LLMs stick to the point, humans to style: Semantic and Stylistic Alignment in Human and LLM Communication

Noé Durandard[1], Saurabh Dhawan[2], Thierry Poibeau[1]
[1]LATTICE, ENS-PSL, U. Sorbonne Nouvelle Paris 3, CNRS, Paris, France,
[2] Technical University of Munich, Munich, Germany
Correspondence: noe.durandard@psl.eu

Abstract

This study investigates differences in linguistic accommodation—changes in language use and style that individuals make to align with their dialogue partners—in human and LLM communication. Specifically, it contrasts semantic and stylistic alignment within question-answer pairs in terms of whether the answer was given by a human or an LLM. Utilizing embedding-based measures of linguistic similarity, we find that LLM-generated answers demonstrate higher semantic similarity—reflecting close conceptual alignment with the input questions—but relatively lower stylistic similarity. Human-written answers exhibit a reverse pattern, with lower semantic but higher stylistic similarity to the respective questions. These findings point to contrasting linguistic accommodation strategies evident in human and LLM communication, with implications for furthering personalization, social attunement, and engagement in human-AI dialogue.

1 Introduction

Human dialogue comes with a range of social and psychological adaptations (Chartrand and van Baaren, 2009; Giles et al., 1991; Niederhoffer and Pennebaker, 2002; Pickering and Garrod, 2004), that are central to mutual understanding and social attunement between two dialogue partners (Chartrand and Bargh, 1999; Garrod and Pickering, 2004; Giles and Ogay, 2007). Much as dialogue with Large Language Models (LLMs) has come to offer a close approximation to that with humans, it remains unclear whether or not such adaptive behaviors might already have been learned by LLMs owing to their training on massive amounts of human-produced texts. As LLMs are increasingly being deployed in what are essentially social settings, such as teaching and mental health counseling, understanding of these socially adaptive behaviors has assumed greater relevance (Belosevic and Buschmeier, 2024). Linguistic accommodation is one such behavior (Giles and Ogay, 2007). It refers to adjustments in language use and style that individuals make to align with their dialogue partners. It has been shown to be a distinctive feature of human dialogue and plays an important role in effective interaction, facilitating social approval and mutual understanding (Giles et al., 1991).

A range of recent studies have compared linguistic differences between human and LLM generated text. Zhou et al. (2023) compared AI-created and human-created misinformation and found that AI-generated misinformation exhibited enhanced emotional content and used more salient expressions. Herbold et al. (2023) compared human-written versus ChatGPT-generated argumentative student essays and found that expert teachers rated the ones by ChatGPT higher. They also found that while humans used more modals and epistemic markers, ChatGPT wrote more complex sentences and used more nominalizations. Muñoz-Ortiz et al. (2024) compared human-written English news text with that generated by LLMs, and found that human texts exhibit more scattered sentence length distributions, a distinct use of dependency and constituent types, and more aggressive emotions (fear, disgust), while LLM outputs showed more markers of objective language. Cai et al. (2024) conducted a range of psycholinguistic tests and found that, compared to LLMs, humans, preferred using shorter words to convey less informative content, and showed higher use of context to resolve syntactic ambiguities.

While the aforementioned studies compared linguistic properties of human and LLM generated text corpora, in this study, we compare linguistic properties within specific question-answer pairs, where the answers were given either by a human or an LLM. Toward this purpose, we extend a dataset of human-human and human-LLM QA pairs (Guo et al., 2023), to a range of new LLMs. Subse-

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 206–213
Aug 25–27, 2025.

quently, using embedding-based measures, we calculate semantic (Solatorio, 2024) and stylistic (Patel et al., 2025) similarities between each QA pair. This allows us to examine whether humans and LLMs make adjustments to the language style and usage in their responses, in order to better align with the style and usage evident in the questions from their dialogue partners.

2 Methods

The developed analytical framework proposes to study linguistic alignment within human-human and human-LLM QA pairs through their vector representations within high dimensional textual embedding spaces.

2.1 Corpus

We start by extending a dataset of human-human and human-LLM QA pairs from Guo et al. (2023), to a range of new LLMs. The Human ChatGPT Comparison Corpus (HC3) (Guo et al., 2023) gathers a set of open-ended questions derived from five different online sources. Each question is associated with a human-written answer, as well as the response given by ChatGPT when prompted with the same question. Thus, this corpus provides an explicit comparison between human-written and LLM-generated responses to questions covering different topics and writing styles.

In the context of this study, 800 questions are randomly sub-sampled from the four splits that contain organic QA pairs: *medicine* (from Medical Dialog dataset (Zeng et al., 2020)), *reddit_eli5* (from ELI5 dataset (Fan et al., 2019)), *finance* (from FiQA dataset (Maia et al., 2018)), and *open_qa* (from WikiQA dataset (Yang et al., 2015)). These samples are then augmented with responses generated by seven LLMs other than ChatGPT. This process results in the subsampled Human LLM Comparison Corpus (H-LLMC2), a parallel corpus of 3 200 questions, balanced between the four selected sources, each associated with a human answer and the outputs of LLMs (compared to 24 322 questions with parallel human and ChatGPT answers for HC3)[1].

2.2 Evaluated Models

H-LLMC2 is assembled by including responses generated by smaller open-weights LLMs from different families, and resulting from different instruction tuning and reinforcement learning paradigms.

In addition to ChatGPT, responses to the questions are generated with Llama 2 Chat models (7B, 13B) (Touvron et al., 2023), Llama 3.1 Instruct models (8B) (Grattafiori et al., 2024), Qwen 2.5 Instruct models (7B, 14B) (Team, 2024), as well as corresponding reasoning-distilled versions from DeepSeek (8B, 14B) (DeepSeek-AI, 2025).

2.3 Encoder-based analytical framework

Modern neural encoder models are leveraged in order to compare human-written questions with human-produced and LLM-generated answers in their similarities and particularities.

2.3.1 Encoders

High-dimensional representations of the texts are obtained with different encoder-only models trained to represent different textual features, while focusing respectively on content and form. Both of the encoder models, selected to represent semantic and stylistic features, rely on contrastive learning but use different data and objectives during training, hence optimizing different properties.

Semantic Space `GIST-Embedding-v0`[2] (Solatorio, 2024) is a model fine-tuned on top of `bge-base-en-v1.5` (Xiao et al., 2023), which demonstrates SoTA performances across various tasks from MTEB (Muennighoff et al., 2023). This model is selected to embed texts within a semantically coherent space.

Stylistic Space `styledistance`[3] (Patel et al., 2025), fine-tuned on top of `roberta-base`, was selected to produce representations of the texts' style as it embeds texts with similar writing styles closely and texts with different styles far apart, regardless of content. The model was trained on 40 stylistic features spanning seven broad groups—syntactic features (e.g., passive vs. active voice), graphical and digital features (uppercase, emoji, text-emoji), emotional & cognitive tone, stylistic/aesthetic devices (formality, metaphors, humour), social and interpersonal features (politeness, offensiveness, self- vs. audience focus), lexical preferences (long words, nominalisations) and temporal/aspectual

[1] The dataset is available at https://huggingface.co/datasets/noepsl/H-LLMC2. The code is available at https://github.com/d-noe/LLM_emb_CAT.

[2] https://huggingface.co/avsolatorio/GIST-small-Embedding-v0

[3] https://huggingface.co/StyleDistance/styledistance

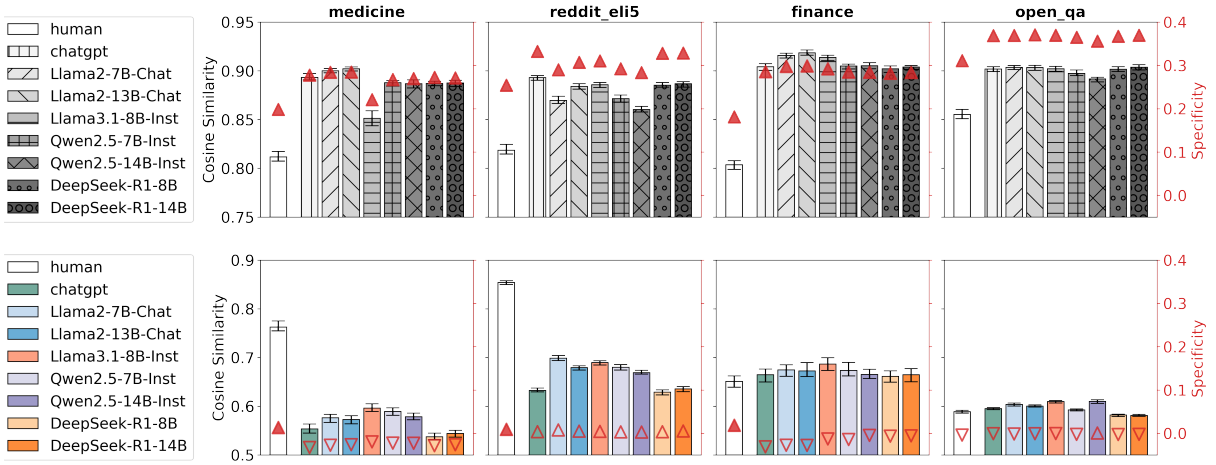

Figure 1: Median pairwise cosine similarities between questions and answers (bars) and specificity scores (in red), grouped by data sources, within the semantic representation space (top) and stylistic one (bottom). Error bars represent 95% bootstrapped confidence intervals. For specificity, in red, triangles up and down represent scores above, resp. below, zero. Filled triangles indicate scores significantly above zero according to a one-sample Wilcoxon signed-rank test ($p < 0.05$).

framing (usage of present vs. past or future focussed language).

2.3.2 Measures in Representation Spaces

The vector representations of the questions and answers are examined using pairwise QA measures, as well as contextualized metrics devised to put the results into perspective within the corpus.

Similarity Building upon traditions in the encoder-based literature, cosine similarity is used as a measure of similarity within the embedding spaces. The similarity scores provide one-to-one comparison between the questions and answers.

Specificity Although similarity scores quantify the amount of resemblance between vector pairs, they do not inform on any potential adaptation within the embedding space. Thus, in an attempt to measure such effects, a specificity metric is devised to contrast pairwise similarity scores with an averaged behavior. The specificity score, defined in Equation 1, quantifies the degree to which an answer a_i ($i = 1, \cdots, N$) is tailored to its corresponding pair q_i, relative to a set of questions $\{q_j\}_{j=1}^{N}$. Conceptually, it addresses the question: *How much better does the answer match its associated question compared to all other questions in the corpus?*

Inspired by contrastive representation learning objectives, particularly InfoNCE (Sohn, 2016), this metric compares the similarity between each answer and its paired question against its similarity to non-target questions in a shared representation space. It is computed as:

$$\mathrm{Spec}(a_i) = \log\left(\frac{e^{m(a_i,q_i)}}{\frac{1}{N-1}\sum_{j\neq i} e^{m(a_i,q_j)}}\right) \quad (1)$$

where m is a similarity function between answer-question pairs. While omitted for simplicity, m usually operates on vector representations derived from a text encoder E, so that $m(a_i, q_j)$ denotes $m(E(a_i), E(q_j))$.

This self-contained, corpus-relative metric enables interpretable evaluation of answer specificity. Positive scores mean that answers are more similar to their target question than to others, indicating more tailored responses, whereas negative scores suggest that the answers align more closely with unrelated questions, revealing more generic or non-specific replies. The magnitude of the score further quantifies the degree of this effect.

In the following, the specificity scores are calculated with encoder models introduced in subsubsection 2.3.1 and m is the cosine similarity. Plus, the specificities are computed within subcorpora, i.e., between questions and answers originating from the same source.

3 Results & Discussion

Figure 1 showcases the results of the analysis. The cosine similarities, displayed as bars, and specificity scores, displayed as red triangles, reveal contrasting linguistic alignment strategies between hu-

man and LLM communication. At the same time, various LLMs (from different model families and sizes) that we tested, do not seem to differ from each other in their alignment strategies.

3.1 Semantic

Figure 1's top row shows that, for every source, LLMs produce responses that are semantically more similar to the questions than humans do. To delve further into this result, we analyzed the average proportion of words in the responses that share the same lemma as words in the questions (see Appendix B). This analysis shows that the main observation highly correlates with, and may stem from, the fact that LLMs re-use larger proportions of words (or lemmas) from the questions.

Likewise, LLMs' responses are more specific than humans' ones within the semantic representation space. Nonetheless, as expected, the one-sample Wilcoxon signed rank tests reveal that all specificity scores' median lie significantly above zero for humans and tested LLMs in all contexts. In terms of QA pairs as information exchange, this is desirable as it signals higher capacity to tackle questions' content, stay on the point, and to provide answers more tailored towards the topic of inquiry.

3.2 Stylistic

Conversely, computations within the stylistic representation space, as shown in Figure 1's bottom row, reveal an opposite pattern.

It is important to note that we interpret cosine proximity in the `styledistance` space solely in reference to the training methodology and results as described in Patel et al. (2025), i.e. we read it as an overlap across the 40-feature style palette that the model was explicitly trained to tease apart. Texts that align on most or all 40 coordinates cluster tightly, whereas divergence on even a handful of features (e.g., the answer is informal and emoji-rich while the question is formal and emoji-free) pushes their vectors apart. In the authors' benchmarks, sentence pairs written in the same style consistently achieve cosine similarities around 0.70 or higher, whereas stylistically divergent paraphrases fall markedly lower. Guided by these results, we interpret higher cosine scores in our QA pairs as evidence that the answer mirrors the stylistic choices of the question across the 40-feature palette (e.g., formality, emoji usage). All similarity statistics are analysed comparatively rather than against a fixed threshold.

For subcorpora where long enough question lengths, and hence sufficient style information, is available (i.e., within *medicine* and *reddit_eli5* that have median questions length of 71 and 43 words), human-written answers exhibit high stylistic similarity to the respective questions. LLM-generated answers show relatively low stylistic similarities across the board. In the subcorpora with short question lengths and hence insufficient linguistic style information, (in *finance* and *open_qa*, which have median question lengths of 12 and 6 words respectively), the higher margin in style alignment for humans disappears.

Moreover, the stylistic specificity computations expose statistical evidences of adaptation for human-written answers in all splits except for *open_qa*, whereas none of the tested LLMs exhibits such dynamic in any of the contexts. LLMs even disclose negative values in all but *reddit_eli5* subcorpus, meaning that the style of LLM-generated responses is generic and does not accommodate to the prompts' styles. This phenomenon may, at least in part, be explained by their training objectives. LLMs are primarily next-token predictors, trained to generate coherent text through extensive exposure to large corpora. This process may lead to a smoothed distribution over the vocabulary (Diehl Martinez et al., 2024; Guo et al., 2024), resulting in more generic discourse styles. Nevertheless, it is important to note that even for humans the specificity scores for style similarities are quite low.

To summarize, the results show that LLMs maintain high semantic alignment with the questions with which they were prompted —generating responses that are straightforward, factual and strictly-framed—, but do so in a generic style that often falls short of human standards in terms of similarity and doesn't exhibit specific adaptation to the questions. Humans, on the other hand, show lower semantic alignment to the questions asked. A close reading of a subsample of responses suggests that this is because of a greater variety in response conceptualization, for instance, use of metaphors, stories and deviations. At the same time, humans show relatively higher stylistic alignment.

4 Conclusion

This study contributes to a growing body of work examining how language models compare to humans in conversational settings. It analyzes ques-

tion answering approaches in LLMs and contrasts them with human behavior by introducing H-LLMC2, an augmented version of a sample of HC3, a corpus of human-written questions and parallel sets of respective answers written by humans and various LLMs. The similarity and specificity of responses are assessed within representation spaces encoding semantics and style. While no major distinctions are observed between various LLMs (of different sizes and model-families), the findings reveal interesting contrasts between LLM-generated and human-written answers. LLMs tend to exhibit higher semantic similarity with high specificity, but don't reach human-level stylistic similarity. Human-written answers exhibit a reverse pattern, with lower semantic but higher stylistic similarity to the respective questions (when questions lengths are long enough). These findings point to contrasting linguistic accommodation strategies evident in human and LLM communication. These results invite further exploration into the conversational dynamics of LLMs, particularly with regard to how they differ from human behavior, and prompt reflection on strategies that could foster more natural communicative practices. Inference from the extensive literature on linguistic accommodation suggests that bringing stylistic and semantic alignment strategies in LLM responses closer to human behavior can enhance social attunement and engagement in human-AI dialogue.

Limitations

This study heavily relies on representation spaces and notions of similarities within such high-dimensional spaces. The notions of semantics and style encoded by the models are learned and may lack transparency and clearer interpretability. Moreover, while largely adopted in the literature and often preferred to other scores, the soundness of using cosine similarity has recently been discussed (Steck et al., 2024). Future work would benefit from incorporating human judgments of perceived alignment and answer quality as a complement or counterpoint to the computational methods. This would help validate or challenge the results, and contribute in bringing valuable context from a human-centered or social perspective.

Furthermore, while H-LLMC2 offers a clear framework to compare in parallel human-written answers and LLM-generated responses to the same questions, the size of the questions, especially within *finance* and *open_qa* subsets, may impair embedding robustness, thus harming interpretation possibilities. Alternatively, the use of deeper categorization based on intent or question types, rather than questions' source, could offer finer granularity to interpret the results and potentially provide new insights into the data.

Acknowledgments

Parts of this research were carried within the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 945304 - Cofund AI4theSciences hosted by PSL University. SD's contribution was supported by a grant from IEAI, TUM.

References

Milena Belosevic and Hendrik Buschmeier. 2024. Calibrating trust and enhancing user agency in llm-based chatbots through conversational styles. *CUI@ CHI 2024: Building Trust in CUIs–From Design to Deployment*.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56, Bangkok, Thailand. Association for Computational Linguistics.

Tanya L. Chartrand and John A. Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910. Place: US Publisher: American Psychological Association.

Tanya L. Chartrand and Rick van Baaren. 2009. Chapter 5 Human Mimicry. In *Advances in Experimental Social Psychology*, volume 41, pages 219–274. Academic Press.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Richard Diehl Martinez, Zébulon Goriely, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. Mitigating frequency bias and anisotropy in language model pre-training with syntactic smoothing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5999–6011, Miami, Florida, USA. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of*

the 57th Annual Meeting of the Association for Computational Linguistics, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Simon Garrod and Martin J. Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1):8–11. Publisher: Elsevier.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In Howard Giles, Justine Coupland, and Nikolas Coupland, editors, *Contexts of Accommodation: Developments in Applied Sociolinguistics*, Studies in Emotion and Social Interaction, pages 1–68. Cambridge University Press, Cambridge.

Howard Giles and Tania Ogay. 2007. Communication Accommodation Theory.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1):18617.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10).

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, 21(4):337–360. Publisher: SAGE Publications Inc.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. Styledistance: Stronger content-independent style embeddings with synthetic parallel examples. *Preprint*, arXiv:2410.12757.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 887–890, New York, NY, USA. Association for Computing Machinery.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In

Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, pages 1–20, New York, NY, USA. Association for Computing Machinery.

A H-LLMC2 Details

A.1 Technical Details

The responses generated with open-weights LLMs to assemble H-LLMC2 was performed on a single machine equipped with NVIDIA GeForce RTX 3090 Ti GPU (24GB), and relied on the `transformers`[4] library, except for the data generated for DeepSeek-R1-8B, which was generated on M2-powered MacBook Pro and relied on `ollama`[5] implementation. The models used for generation were quantized, loaded in 8 bits, or converted to 4-bits for ollama (`Q4_K_M` version). The version of ChatGPT used to create HC3 is not mentionned explicitly, but is prior to January 2023 (date of publication on HuggingFace).

A.2 Summary statistics

Table 1 provides an overview of the word count distribution per split within H-LLMC2, both for question and answers from human and LLMs. The number of words is approximated using whitespace splitting.

B Textual Analysis

Figure 2 shows the average proportion of words in the responses that share the same lemma[6] as words in the questions. Higher values mean that responses tend to be closer semantically to the questions. As can be observed on Figure 2, the shared proportion is consistently higher for LLMs, compared to human-written answers. Across the whole corpus, the median among LLM is 60%, compared to 38% for humans.

		Percentiles				
		0	25	50	75	100
medicine	question	60	65	71	79	97
	human	10	53	74	102	336
	chatgpt	76	155	185	216	407
	L2-7B-Chat	162	301	319	333	404
	L2-13B-Chat	201	304	323	337	422
	L3.1-8B-Inst	14	27	42	327	411
	Q2.5-7B-Inst	4	303	343	393	1596
	Q2.5-14B-Inst	1	218	278	316	578
	DS-R1-8B	59	191	225	255	556
	DS-R1-14B	26	210	244	276	465
reddit_eli5	question	31	37	43	48	54
	human	8	56	100	189	1707
	chatgpt	52	141	176	210	639
	L2-7B-Chat	88	247	290	327	421
	L2-13B-Chat	9	257	300	339	416
	L3.1-8B-Inst	4	286	343	392	438
	Q2.5-7B-Inst	5	216	262	313	688
	Q2.5-14B-Inst	1	175	211	247	509
	DS-R1-8B	10	119	166	203	402
	DS-R1-14B	50	135	185	225	512
finance	question	8	10	12	15	31
	human	2	75	138	226	1629
	chatgpt	22	173	208	247	427
	L2-7B-Chat	42	310	345	367	423
	L2-13B-Chat	79	321	348	368	417
	L3.1-8B-Inst	4	329	370	392	441
	Q2.5-7B-Inst	2	288	363	434	1487
	Q2.5-14B-Inst	1	240	307	363	572
	DS-R1-8B	14	198	243	286	756
	DS-R1-14B	13	214	261	305	488
open_qa	question	5	5	6	8	17
	human	2	20	27	38	206
	chatgpt	5	75	104	152	595
	L2-7B-Chat	7	131	249	322	426
	L2-13B-Chat	4	115	236	317	415
	L3.1-8B-Inst	6	145	245	340	454
	Q2.5-7B-Inst	8	59	122	231	1518
	Q2.5-14B-Inst	1	44	80	163	473
	DS-R1-8B	7	63	111	215	966
	DS-R1-14B	6	54	92	192	466

Table 1: Word count percentiles per data source and input/output types ('L' stands for Llama, 'Q' for Qwen and 'DS' for DeepSeek).

[4] https://github.com/huggingface/transformers

[5] `https://ollama.com`

[6] obtained with spaCy library (`en_core_web_sm` version)

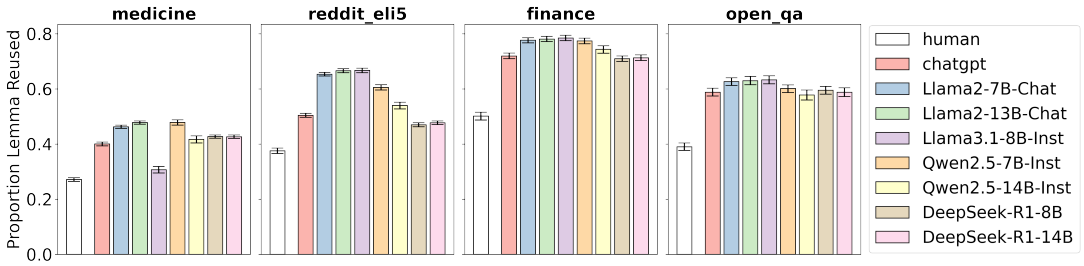

Figure 2: Mean proportion of lemma reused in answer provided by humans and LLMs with 95% bootstrapped confidence intervals, grouped by data sources.