy of 89.0% accuracy on TOEFL11 was achieved by fine-tuning GPT-2 [8]. Next, the application of GPT-3.5 and GPT-4 for zero-shot learning experiments on the TOEFL11 corpus set a new precedent in accuracy, achieving 91.7% with GPT-4 [14]. This exploration into the

[*] Corresponding Author. Email: r.nicholls@sms.ed.ac.uk.
[**] Corresponding Author. Email: kenneth.alperin@ll.mit.edu.

capabilities of LLMs revealed the potential for significant improvements in NLI accuracy. To our knowledge, LLMs have not been used yet on the Reddit-L2 dataset for NLI classification.

Further extending the boundaries of current methodologies, Ng and Markov [11] embarked on the approach of fine-tuning various open-source LLMs utilising 4-bit Quantization-aware training for Low Rank Adaptation (QLoRA) [5]. Their findings suggest a narrowing accuracy gap between fine-tuned LLMs and the zero-shot capabilities of GPT-4, with 8 billion parameter models nearly matching the performance of GPT-4 on the TOEFL11 dataset and surpassing it on the ICLE-NLI dataset. Such advancements underscore the rapidly evolving landscape of NLI research and its increasing reliance on the sophisticated capabilities of large language models.

3 Data and Models

The foundation of our study rests on two primary datasets: the ETS Corpus of Non-Native Written English (TOEFL11) and the Reddit-L2 corpus. To ensure a comprehensive evaluation, we utilise two LLMs of varying sizes. These were selected based on their strong performance on the TOEFL11 test set. This approach allows for an exploration of how model size impacts performance on non-native English text, providing insights into the scalability and efficiency of LLMs in handling linguistic diversity.

3.1 Data

ETS Corpus of non-Native Written English (TOEFL11) [2]: comprises 12,100 essays written by individuals across 11 L1 backgrounds (Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish) and provides a rich resource for analysing written proficiency in academic English. This corpus is divided into training, validation, and testing sets, containing 9,900, 1,100, and 1,100 essays, respectively balanced between the languages. This ensures an even distribution of L1 groups, which facilitates a balanced analysis of linguistic features across different language backgrounds.

Reddit-L2 Corpus [12]: represents a more informal register of English, encompassing approximately 250 million sentences written by over 45,000 authors. These authors were identified through the use of flairs, a metadata attribute in subreddits, which approximated the authors' L1 based on the national language of their indicated country. Although this method may introduce some inaccuracies in L1 attribution (such as with countries that have multiple national languages), due to misleading flairs or instances where an author's true L1 does not align with their country's primary language, the sheer volume of data is expected to mitigate the impact of such anomalies.

WI-LOCNESS [3]: was initially developed to support research in Grammatical Error Correction (GEC), as it comprises a total of 350 essays authored by both native English speakers and English language learners. Given that a portion of the essays originates from non-native English learners, the dataset is also suitable for tasks involving native versus non-native classification. We use approximately 150 tokens per essay to ensure consistency in document length, and a subset of 100 essays to maintain a balanced distribution between native and non-native authors.

To analyse these corpora, this study adopts a methodology [12] focusing on the masking of nouns through named entity recognition (NER) using the spaCy English core web text Transformer model. This approach aims to obscure specific lexical items, thereby compelling the analytical model to emphasise semantic understanding over mere lexical recognition. We hypothesise this strategy enhances the model's robustness by reducing its reliance on identifiable and proper nouns, which may vary significantly across L1s.

Table 1: Comparison of TOEFL11 and Reddit-L2

Measure	TOEFL11	Reddit-L2
Average Sentence Length	24.206±15.436	15.858±3.986
Proportion Unique Tokens	0.498±0.086	0.436±0.040
First Order Coherence	0.467±0.077	0.328±0.037
Second Order Coherence	0.456±0.097	0.314±0.039
Flesch Reading Ease	72.700±18.450	81.303±7.447

Table 1 shows some linguistic statistics of the test sets used from the TOEFL11 and Reddit-L2 datasets. Reddit has a much lower average sentence length than TOEFL does, which indicates Reddit has simpler sentences with less syntactic complexity. It also has a lower proportion of unique tokens, which indicates Reddit also has a simpler vocabulary and simplifed content. These measures definitely effect the coherence measures for the two datasets, as having shorter/choppier text can lead to more abrupt transitions and less flow, yielding the lower coherence measures for Reddit. Due to the simpler language employed on Reddit, it is easier to read than the TOEFL essays, indicated by the higher Flesch Reading Ease. Overall, these measures support the two datasets vary significantly in their linguistic properties.

Further, to assess the influence of text length on the accuracy of the model, subsets of the Reddit-L2 dataset were curated with fixed document lengths. This aspect of the study acknowledges that text length can be a confounding variable, potentially impacting the model's performance in identifying linguistic features characteristic of non-native English writing. This was done though to remove document length as a variable of comparing TOEFL11 and Reddit-L2, so the focus is on the writing of the text.

In the comparative analysis between the Reddit-L2 corpus and the TOEFL11 dataset, it is noteworthy that the Reddit-L2 corpus includes authors from an extensive array of 50 countries, whereas only a subset consisting of seven countries corresponds to the language backgrounds represented in the TOEFL11 corpus. Among these seven countries, the representation from China was deemed insufficient for meaningful analysis, thus necessitating its exclusion from certain comparative studies within this research. Hindi and Telugu, the two predominant languages in India, are both represented in TOEFL11. However, in Reddit-L2, Indian authors do not affiliate with a specific language. To address this discrepancy for validation tests involving the TOEFL11 dataset, predictions made by the model that assign either Hindi or Telugu to an author of Indian origin are considered accurate.

3.2 Dataset Topics

We conduct a qualitative analysis using Latent Dirichlet Allocation (LDA) topic modelling to evaluate the potential overlap of lexical-based features between the two datasets. We use LDA both on the overall datasets and on the L1 groups within them. Such an analysis is critical to ensure that topic bias does not influence feature selection. Prior to examining the topics identified through LDA, it is important to highlight the fundamental differences in the content of the two datasets. The TOEFL11 dataset consists of responses to eight standardised writing prompts, designed to ensure balanced representation of topics across the dataset. These prompts reflect the diversity of themes typically encountered in the TOEFL writing section. In contrast, the Reddit-L2 dataset consists of comments made in response to posts within specific subreddits related to Europe.

Generally, the TOEFL11 dataset includes topics such as education, travel, social life, and transportation, with similar thematic patterns observed across individual L1 groups. This consistency is likely attributable to the standardised nature of the prompts: while authors may draw upon their personal experiences to respond, their answers are constrained by the predefined topic of the questions. On the other hand, the Reddit-L2 dataset predominantly features discussions of public issues popular at the time the comments were posted. For instance, topics such as the Greek economy, Turkish-European relations, and terrorism were frequently observed. Importantly, within Reddit-L2, the topics vary significantly across individual L1 groups, aligning closely with issues that might be expected to be of particular interest to speakers of those languages. For example, German authors frequently discussed political ideologies, refugees, and the German language, whereas Turkish authors often focused on Islam, history, and Europe.

These differences in topic distributions across the two datasets support the hypothesis that classification is not solely driven by lexical-based features. However, the variation in topics within the Reddit-L2 L1 groups raises the concern that models trained on this dataset may become overly reliant on lexical features during fine-tuning. This highlights the importance of applying noun masking to the Reddit-L2 data, as this technique prevents the model from biasing toward topic-specific lexical features.

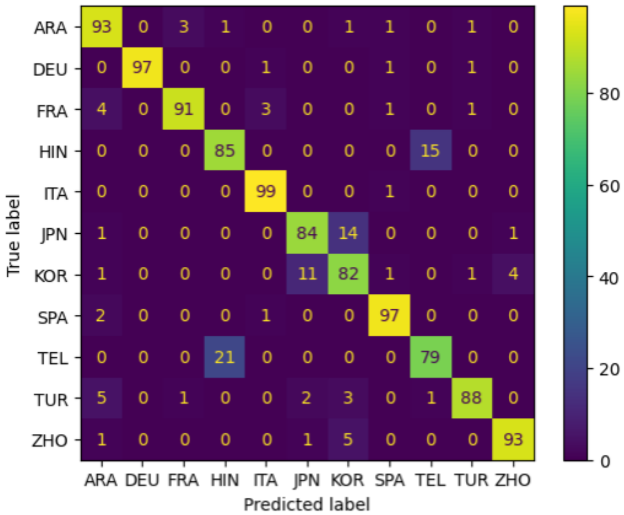
3.3 Models

As shown in Table 2, we compare the results of Ng and Markov [11] to our own study of similar open-source LLMs: Llama-3.2 (3B), Llama-3.1 (8B), Llama-3.1 (70B) [7], and Mixtral (8x7B) [13]. For all results, we provide the average accuracy score and standard deviation over three runs. All model temperatures and top-p values were set to 0.95 and 0.7. We used Llama-Factory [15] for fine-tuning and validation of all models.

Table 2: Comparative analysis of foundation and fine-tuned open-source LLM performance on TOEFL11 in terms of classification accuracy (%).

Model	TOEFL11 (11 L1s, test set) Closed-set
foundation	
Llama-3.2 (3B)	14.4 ± 0.0
Llama-3 (8B) Ng and Markov [11]	56.8 ± 1.1
Llama-3.1 (8B)	59.2 ± 0.0
Llama-3.1 (70B)	84.0 ± 0.0
Mixtral (8x7B)	67.2 ± 0.8
Gemma (7B) Ng and Markov [11]	13.6 ± 0.0
Phi-3 (3.8B) Ng and Markov [11]	18.2 ± 0.3
fine-tuned	
Llama-3 (8B) Ng and Markov [11]	85.3 ± 0.1
Llama-3.2 (3B)	86.8 ± 0.2
Llama-3.1 (8B)	90.0 ± 0.3
Gemma (7B) Ng and Markov [11]	90.3 ± 1.2

The selection criteria for advancing specific models to the fine-tuning phase are based on both baseline performance scores and practical considerations regarding model deployment. We observe that the 8B and 70B variants of the Llama model exhibit strong baseline performance. However, the decision to proceed with the fine-tuning of the 3B and 8B models, while excluding the 70B variant, is informed by a strategic preference for models that are compatible with typical consumer-grade systems. This choice reflects a pragmatic approach to model selection, aiming to balance the pursuit of high accuracy with the constraints imposed by the computational resources commonly available to end-users. Although our fine-tuned model does not achieve the same performance level as the fine-tuned Gemma 7B model [11], the advantage of using models from the same family is a significant consideration in our decision-making process.



True label	ARA	DEU	FRA	HIN	ITA	JPN	KOR	SPA	TEL	TUR	ZHO
ARA	93	0	3	1	0	0	1	1	0	1	0
DEU	0	97	0	0	1	0	0	1	0	1	0
FRA	4	0	91	0	3	0	0	1	0	1	0
HIN	0	0	0	85	0	0	0	0	15	0	0
ITA	0	0	0	0	99	0	0	1	0	0	0
JPN	1	0	0	0	0	84	14	0	0	0	1
KOR	1	0	0	0	0	11	82	1	0	1	4
SPA	2	0	0	0	1	0	0	97	0	0	0
TEL	0	0	0	21	0	0	0	0	79	0	0
TUR	5	0	1	0	0	2	3	0	1	88	0
ZHO	1	0	0	0	0	1	5	0	0	0	93

Figure 1: Confusion matrix for L1 accuracy per L1 on TOEFL11

3.4 L1 Analysis

Figure 1 shows the confusion matrix of the classification results for each language in the TOEFL test set on the fine-tuned Llama-3.1 8B model. For most of the languages, the model does a good job of correctly predicting the L1. The model performances lower on Telugu (79 %) than the others. This may be due to high confusion of Telugu with Hindi since they are both languages primarily in India. We also observe some minor overlap of Japanese and Korean, which makes sense as they are both East Asian languages and have a lot of similar syntactic features.

3.5 Prompting Technique

For the closed-set NLID task, we chose to use the prompts provided by Zhang and Salle [14]. For the native vs. non-native English classification, we modify their prompt and provide it in Figure 2. For all experiments, we employ iterative prompting. This allows us to continue to prompt the model, until the LLM returns an answer within the accepted criteria, or a maximum of five attempts have been made.

4 Reddit-L2 for NLI

4.1 Effect of document length

For our initial set of experiments, we utilise the subset of Reddit-L2 data described in the data section, which include German, French, Italian, Turkish, Indian (Hindi/Telugu), and Spanish. The dataset consists of 3,632 training documents and 909 testing documents. As illustrated in Figure 3, we conducted a comprehensive series of

```
<system>
You are a forensic linguistics expert that reads English texts written by
native and non-native authors in order to classify the authors as either:
"NATIVE: native English author
or
"NONNATIVE": Non native English author

Use clues such as spelling errors, word choice, syntactic patterns, and
grammatical errors to decide.
DO NOT USE ANY OTHER CLASS.

Valid output formats:
Class: "NATIVE"
Class: "NONNATIVE"

<user>
[document]

<response>
[predicted label]
```

Figure 2: Prompting template for native vs. non-native English classification.

experiments across varying document lengths to elucidate the significant influence that document length exerts on the model's proficiency in accurate L1 classification. The findings indicate that document length strongly correlates with an enhanced likelihood of feature manifestation. This correlation remains consistent up to an approximate threshold of 1200 tokens, beyond which the benefits of increased document length begin to exhibit diminishing returns.

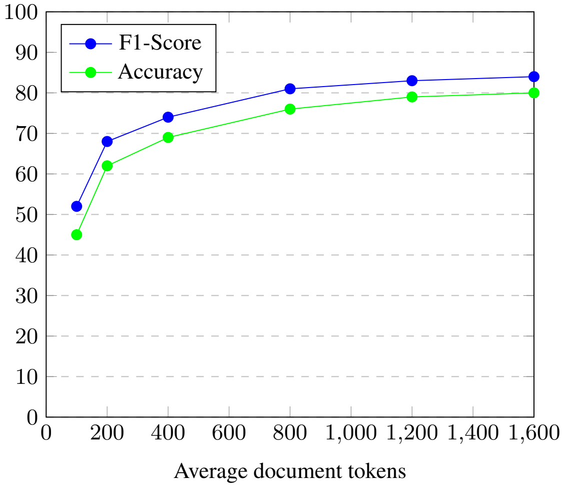


Figure 3: Effect of document length on Reddit-L2

4.2 LLMs fine-tuned on TOEFL11 and tested on Reddit-L2

When considering the results for document lengths of approximately 1200 tokens, the fine-tuned models exhibit a commendable ability to generalise to the Reddit authors. This achievement is particularly notable given the differences in genre and the application of noun masking. Specifically, the TOEFL dataset comprises short essays, whereas our construction of the Reddit data involves concatenated short-form comments. These comments, due to their nature, do not typically conform to a single coherent conversation, making them harder to follow as a unified passage. As noted in the data description, aggressive noun masking was applied to the Reddit data to ensure that semantic understanding, rather than mere lexical recognition, was required. This approach is particularly critical given the method used to identify Reddit authors. As many authors were sourced from country-specific subreddits, this frequently led to discussions about their home countries, potentially revealing their presumed native language to the language model.

Table 3: Comparative analysis of foundation and TOEFL11 fine-tuned models on Reddit-L2, 6 L1s, average 1200 tokens.

Model	Accuracy (%)	F1-score (%)
Llama-3.2 (3B)	19.3 ± 0.4	17.9 ± 0.5
Llama-3.1 (8B)	46.3 ± 1.7	53.1 ± 1.7
Llama-3.2 (3B) (fine-tuned)	66.7 ± 0.8	71.6 ± 0.8
Llama-3.1 (8B) (fine-tuned)	78.7 ± 0.2	83.1 ± 0.1

Given the limitation of using six L1s for this analysis, a random guess of the classification would yield an accuracy of approximately 16.7%. Table 3 shows that while the zero-shot Llama-3.1 (8B) model comfortably surpasses this baseline, the fine-tuned models improve on this by an additional 20% to 30%. This substantial enhancement clearly demonstrates that the linguistic features present in the TOEFL11 documents are also discernible in the Reddit-L2 data.

4.3 Reddit-L2 as a training set

Next, we perform the inverse of the first experiment to evaluate how a model fine-tuned on Reddit would perform on the TOEFL11 documents. For this purpose, we choose to fine-tune the model on five languages (French, Italian, Turkish, German, and Spanish), excluding Indian languages. This exclusion is due to the insufficient representation of Indian authors, with fewer than 100 authors, which is not adequate for fine-tuning.

As shown in Table 4, the baseline models struggle to classify accurately, with the Llama-3.2 (3B) model performing no better than a random guess (20%). Upon fine-tuning, both models significantly improve one their performance, with the 3B model nearly matching the baseline scores of the 8B model. Most notably, the fine-tuned 8B model achieves accuracy and F2 scores exceeding 90% on both the Reddit-L2 and TOEFL11 datasets, with the TOEFL scores being higher than those for Reddit-L2. This indicates that when trained on the Reddit authors, the model generalises exceptionally well to the TOEFL data. One plausible explanation for this is that TOEFL authors are generally intermediate learners, and as such, they may make more discernible errors more frequently. This characteristic makes TOEFL11 a relatively easier dataset for NLI tasks.

The results of the analysis suggest that noun masking is effective in mitigating the potential lexical-based bias. Additionally, the findings suggest that models trained on Reddit-L2 are able to classify TOEFL11 authors with a high degree of accuracy, relying on more than merely lexical topic-based features. This demonstrates the overall robustness of the proposed approach.

It is important to note that while our results exceed 90%, they should not be directly compared with previous works [14] and [11] since their research included the full set of TOEFL11 L1s. To make a proper comparison, we would need to source authors from the missing L1s.

4.4 Reddit-L2 for native vs nonnative classification

The identification of an author as either native or non-native can be beneficial across various fields, including educational strategies

Table 4: Comparative analysis of foundation and Reddit-L2 fine-tuned models on Reddit-L2 average 1200 tokens and TOEFL11 (test-set).

Model	Reddit-L2 (5 L1s)		TOEFL11 (5 L1s)	
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
Llama-3.2 (3B)	15.6 ± 0.0	15.2 ± 0.0	21.8 ± 0.0	19.5 ± 0.0
Llama-3.1 (8B)	48.4 ± 0.2	58.1 ± 0.1	68.4 ± 0.0	70.9 ± 0.0
Llama-3.2 (3B) (fine-tuned)	79.3 ± 0.1	85.5 ± 0.1	67.6 ± 0.1	65.5 ± 0.1
Llama-3.1 (8B) (fine-tuned)	92.0 ± 0.1	92.0 ± 0.1	93.8 ± 0.1	93.6 ± 0.1

Table 5: Comparative analysis of foundation and Reddit-L2 fine-tuned models on Reddit-L2 for native vs non-native, average 1200 tokens.

Model	Accuracy (%)	F1-score (%)
Llama-3.2 (3B)	51.5 ± 0.0	38.3 ± 0.0
Llama-3.1 (8B)	50.8 ± 0.2	39.8 ± 0.1
Llama-3.2 (3B) (fine-tuned)	97.6 ± 0.1	97.6 ± 0.1
Llama-3.1 (8B) (fine-tuned)	98.7 ± 0.1	98.7 ± 0.1

and security. The Reddit-L2 dataset comprises over 40,000 authors, with approximately 12,000 originating from native English-speaking countries. This substantial representation enables us to curate a subset of the dataset containing 24,000 authors, evenly divided between native English authors and a random sample of non-native English authors. We subsequently partition the data into training and testing sets with a 70:30 split.

Table 5 presents the results obtained using the base models and after fine-tuning. The base models perform no better than random guessing, predominantly classifying the majority of documents as non-native. However, once fine-tuned, the models exhibit remarkable performance on the testing set, achieving near-perfect accuracy.

Table 6: Comparative analysis of Reddit-L2 and WI-LOCNESS for native vs non-native. Model used: Llama3.1 (8B) fine-tuned on Reddit-L2.

Dataset	Accuracy	F1-score
Reddit-L2	$87.6\,\% \pm 0.0\,\%$	$87.6\,\% \pm 0.0\,\%$
WI-LOCNESS	$75.0\,\% \pm 0.0\,\%$	$73.3\,\% \pm 0.0\,\%$

To ensure that the models do not overfit on the data or merely identify lexical features, we employed the WI-LOCNESS [3] dataset as an evaluation set. For a fair assessment of the models' capabilities, we compared the validation results with those of the Reddit-L2 test set, limiting the document length to 150 tokens. This constraint was applied to ensure all documents were of similar length, providing the models with an equivalent number of tokens to analyse. As Table 6 shows, although there was a performance drop, the models still achieve reasonable scores on the WI-LOCNESS dataset.

This observation is particularly noteworthy given that, similar to the TOEFL11 corpus, the document style of WI-LOCNESS differs significantly from the concatenated short-form comments of Reddit-L2.

5 Conclusion

In this study, we explored the effectiveness of open-source LLMs in identifying native languages from short-form comments on Reddit, using both the TOEFL11 and Reddit-L2 corpora. Our findings highlight several key insights that contribute to the ongoing research in NLI.

Firstly, our experiments demonstrate that fine-tuning smaller Llama models (3B and 8B) on TOEFL11 can yield significant improvements in accuracy when applied to Reddit-L2 data. This suggests that the linguistic features captured from structured, academic English texts can generalise well to the more informal and varied language use on social media platforms like Reddit. The fine-tuned models significantly outperform baseline models, with accuracy improvements of 20% to 30%, indicating the transferability of learned linguistic characteristics across different text genres.

The inverse experiment of fine-tuning on Reddit-L2 and validating on TOEFL11 shows that models trained on informal text can also generalise effectively to more structured academic writing. The fine-tuned 8B model achieves accuracy and F1 scores exceeding 90% on both datasets, with higher performance on TOEFL11. This outcome underscores the robustness of the model in handling diverse linguistic contexts and suggests that models trained on a wide range of informal texts can successfully adapt to more formal writing styles.

Additionally, our study on classifying authors as native or non-native English speakers reveals that fine-tuned models could achieve near-perfect accuracy on the Reddit-L2 dataset. This classification task is crucial for various applications, including educational strategies and security measures. The models retain reasonable performance on the WI-LOCNESS dataset, further validating their generalisation capability.

One notable observation from our experiments is the influence of document length on model performance. We find that longer documents tend to provide more linguistic features that aid in accurate L1 classification, up to a threshold of approximately 1200 tokens. Beyond this point, the benefits of increased document length diminish. This finding is critical for future NLI research and practical applications, as it emphasises the need to balance document length with computational efficiency.

Our research contributes to the field of NLI by demonstrating the potential of open-source LLMs in handling diverse and informal text genres while maintaining high accuracy. The ability of these models to generalise across different datasets and writing styles highlights their versatility and applicability in real-world scenarios.

6 Limitations and Next Steps

While our study demonstrates the effectiveness of open-source LLMs for NLI, several limitations persist. Firstly, the Reddit-L2 dataset may overlap with the Llama models' pre-training data, potentially influencing the observed performance improvements. A curated dataset collected post-Llama release could mitigate this issue. Additionally, our exploration of document length effects is dataset-specific, requiring further validation across diverse text genres.

The scope of native languages (L1s) in our study is limited, restricting the generality of findings. Expanding the range of L1s, particularly underrepresented ones, is essential for broader applicability. Lastly, our focus on English as the target L2 leaves open the challenge of extending NLI to other L2s, particularly those with fewer resources and greater linguistic variation.

To address these limitations, we propose several directions for future work:

Curate a Post-Llama Reddit-L2 Dataset: Collect Reddit comments posted after Llama's release using the same collection method as the Reddit-L2 dataset to eliminate pre-training data overlap.

Expand Short-Form Sources: Evaluate models on short-form text from platforms like X or Discord to test robustness to different styles of writing (e.g. formality, target audience).

Increase L1 Diversity: Include low-resource native languages to improve multilingual applicability. Additionally, focus on collecting Reddit comments from a more diverse set of languages beyond primarily European languages.

Extend to Non-English L2s: Focus on NLI tasks for other L2s with limited data and greater linguistic diversity to see how methods generalize to other L2s.

Refine Document Length Insights: Investigate the impact of document length across varied text genres for optimal input design.

Conduct Native Language Style Transfer: Use a style transfer pipeline such as the one from Alperin et al. [1] to evaluate the quality of generating text to look like particular L2s (i.e. make non-native Spanish look native or vice versa).

These steps collectively aim to enhance the generalisation, robustness, and applicability of LLMs in NLI tasks.

References

[1] K. Alperin, R. Leekha, A. Uchendu, T. Nguyen, S. Medarametla, C. L. Capote, S. Aycock, and C. Dagli. Masks and mimicry: Strategic obfuscation and impersonation attacks on authorship verification. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 102–116, 2025.

[2] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15, 2013. doi: https://doi.org/10.1002/j.2333-8504.2013.tb02331.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.2013.tb02331.x.

[3] C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe. The BEA-2019 shared task on grammatical error correction. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, and T. Zesch, editors, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4406. URL https://aclanthology.org/W19-4406/.

[4] A. Cimino and F. Dell'Orletta. Stacked sentence-document classifier approach for improving native language identification. In J. Tetreault, J. Burstein, C. Leacock, and H. Yannakoudakis, editors, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 430–437, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5049. URL https://aclanthology.org/W17-5049/.

[5] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL https://arxiv.org/abs/2305.14314.

[6] D. Goswami, S. Thilagan, K. North, S. Malmasi, and M. Zampieri. Native language identification in texts: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3149–3160, 2024.

[7] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] E. Lotfi, I. Markov, and W. Daelemans. A deep generative approach to native language identification. In D. Scott, N. Bel, and C. Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1778–1783, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.159. URL https://aclanthology.org/2020.coling-main.159/.

[9] S. Malmasi, S.-M. J. Wong, and M. Dras. NLI shared task 2013: MQ submission. In J. Tetreault, J. Burstein, and C. Leacock, editors, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-1716/.

[10] S. Malmasi, K. Evanini, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, and Y. Qian. A report on the 2017 native language identification shared task. In J. Tetreault, J. Burstein, C. Leacock, and H. Yannakoudakis, editors, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5007. URL https://aclanthology.org/W17-5007/.

[11] Y. M. Ng and I. Markov. Leveraging open-source large language models for native language identification. In Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, and M. Zampieri, editors, *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.vardial-1.3/.

[12] E. Rabinovich, Y. Tsvetkov, and S. Wintner. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342, 2018. doi: 10.1162/tacl_a_00024. URL https://aclanthology.org/Q18-1024/.

[13] M. A. team. Mixtral of experts. https://mistral.ai/news/mixtral-of-experts, 2023. Accessed: 2025-04-22.

[14] W. Zhang and A. Salle. Native language identification with large language models, 2023. URL https://arxiv.org/abs/2312.07819.

[15] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models, 2024. URL https://arxiv.org/abs/2403.13372.

Climate Change Discourse Over Time: A Topic-Sentiment Perspective

Chaya Liebeskind[a,*,1] **and Barbara Lewandowska-Tomaszczyk**[b,1]

[a]Department of Computer Science, Jerusalem College of Technology, Israel
[b]Department of Language and Communication, University of Applied Sciences in Konin, Poland
ORCID (Chaya Liebeskind): https://orcid.org/0000-0003-0476-3796, ORCID (
Barbara Lewandowska-Tomaszczyk): https://orcid.org/0000-0002-6836-3321

Abstract. The present paper focuses on the study of opinion dynamics and opinion shifts in social media in the context of climate change discourse in terms of the quantitative NLP analysis, supported by a linguistic outlook. The research draws on two comparable collections of climate-related social media data from different time periods, each based on trending climate-related hashtags and annotated for relevant sentiment values. The quantitative computer-based research methodology has been supported by a language-based perspective in the pragma-linguistic form. The research shows that the latter data source, for the majority of identified topics, exhibits a significant reduction in negative sentiment and a dominance of positive sentiment, i.e., a potential temporal evolution in public sentiment toward climate change. To achieve this, we used a BERT-based clustering approach to identify dominant themes within a combined dataset of tweets from both periods. Subsequently, a unified sentiment classification framework using a Large Language Model (LLM) was applied to reclassify all tweets, ensuring consistent and climate-specific sentiment analysis across both datasets. This methodology allowed for a coherent comparison of public attitudes and their evolution in different time periods and thematic structures.

1 Introduction

Understanding human communication in its full richness requires delving beyond the mere factual content of utterances. It requires an exploration of the underlying layers of opinion, sentiment, and attitude [10] that imbue discourse with meaning, intention, and interpersonal resonance. These three interconnected yet distinct concepts form the background of subjective expression, shaping how individuals perceive, interpret, and react to the world around them. The manifestation and interplay of opinions, sentiments, and attitudes are ubiquitous, fundamentally influencing the dynamics of human interaction and the propagation of ideas.

The study of opinion, sentiment, and attitude in discourse has emerged as an increasingly critical area within diverse academic disciplines, including linguistics, computational linguistics, and others. This interdisciplinary focus reflects the impact these objective and subjective elements have on information processing, decision-making, and social cohesion.

Opinion can be broadly defined as a belief or judgment held by an individual about a particular subject. It represents a cognitive stance, intermixed with existential, epistemological and moral stances [11], perceived in terms of propositional structures, or else as proposed in Lewandowska-Tomaszczyk et al. [12] perceived in terms of a whole Opinion Event.

The topic of this study undertakes research in the dynamicity of opinions. Opinions are dynamic and can evolve over time as individuals and communities acquire new information, engage in critical reflection, or are exposed to diverse perspectives. In discourse, opinions are frequently expressed through explicit statements, but can also be inferred from argumentative structures, rhetorical devices, patterns of reasoning in terms of figurative, indirect or implicit structures [13].

Sentiment refers to the emotional tone or feeling conveyed by a piece of text or speech. It encapsulates the affective dimension of discourse, reflecting positive, negative, or neutral emotional or, more generally, affective, states [5]. For example, saying "I love this new policy" expresses a positive sentiment, while "I'm frustrated with the current situation" conveys a negative one. Sentiment is often, though not always, expressed through emotionally charged vocabulary (e.g., "amazing," "terrible," "joy," "anger"), but can also be conveyed through intonation, facial expressions, or linguistic cues like intensifiers or hedges. The granularity of sentiment analysis can vary, from broad polarity (positive/negative/neutral) to more fine-grained emotions (e.g., joy, sadness, anger, fear, surprise, disgust) and affective states such as curosity, reserve, boredom [5]. The prevalence of sentiment in online reviews, social media posts, and customer feedback has made sentiment analysis a cornerstone of natural language processing (NLP), enabling businesses to gauge public perception, track brand reputation, and identify emerging trends.

Attitude represents a more stable and enduring predisposition towards a person, object, idea, or issue [3]. It is a psychological construct that reflects an individual's overall evaluative stance, encompassing cognitive, affective, and behavioral components. Attitudes are often seen as the underlying drivers of opinions and sentiments. For example, a pro-environmental attitude might lead to the opinion that renewable energy is essential and positive sentiments toward green initiatives. Attitudes are more deeply ingrained than sentiments or even situation-specific opinions; they shape an individual's worldview and influence their long-term behaviors and decision-making. In discourse, attitudes are often less explicitly stated than opinions

[*] Corresponding Author. Email: liebchaya@gmail.com
[1] Equal contribution.

or sentiments, but can be inferred from consistent patterns of expression, recurring themes, and the overall evaluative orientation of an individual's communication. Analyzing attitudes in discourse involves a more holistic and interpretive approach, often drawing on psychological theories and qualitative methodologies.

The relationship between these three concepts is hierarchical and interdependent. Attitudes form the broadest and most stable foundation, influencing the opinions an individual holds, which in turn are often expressed with a particular sentiment. For example, a deeply ingrained attitude of skepticism toward government intervention might lead to the opinion that a new social welfare program is flawed, which could then be articulated with sentiment of anger or frustration. However, this relationship is not always linear or one-directional. A particularly strong sentiment or a newly formed opinion, especially when reinforced by social interaction, can sometimes contribute to the formation or modification of an attitude. The dynamic interplay between these layers makes their individual and collective analysis crucial for a comprehensive understanding of human communication.

To sum up, opinions tend to be more rooted in cognitive processes - they are often the result of consideration, evaluation, and assessment. They can be can be logical, informed, or even biased and can be strongly held or loosely formed. Sentiment, in this context, refers to the underlying emotional tone or feeling associated with an opinion, statement, or topic. It is the affective component of an opinion. Thus, it is most often a complex combination of affect (emotions, feelings or moods) and a cognitive process of thinking.

2 Influence of external events

The dynamism between opinions, emotions, and attitudes with respect to climate change, could be influenced externally both by external weather conditions and by political events. The period between 2015 and 2019 from which we draw our data stands as a pivotal time in the evolution of global climate consciousness, as it was during these five years that the tangible impacts of a warming planet became increasingly undeniable and visible to the broader public due to a series of major climate-related events, coupled with significant policy milestones, served as stark reminders of the escalating crisis, effectively shifting public perception from abstract threat to immediate reality. One of the most significant overarching trends during this period was the consistent shattering of temperature records and the increasingly evident link between these extreme events and human-induced climate change became a central theme in public discourse. The political landscape also played a crucial role in shaping climate consciousness during this period. The Paris Agreement on climate change[2], adopted in December 2015, marked a landmark moment. Furthermore, the Intergovernmental Panel on Climate Change (IPCC) Special Report on Global Warming of 1.5°C, released in October 2018, delivered a stark warning. On the other hand, Donald Trump's presidency (which began in January 2017) was marked by a significant departure from previous U.S. climate policy, actively seeking to dismantle environmental regulations and withdraw from international climate agreements[3]. Paradoxically though, Trump's actions and controversy also exerted impact on climate change awareness. And it was then, in August 2018, that Greta Thunberg[4] emerged as an unparalleled global icon for climate action during the latter part of this period, dramatically amplifying climate consciousness, especially among young people.

3 Computational and linguistic approaches

The rise of computational approaches, particularly in natural language processing (NLP) and machine learning, has revolutionized the study of opinion, sentiment, and attitude in discourse. The sheer volume of digitally available text data – from social media feeds and online forums to news articles and customer reviews – has created an unprecedented opportunity to analyze subjective language at scale. Early efforts in sentiment analysis focused primarily on the lexicons of positive and negative words, but more sophisticated techniques now employ machine learning algorithms, deep learning models, and contextual embeddings to capture the nuances of human emotion and subjective expression. These computational tools enable researchers and practitioners to automatically detect sentiment [4], identify opinion holders, track opinion evolution, and even infer underlying attitudes from large corpora of text.

However, the computational analysis of opinion, sentiment, and attitude is not without its challenges. The inherent subjectivity and context-dependency of human language pose significant hurdles. Sarcasm, irony, negation, and implicit expressions of sentiment can easily mislead automated systems. For example, "This movie was so good I fell asleep" expresses negative sentiment despite the positive word "good." Furthermore, differentiating between factual statements, expressions of opinion, and expressions of sentiment can be complex, as these often intertwine in natural discourse. The development of robust and accurate computational models requires sophisticated linguistic knowledge, large annotated datasets, and continuous refinement to account for the complexities of human communication.

Beyond computational approaches, traditional linguistic and discourse analytic methodologies remain indispensable for a deeper qualitative understanding. Discourse analysis and rhetorical strategies, for example, examines how opinions, sentiments, and attitudes are constructed, negotiated, and contested within specific communicative contexts. Semantics and pragmatics provide frameworks for understanding the meaning and intended effect of subjective language, considering factors such as the intention of the speaker, the shared knowledge, and the social norms. By integrating computational and qualitative approaches, researchers can achieve a more comprehensive and nuanced understanding of these multifaceted phenomena and uncover stances [1, 8, 9] towards climate change.

4 Aims of the analysis

The aims of our analysis are to uncover the social media users' attitudes, embracing sentiment, towards the climate change in two social media datasets from different periods of time and investigate to what extent the public opinions and attitudes towards this issue change within that period of time. As individuals categorize information based on their existing beliefs and attitudes, new information is evaluated in relation to these existing attitudes, leading to either the new information assimilation (acceptance) or contrast (rejection) [7]. A general theoretical approach adopted here refers to the identification of the users' opinions and attitudes via the sentiments expressed with reference to the opinions and argumentation persuasive effects,

[2] https://unfccc.int/process-and-meetings/the-paris-agreement

[3] https://www.nytimes.com/2017/06/01/climate/trump-paris-climate-agreement.html

[4] https://awpc.cattcenter.iastate.edu/2019/12/02/speech-at-cop24-dec-4-2018/

also in terms of Aristotle's persuasive appeals. An important part of this theory is played by three main factors - logic, emotion and trust-based types of persuasion. Logos is the appeal to the logic or reasoning by referring to evidence. Pathos is the appeal to the emotions of the audience. Trust-based ethos appeal invokes the credibility or trustworthiness of individuals considered the source of information. Sentiment marking in datasets can refer to any of the three appeals while opinions have specific content orientation (Mao et al. 2024). That is why some opinionated utterances can be ambiguous between sentence polarity and opinion polarity or between literal and non-literal (eg ironic) utterances such as e.g., You'd rather drown in these floods than admit global warming.

Previous attempts at opinion sentiment analysis are multiple and varied. In their (2024) review paper Mao et al. [15] critically assess available systems and highlight their challenges and limitations. To minimize the limitation of these methodologies, two types of sentiment analysis are used in the present study, one based on human marking, which can be more reliable but time-consuming and hence inefficient, and the other - on automatic lexicon-based labels - which is fast but can provide erroneous results with regard to the results of the opinion sentiment marking due to the above mentioned possible distinction between appeal polarity and opinion stance sentiments [12, 11]. As the results obtained were not fully consistent, we used LLMs to assist in the clarification of the distinction and to scrutinize a possible statistically significant sentiment flow between the earlier and later public opinion expression.

5 Literature review

The topic of automated sentiment analysis is not new in computational linguistics. The reference to and basing the identification on emotion word dictionaries is typically the first method to think of. However, the emotion sentiment values are often ambiguous outside context. Prediction of sentiment data with the use of machine-learning algorithms required heavy pre-processing preparations but it is machine learning that is regarded as being one of the latest and most prevalent techniques in sentiment analysis [2, 17]. Another related problem refers to an unreflexive use of lexicon sentiment analysis for opinion sentiment which could lead to erroneous and untrusthworthy results due to considering sentiment identification synonymous with opinion polarity marking. The most recent model of sentiment analysis and opinion mining is developed by Maruthupandi et al. [16]. This framework (SemAI) analyzes opinion sentiment by applying AI-based semantic analysis to assess and classify user views from social data. The authors use the Degree of Correlation Network Model (DCNM) to extract the subset of basic features from social network data with the meta-heuristics model - Heap based Optimization (HbO), to reduce the dimensionality of retrieved features. Finally, the unique classification algorithm, Self-Attention based Deep Analyzing Network (SA-DAN), is used to identify and categorize attitudes into positive, neutral, and negative. The SemAI model achieves particularly good results for accuracy, precision, as well as F1-scores in a variety of social data sets. Our approach is an attempt to reconcile the lexicon- and text-based sentiment value approaches via contextual BERT text embeddings, their clustering and topic labeling and interpretation to observe the dynamicity of opinion shift in public opinion sentiment values related to climate change.

6 Experiments

6.1 Datasets

To investigate the evolution of public opinion regarding climate change across different time periods, we employed two datasets composed of climate-related tweets. Each dataset captures sentiment toward climate change but differs in its time range, method of data collection, and sentiment labeling.

6.1.1 Twitter Climate Change Sentiment Dataset (2015–2018)

The first dataset, titled **Twitter Climate Change Sentiment Dataset**[5], was collected between April 27, 2015, and February 21, 2018, with support from the Canada Foundation for Innovation JELF Grant to Chris Bauch at the University of Waterloo. It consists of 43,943 tweets related to climate change, each manually annotated by three independent reviewers. Only tweets for which all three annotators reached unanimous agreement were retained, while others were discarded to ensure label reliability.

Each tweet in this dataset is assigned one of four sentiment categories:

- 2 (News): Links to factual news about climate change
- 1 (Pro): Supports the belief in anthropogenic climate change
- 0 (Neutral): Neither supports nor refutes the belief in anthropogenic climate change
- -1 (Anti): Rejects the belief in anthropogenic climate change

This dataset provides high-quality, human-annotated sentiment labels that are well-suited for training and evaluating classification models over historical climate discourse.

6.1.2 TwitterSocialMediaAnalysis_ClimateChange (2019)

The second dataset, referred to as **TwitterSocialMediaAnalysis_ClimateChange**[6], comprises tweets collected throughout 2019 via the Twitter API. More than 50,000 tweets were gathered from major cities across the United States and various international locations. Tweets were obtained using advanced search strategies, focusing on trending climate-related hashtags, including:

> #climateStrike, #climatestrike, #climatechange, #GreenNewDeal, #climatecrisis, #climateAction, #FridaysForFuture, #environment, #globalwarming, #GlobalWarming, #ActOnClimate, #sustainability, #savetheplanet, #bushfiresAustralia, #bushfires

To analyze public sentiment, two sentiment analysis tools were applied: TextBlob[7] and VADER (Valence Aware Dictionary and sEntiment Reasoner)[8]. In particular, VADER's compound sentiment score was averaged for each hashtag to quantify general sentiment trends. Unlike the first dataset, which contains categorical human labels, this dataset provides continuous sentiment scores based on automatic, lexicon-based sentiment analysis methods.

Together, these two datasets enable a comparative analysis of climate change sentiment across different years and geographic regions, while also contrasting manual and automated sentiment annotation approaches.

[5] https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset

[6] https://github.com/tasvora/TwitterSocialMediaAnalysis_ClimateChange

[7] https://textblob.readthedocs.io/en/dev/

[8] https://github.com/cjhutto/vaderSentiment

6.2 Topic Identification via BERT-Based Clustering

To explore how public sentiment toward specific climate-related topics evolved over time, we first aimed to identify dominant themes within these dataset. For this purpose, we employed a BERT-based clustering approach that allows for the extraction of semantically coherent topics from unstructured text.

Our method consists of the following key steps:

1. **Text Embedding with BERT** Each tweet was encoded into a high-dimensional dense vector using a pre-trained BERT model[6]. These embeddings capture contextual semantic information, allowing similar tweets to be represented by similar vectors in the embedding space. This step enables clustering to be based on meaning rather than surface-level lexical similarity.
2. **Clustering with k-means** The resulting tweet embeddings were clustered using the k-means algorithm. This unsupervised method partitions the data into k clusters by minimizing intra-cluster distances. Each cluster is interpreted as representing a distinct topic, with the assumption that tweets within the same cluster discuss related subject matter.
3. **Topic Labeling and Interpretation** To interpret each cluster, we extracted the most representative terms and manually reviewed example tweets. This allowed us to assign meaningful labels to clusters, such as "climate policy," "natural disasters," or "climate activism." These topics were then tracked across datasets to assess how sentiment shifted around each theme over time.

It is crucial to note that we first identify these topics from the combined dataset of both periods. Subsequently, we analyze the dominance of each identified topic within each individual period. This approach avoids the complex challenge of needing a mapping function to align similar topics across different periods. Such a mapping function could potentially complicate the comparison, as it might not be a one-to-one correspondence between a topic's representation in one period and its equivalent in another. By identifying topics from the outset on the aggregated data, we ensure a consistent thematic framework for cross-period analysis.

This BERT-based clustering framework enables nuanced detection of thematic structures in social media discourse and provides a robust basis for analyzing topic-specific opinion dynamics.

6.3 Unified Climate-Focused Sentiment Analysis

To enable a coherent comparison of public attitudes toward climate change across datasets, we applied a unified sentiment classification framework as a method to uncover public opinions. While both datasets contain some form of sentiment annotation, they differ in scope and interpretation. Our goal was to standardize sentiment analysis so that, for every tweet, the sentiment reflects its stance *specifically toward climate change*—independently of any associated political, social, or contextual references.

We adopted a three-class sentiment scheme: *positive*, *negative*, and *neutral*, referring strictly to the tweet's attitude toward climate change. However, a review of the original sentiment annotations revealed two key limitations. First, as noted in Section 6.1.1, the first dataset encodes sentiment in terms of belief in anthropogenic climate change, while the second uses general sentiment labels without topic-specific context. This mismatch complicates direct cross-dataset analysis.

Second, neither dataset consistently isolates sentiment toward climate change itself. Many tweets express sentiment toward peripheral topics such as politicians, movements, or media coverage, which may not reflect the author's view on the climate issue directly.

To resolve these issues, we employed a large language model (LLM) — specifically, `Meta-Llama-3-8B-Instruct`[9] — to reclassify all tweets under a consistent and climate-specific sentiment framework. We designed a prompt (Fig. 1) that instructs the model to assess each tweet's sentiment solely in relation to climate change. The model then outputs one of the three standardized sentiment labels.

> Sentiment Annotation Prompt for Climate Change Texts
>
> You are tasked with performing sentiment analysis on texts referencing climate change (e.g., "climate," "weather," "global warming"). Assign one of the following sentiment labels:
> **Positive:** Uses affirmative language about climate change without negation. Examples: "Global warming is harmful," "World leaders must act on climate change," or supportive figurative expressions.
> **Negative:** Denies or downplays climate change. Includes negation (e.g., "There is no global warming"), or sarcasm dismissing concerns.
> **Neutral:** Ambiguous or non-committal stance (e.g., "I don't know if climate change is taking place") or lacking clear opinion.
> Label each sample with only one word: **Positive**, **Negative**, or **Neutral**.
> Text to analyze: `{text}`

Figure 1. Prompt used for annotating sentiment in climate change-related texts.

This process allowed us to harmonize sentiment interpretation across both datasets and ensured that all subsequent analyses reflect genuine sentiment toward climate change, regardless of the dataset's original structure or annotation scheme.

By applying this unified sentiment framework to tweets clustered by topic—across both datasets—we are able to track how public sentiment toward each climate-related theme evolved between the earlier (2015–2018) and later (2019) periods. This sets the foundation for our temporal analysis of climate opinion dynamics, which is presented in the following section.

7 Results

7.1 Estimating the Number of Topics

As a preliminary step in the analysis, we aimed to estimate a suitable number of dominant topics represented in the combined tweet corpus. To that end, we applied the K-Means clustering algorithm with varying values of K, representing different potential topic counts.

To qualitatively assess the coherence and separability of the resulting clusters, we visualized the tweet embeddings using t-distributed Stochastic Neighbor Embedding (t-SNE). This dimensionality reduction technique maps high-dimensional vectors into a two-dimensional space while preserving the local structure of the data, thereby allowing us to visually evaluate the clustering structure.

[9] https://llama.meta.com

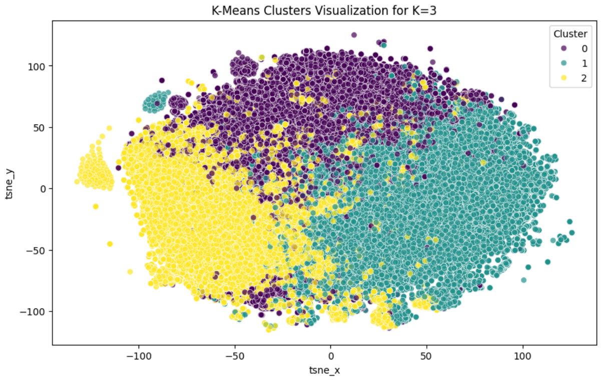


Figure 2. t-SNE visualization of tweet embeddings clustered using K-Means with $K = 3$.

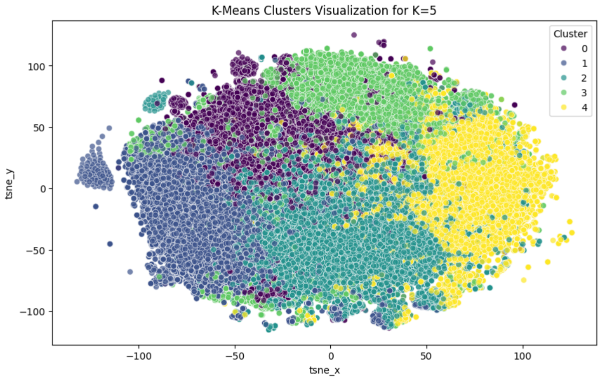


Figure 3. t-SNE visualization of tweet embeddings clustered using K-Means with $K = 5$.

Figures 2 and 3 present the clustering results for $K = 3$ and $K = 5$, respectively. The visualization for $K = 3$ reveals relatively well-separated clusters, suggesting distinct thematic groupings in the data. In contrast, the $K = 5$ configuration appears more fragmented, with less distinct boundaries between some clusters.

These visualizations served as an initial guide in selecting an appropriate number of topics for further analysis.

The distribution of tweets across the three clusters identified through K-Means clustering with $K = 3$ is as follows: Cluster 0 accounts for approximately 28.8% of the tweets, Cluster 1 for 38.8%, and Cluster2 for 32.4%. The specific thematic content of each cluster is interpreted in the next section.

7.2 Topic Interpretation Based on Clustered Keywords

To interpret the nature of each identified topic, we analyzed the most (30) frequent and representative words within each cluster. These clusters were extracted from the combined dataset of both periods. Table 1 lists the top terms for each cluster when applying K-Means with $K = 3$. The vocabulary associated with each cluster offers insight into the thematic content represented by the tweets within it.

Based on the prominent words in each cluster, we propose the following thematic interpretations:

- **Cluster 0 – Climate Change Controversy and Politics:** This cluster includes terms such as *trump*, *believe*, *real*, *president*, and *science*, suggesting a discourse focused on political debate and skepticism or affirmation regarding climate change. The presence of *epa* and *scientists* further indicates that this topic may center on the legitimacy of climate science and institutional responses.

Table 1. Top Words per Cluster ($K = 3$), extracted from the combined dataset of both periods.

Cluster 0	Cluster 1	Cluster 2
climate	climatechange	climate
change	climate	change
global	climatestrike	climatestrike
warming	climateaction	global
trump	climatecrisis	warming
climatechange	change	greennewdeal
climatestrike	amp	climatechange
amp	sustainability	today
world	environment	people
environment	globalwarming	amp
new	today	climateaction
people	people	like
real	greennewdeal	globalwarming
believe	world	believe
like	action	trump
science	new	environment
president	need	going
says	future	world
climatecrisis	global	need
epa	time	time
climateaction	planet	new
need	earth	actonclimate
fight	great	future
action	savetheplanet	day
scientists	youth	planet
time	day	la
energy	like	real
going	energy	strike
years	actonclimate	thank
planet	green	right

- **Cluster 1 – Climate Activism and Global Sustainability:** This cluster features frequent hashtags and terms associated with organized movements and environmental advocacy, such as *climatestrike*, *climateaction*, *sustainability*, *greennewdeal*, and *savetheplanet*. It likely represents the narrative of global climate activism, particularly involving youth movements and calls for systemic change.
- **Cluster 2 – Climate Events and Urgency:** With terms like *today*, *going*, *strike*, *thank*, and *right*, this cluster appears to reflect real-time reactions to climate-related events, including public demonstrations and environmental crises. The co-occurrence of *greennewdeal* and *actonclimate* supports the interpretation that this topic captures urgent, event-driven discourse.

7.3 Sentiment Dynamics Within Topics

The stacked bar chart (Figure 4 illustrates a comparative analysis of sentiment distribution across three distinct topics (labeled 0, 1, and 2) between two distinct data sources (referred to as Source 1 and Source 2). Each bar represents a specific topic within a given data source, segmented to display the proportions of negative, neutral, and positive sentiment. A general observation reveals a prevalence of positive sentiment across most topic-source combinations. However, notable variations emerge upon closer examination. For Topic 0, Source 1 exhibits a considerably higher proportion of negative sentiment compared to Source 2, which displays a markedly lower negative sentiment and a correspondingly higher positive sentiment. In contrast, Topic 1 demonstrates a relatively consistent sentiment distribution across both sources, with a dominant positive sentiment

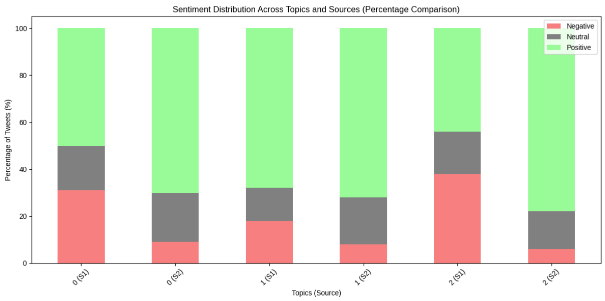


Figure 4. Sentiment distribution across three topics for two data sources, highlighting temporal differences in public sentiment.

and similar, albeit slightly elevated in Source 1, levels of negative and neutral sentiment. The most pronounced difference is observed in Topic 2, where Source 1 is characterized by the highest proportion of negative sentiment among all combinations, alongside a substantial neutral sentiment. Conversely, Source 2 for Topic 2 shows a significant reduction in negative sentiment and a dominance of positive sentiment. In conclusion, the observed variations in sentiment distribution between Source 1 and Source 2 for each topic suggest a potential temporal evolution in public sentiment, as each source represents a distinct time period.

To determine whether the distribution of sentiment categories (positive, neutral, negative) significantly differs between the two sources within each topic cluster, we conducted a Chi-Square Test of Independence for each cluster separately. This test evaluates whether the proportions of sentiment labels are independent of the data source, while accounting for the differences in sample size between sources. The results revealed statistically significant differences in sentiment distribution across all three clusters. In each case, the test yielded a p-value less than 0.01, indicating that the proportion of sentiment types (e.g., the ratio of positive to negative tweets) changed meaningfully between the two time periods represented by the sources. These findings suggest a temporal shift in sentiment expression within each identified topic.

8 Discussion and conclusions

The main aim of this study was to investigate how public opinions expressed in terms of sentiment toward climate change evolved over time. As is known sentiment analysis can be performed at various granularities, each offering different insights, such as the word level, document-, sentence- or aspect-levels. The closest brief description of our methodology employed in the study is a hybrid approach, i.e., a combination of some of these levels. We first specified particular dominant climate-related topics that evolved over time in the two datasets we explored by employing a BERT-based clustering approach to extract semantically coherent topics from unstructured texts. Each tweet was encoded into a BERT-related vector, capturing contextually relevant semantic information. The problem with the two datasets we used was that they were not identical in scope and in the contextual information annotation. Some of the sentiment values were assigned not to general topics or main - climate change - targets, but to some lexical entities, peripheral from the point of view of a particular text opinion sentiment. The BERT transformer advanced the study, more effectively capturing long-range dependencies and contextual relationships.

The analysis revealed dynamic patterns in climate change discourse, frequently characterized by polarization and shifts in dominant narratives. A notable observation was how public expressions of sentiment modulates, often reflecting concern, anxiety, or fear, probably intensified during extreme weather events or following significant climate-related news. This suggests a direct link between real-world occurrences and shifts in public opinion, demonstrating the dynamic, responsive nature of public sentiment and highlighting the importance of real-time monitoring for climate communication strategies.

An important part of our study was a distinction between lexical emotion sentiment and more general, opinion thematic sentiments that might often contradict. This layering of opinion via the opinion sentiment targets might be an important site of opinion and attitude sentiment distinction in our further studies. There are new proposals to extend the effectiveness of sentiment identification in data such as e.g., semantic analysis based on AI (SemAI) for assessing and categorizing user views from social data mentioned in the Literature review secton here, which applies algorithms in the stages of feature extraction, optimization, and sentiment prediction [16] with The Degree of Correlation Network Model (DCNM) and the Heap based Optimization model, supported by Self-Attention based Deep Analyzing Network (SA-DAN). And yet, despite significant advancements, opinion-related sentiment analysis continues to face several challenges that would be undertaken in the future studies. They are primarily related to implicit meanings such as figurative language or sarcasm and irony detection. Another hurdle, often requiring deep contextual and commonsense understanding, is handling the scope of negation and intensification. As is also encountered in our analyzed data, there are still problems with context and domain dependency as the opinion sentiment can change across contexts and domains as well as target orientation that can vary depending on whether we identify sentiments towards particular entities expressed in lexical units or whether the sentiment of the whole entity (e.g., post or document).

An important methodological asset of our study lies in the integration of lexicon-based tools, human annotation, and LLM-powered reclassification within a unified framework for sentiment analysis. The use of BERT embeddings allowed for a semantically nuanced clustering of topics beyond surface-level lexical similarity, while LLMs enabled a more contextualized and climate-specific reinterpretation of sentiment. This computational pipeline ensured interpretability and adaptability across datasets with divergent structures.

Comparing Source 1 and Source 2 climate change Twitter data, some of the opinion sentiment marking agrees with human contextual sentiment recognition (see Table 2). For example:

> @GretaThunberg Thank you for standing up #ClimateStrike

The two sentiment models—VADER and LLMs—mark the positive sentiment of the above opinion consistently.

However, several challenges remain: sentiment models, especially lexicon-based ones like VADER, are sensitive to contextual ambiguity, figurative language, and sarcasm. Both VADER and even LLMs, despite their contextual strength, may struggle with stance disambiguation when multiple sentiment targets (e.g., politicians vs. climate change) co-occur. An example of such opinion sentiment recognition problems can be encountered in the following case:

> @tiniebeany climate change is an interesting hustle as it was global warming but the planet stopped warming for 15 yes while the suv boom (Dataset 1)

Such and similar posts are indirect and are interpreted as denying the reality of global climate change, based on the uninformed

Dataset	#	Example	Sentiment	LLM	VADER
1	1	Fabulous! Leonardo #DiCaprio's film on #climate change is brilliant!!! Do watch. https://t.co/7rV6BrmxjW via @youtube	Positive	Positive	Positive
1	2	@tiniebeany climate change is an interesting hustle as it was global warming but the planet stopped warming for 15 yes while the suv boom	False	Neutral	Positive
2	1	The people on #ClimateStrike in DC are marching to confront the banks funding the climate crisis! #FireDrillFriday #ShutDownDC	Negative	Positive	Negative
2	2	@realdonaldtrump is presiding over a mass extinction and climate catastrophe. We need to take action now! #ClimateStrike shutdowndc923 @Trump International Hotel Washington, D.C.	Negative	Positive	Negative
2	3	@GretaThunberg Thank you for standing up #ClimateStrike	Positive	Positive	Positive

Table 2. Sentiment comparison between LLM and VADER across two datasets

or disinformed types of opinion. The opinion sentiment is **Negative**, even though the positive lexicon sentiment lexis—"interesting" or "stopped warming"—is used. Such contextual clues may be misleading and cause LLM **Neutral** marking and **Positive** by VADER. Problems with LLM-generated correct identification of implicit, as well as some of the indirectly conveyed opinions, were first reported in Liebeskind and Lewandowska-Tomaszczyk [14]. In our present paper more nuanced opinion recognition problems are noted. Neither VADER nor LLM recognized the anti-climate change opinion sentiment, although the lexical form "hustle" may function as an interpretative indicator.

Another example (Dataset 2) also presents questionable climate change sentiment recognition:

> The people on #ClimateStrike in DC are marching to confront the banks funding the climate crisis! #FireDrillFriday #ShutDownDC

VADER marks the post sentiment as **Negative**, and LLM as **Positive**, while the post is immersed in a more complex outside context—#ClimateStrike confronting bank funding of investments which accelerate climate change. In other words, the post includes two opinion sentiment stances: climate change recognition (**Positive**) and opposition to the institutions contributing to the process (**Negative**).

Thus, while our methodology enables a richer analysis of opinion dynamics, it also underscores the need for further refinement in handling subtle pragmatic cues and layered sentiment targets. At the same time, this research unequivocally demonstrates that universal opinion is not solely shaped by one mode of persuasion, but rather by a dynamic interplay of appeals. While logos-based arguments provide the foundational framework of factual evidence and logical reasoning that underpins informed perspectives, they are often insufficient on their own to influence widespread belief. We found that emotional appeals are equally critical, tapping into shared human experiences, values, and sentiments to foster resonance and engagement. Furthermore, the ethos of the source—its credibility, authority, and perceived trustworthiness—plays a pivotal role in determining the receptiveness of an audience to any given message such as eg. in the case of Greta Thunberg.

Our analysis, while comprehensive within its defined parameters, may be limited by its focus on specific social media platforms (Twitter) and a fairly short time frame. This can affect the generalizability of findings across more diverse textual sources (e.g. news articles, forums), different cultural contexts, or linguistic variation. One can also indicate a general lack of consideration for user demographics (inaccessible in social media data) and entirely external factors that shape opinions in climate change discourse, a limitation that may also apply to aspects of this study. Yet, in our attempt to use topic cluster modeling and fairly intensive scrutiny, some of these threats may have been overcome.

There are also specific linguistic-pragmatic features that signal subtle shifts in meaning or stance [8]. They are hedging devices (e.g., "perhaps," "it seems") and modal verbs (e.g., "may," "might," "could"), which signal caution, uncertainty, or indirectness in expressing claims, first-person pronouns, indicating a subjective stance and authorial presence, nominalizations (such as "global warming") and passive voice constructions (e.g., "it was observed"), which contribute to informational density and objectivity, often favored in factual or scientific discourse. Furthermore, the pragmatic analysis can delve more deeply into the contextual and consequential dimensions of language (presupposition, implicature, speaker's intent) and, finally, implicit meanings detection (figurative language, irony, sarcasm).

Needless to say, our analysis mirrored to a large extent all those important political events that took place in the period between 2015 - 2019. The generated thematic clusters contained items such as TRUMP and GRETA as their main headwords. Hence, future investigations should also extend research towards more effective modeling of real-world debates on complex societal issues. Conducting comprehensive cross-cultural and cross-linguistic analyses of climate narratives and opinion change is also vital, exploring how language, cultural norms, and communication styles influence opinion dynamics globally.

Incorporating more realistic social network structures into opinion dynamics simulations will better reflect how opinions propagate and change within diverse social groups. Furthermore, a significant leap from merely detecting opinion change to understanding how to influence it responsibly involves investigating causal relationships between narratives, linguistic shifts, and real-world outcomes, moving beyond mere correlation to understand how specific narratives directly influence behavior and societal change.

Future research might also prioritize refining LLMs by fine-tuning them with more diverse and real-world human discourse data, including content that reflects polarized or fact-resistant opinions. This is crucial to overcome current LLM biases, negatvely influencing e.g., word embeddings, thereby enabling more accurate simulations and analyses of complex opinion dynamics.

Further exploration of hybrid computational-linguistic models is warranted, combining the strengths of advanced LLMs with explicit linguistic and pragmatic theories to enhance accuracy, interpretability, and detection of subtle opinion shifts. Developing more nuanced and multidimensional representations of opinion, moving beyond scalar sentiment or stance to capture the complexity of human beliefs and attitudes, is also a key direction.

Following this line of enquiry, our further research in opinion-related sentiment analysis is thus likely to use a hybrid linguistic

NLP research methodology and focus on more robust contextual understanding that can better capture implicit and nuanced language. Furthermore, it goes without saying, that incorporating deeper aspects of common sense reasoning would improve opinion sentiment understanding.

For future work, exploring alternative topic modeling approaches beyond BERT-based clustering, such as Latent Dirichlet Allocation (LDA), could offer complementary insights into the underlying thematic structures of climate change discourse. While BERT excels in capturing semantic nuances through contextual embeddings, probabilistic topic models like LDA could provide a different perspective on topic composition by identifying latent topics based on word co-occurrence patterns, potentially revealing distinct and interpretable themes. This comparative analysis would enrich our understanding of topic dynamics and validate the robustness of identified themes across different methodological paradigms.

Acknowwledgements The study was prepared in the Linguistics group of COST Action CA 21129 *What are Opinions? Integrating Theory and Methods for Automatically Analyzing Opinionated Communication* (OPINION). The authors thank Avigail Haliva and Shilat Shimon for their significant assistance in implementing the code and running the experiments. Their contribution played an important role in the technical execution of the study.

References

[1] A. Addawood, J. Schneider, and M. Bashir. Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th international conference on social media & society*, pages 1–10, 2017.

[2] H. J. Alantari, I. S. Currim, Y. Deng, and S. Singh. An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *International Journal of Research in Marketing*, 39(1):1–19, 2022.

[3] D. Albarracin and S. Shavitt. Attitudes and attitude change. *Annual review of psychology*, 69(1):299–327, 2018.

[4] A. Chourasiya, A. Khan, K. Bajaj, M. Tomar, T. Kohli, and D. Chauhan. A review of sentiment analysis and emotion detection from text using different models. 2025.

[5] G. Colombetti. *The feeling body: Affective science meets the enactive mind.* MIT press, 2014.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[7] M. E. Doherty and E. M. Kurz. Social judgement theory. *Thinking & Reasoning*, 2(2-3):109–140, 1996.

[8] J. W. Du Bois. The stance triangle. In *Stancetaking in discourse: Subjectivity, evaluation, interaction*, pages 139–182. John Benjamins Publishing Company, 2008.

[9] L.-W. Ku, C.-Y. Lee, and H.-H. Chen. Identification of opinion holders. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 4, December 2009*, 2009.

[10] B. Lewandowska-Tomaszczyk. Online interconnectivity and negative emotion patterning. *Sociedad de la Información*, 44:76–109, 2013.

[11] B. Lewandowska-Tomaszczyk and C. Liebeskind. Opinion events and stance types: advances in llm performance with chatgpt and gemini. *Lodz Papers in Pragmatics*, 20(2):413–432, 2024. doi: doi:10.1515/lpp-2024-0039. URL https://doi.org/10.1515/lpp-2024-0039.

[12] B. Lewandowska-Tomaszczyk, C. Liebeskind, A. Bączkowska, J. Ruzaite, A. Dylgjeri, L. Kazazi, and E. Lombart. Opinion events: Types and opinion markers in english social media discourse. *Lodz Papers in Pragmatics*, 19(2):447–481, 2023.

[13] C. Liebeskind and B. Lewandowska-Tomaszczyk. Navigating opinion space: A study of explicit and implicit opinion generation in language models. In *Proceedings of the First LUHME Workshop*, pages 28–34, 2024.

[14] C. Liebeskind and B. Lewandowska-Tomaszczyk. Opinion identification using a conversational large language mode. In *The International FLAIRS Conference Proceedings*, volume 37, 2024.

[15] Y. Mao, Q. Liu, and Y. Zhang. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University-Computer and Information Sciences*, page 102048, 2024.

[16] J. Maruthupandi, S. Sivakumar, V. S. Kumar, and P. B. Srikaanth. Semai: A unique framework for sentiment analysis and opinion mining using social network data. *SN Computer Science*, 6(2):99, 2025.

[17] H. Zhao, Z. Liu, X. Yao, and Q. Yang. A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. *Information Processing & Management*, 58 (5):102656, 2021.