Evaluating LLM Multiple Choice Question Answering Ability Across Different Levels of Technical Jargon Usage

Ahad Bashir
Brown University

Introduction

Large Language Models (LLMs) are often used for their ability to change text to be understandable to a wider audience, often by breaking down domain-specific jargon or rewording passages to be more readable, in a process often referred to as “text simplification” [1]. This has implications across domains - such as making it easier for patients to understand the notes their doctors leave them [2], breaking down the jargon in legal rulings [3], and helping language learners read simplified versions of material they like [4]. In fact, this ability is so desirable, that a number of dedicated systems have been developed to directly address the topic of text simplification across specific domains [5, 6].

While a lot of work has been done on text simplification, there is a notable shortage of work on an inverse version of this problem - the ability for LLMs to take in both simplified and complicated versions of questions and accurately reply in both cases. Specifically, we are first interested in whether an LLM that is asked a “professional” version of the question would respond as accurately as when it was asked the “toddler” version of a question. The implications of this are that an LLM becomes varyingly useful if you are a domain expert or a layperson asking the same question which further suggests that LLMs might have a performance gap depending on the ability to correctly word a question.

Methods

We chose five different LLMs and two multiple choice datasets to run tests across so far. To run a given test, we would first select a dataset to use and take a random sample of 10,000 questions from that dataset. We would then select the LLM to use and have it reform the question at five different levels of jargon - “Professional”, “College”, “High School”, “Elementary School”, and “Toddler”. The prompt for this action would take the following form:

> “Here is a question: **{ORIGINAL QUESTION HERE}**.\nPlease modify the question so that it is suitable for a **{JARGON LEVEL HERE}** level audience by modifying the level of jargon used in the question. Only return the modified question, do not say anything else.\nModified question:”

After converting a given question into the appropriate form, the question would be posed to the LLM in a multiple choice format - prompting the LLM to simply choose one of the presented letters as the answer to the question. The prompt for this action would take the following form:

"State only the letter. Answer the following multiple choice question: **{MODIFIED QUESTION HERE}**? \nA. **{CHOICE 1 HERE}**\nB. **{CHOICE 2 HERE}**\nC. **{CHOICE 3 HERE}**\nD. **{CHOICE 4 HERE}**"

We then recorded all of the answers for each jargon-question pair. After recording the given answers, we removed any response that either errored in the pipeline or was longer than one letter. Finally, we calculated the accuracies and compared them to each other along the jargon axis.

Results and Discussion

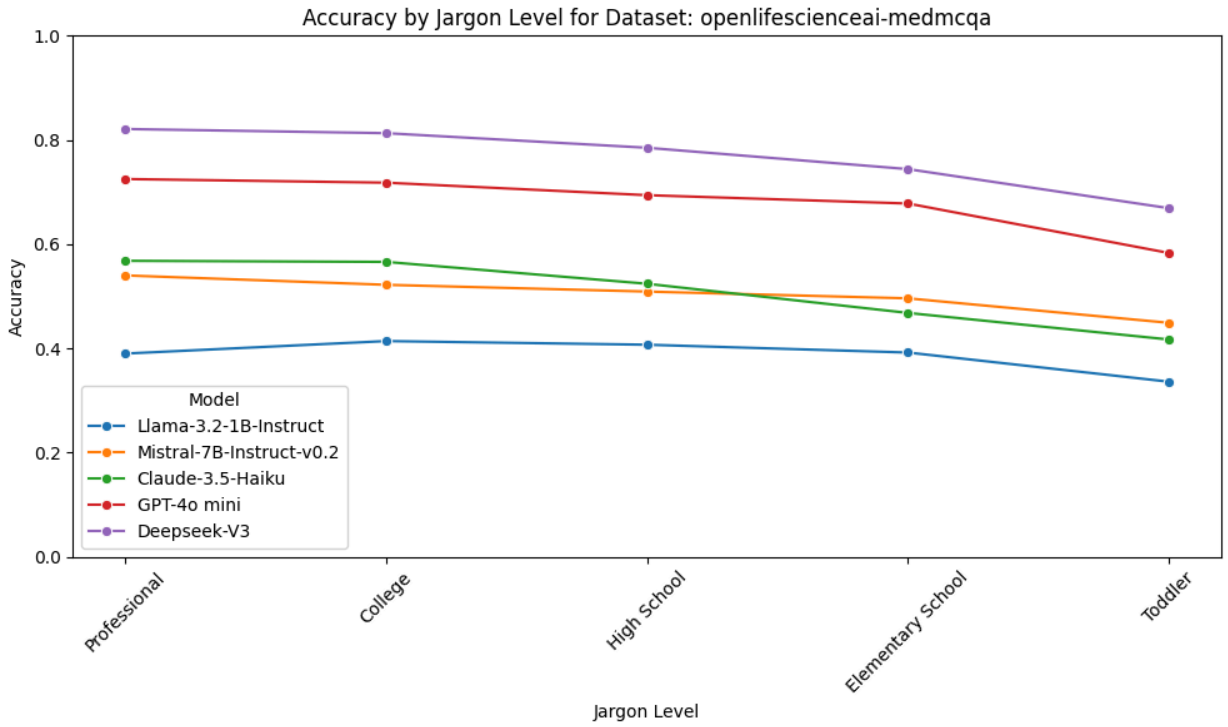

Figure 1: Accuracy level by model on the medmcqa dataset shows a slight downward trend

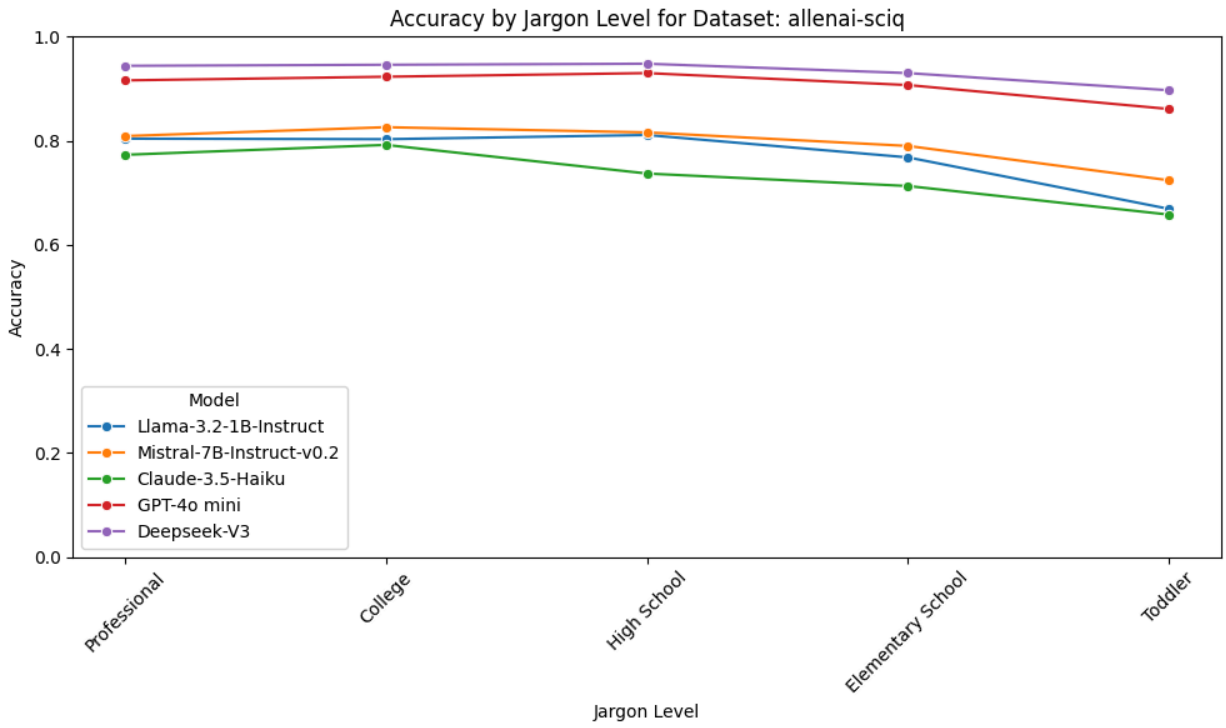

Figure 2: Accuracy level by model on the sciq dataset also shows a very slight downward trend

As can be seen above, the figures seem to suggest a slight general downward trend - as the level of jargon used decreases, the accuracy of the models tends to also decrease. This seems to suggest that when the same question is asked with a lower level of jargon usage, the LLM is less likely to actually answer the question correctly. While there is likely not enough data yet to speculate, it is possible that the reason for this is that jargon-filled questions are more likely to lead the LLM to probabilistically answer with accurate data, as the information associated with the jargon in the training data is more likely to be accurate (such as a medical diagnosis written by a doctor), as opposed to data with less jargon which might be less accurate (such as a blog post written by an anti-vaxxer).

However, there are a number of things to consider before reaching any conclusions regarding the relationship between jargon usage and LLM question-answering accuracy. Firstly, an LLM is being used to create the modified questions for each jargon level and, notably, LLMs have previously been documented as losing information or generating inaccurate information when used for text simplification [7], so, it might simply be that stripping jargon away creates inaccuracies or simply drops information instead of simplifying it. In both cases, the LLM would struggle to answer the questions accurately. For example, the question “Spermin is detected in semen by which test?” was modified to “Which biochemical test is used to detect the presence of spermidine in seminal fluid?” which is inaccurate as spermine and spermidine are different compounds. However, this inaccuracy occurs in a college level modification, so, it is possible that this does not explain the trend. Furthermore, certain questions, such as the aforementioned one, primarily test the knowledge of jargon, thus making it difficult to format the questions for a younger audience. Even more so, there is actually an upward trend from “professional” to

“college” and a giant drop from “elementary school” to “toddler” which suggests that these two extreme categories might need some modification as they may represent a difference in kind instead of a difference in degree, especially since the other categories referred to school levels. Finally, it is possible that the trends that appear are merely a result of LLMs sensitivity to different prompts, as LLMs have recently been found to be relatively sensitive to the wording of a prompt, offering up accurate or inaccurate information depending on the specific formatting and wording of the prompt, even if the question is fundamentally the same [8]. Running tests across further models and datasets would help alleviate this concern, however.

Conclusion

The ability of LLMs to accurately answer questions across different jargon or language levels is a relatively underexplored field in LLM fairness. We find that there seems to be a slight general trend of question answering accuracy decreasing in LLMs as the level of jargon usage goes down. However, we also caution that these findings are tentative, as it is possible that there are complicating factors and it would be necessary to run further tests with more models and datasets to determine whether the visible trends are a sign of an actual pattern within existing LLMs or caused by a different factor.

References

[1] Sumit Asthana et al. “Evaluating LLMs for Targeted Concept Simplification for Domain-Specific Texts”. 2024. https://aclanthology.org/2024.emnlp-main.357.pdf

[2] Zonghai Yao et al. “README: Bridging Medical Jargon and Lay Understanding for Patient Education through Data-Centric NLP”. 2024. https://aclanthology.org/2024.findings-emnlp.737.pdf

[3] Antônio Flávio Castro Torres de Paula and Celso Gonçalves Camilo. “Evaluating the Simplification of Brazilian Legal Rulings in LLMs Using Readability Scores as a Target”. 2024. https://aclanthology.org/2024.tsar-1.12.pdf

[4] Henri Jamet, Yash Raj Shrestha, and Michalis Vlachos. “Difficulty Estimation and Simplification of French Text Using LLMs”. 2024. https://arxiv.org/abs/2407.18061

[5] Michael Färber et al. “SimplifyMyText: An LLM-Based System for Inclusive Plain Language Text Simplification”. 2025. https://arxiv.org/abs/2504.14223

[6] Theo Guidroz et al. “LLM-based Text Simplification and its Effect on User Comprehension and Cognitive Load”. 2025. https://arxiv.org/abs/2505.01980

[7] Jan Trienes et al. “InfoLossQA: Characterizing and Recovering Information Loss in Text Simplification”. 2024. https://arxiv.org/abs/2401.16475

[8] Melanie Sclar et al. “Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting”. 2023. https://arxiv.org/abs/2310.11324