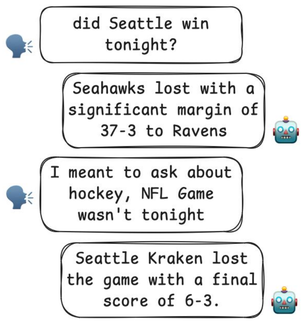
Disambiguating Human Language in LLMs and Analyzing its Effect to Improve NLP Task Accuracy

Aryan Vinod Keluskar, Computer Science
Mentor(s): Dr. Huan Liu, Regents Professor and Amrita Bhattacharjee, PhD Student
School of Computing and Augmented Intelligence

Abstract

Ambiguity in natural language poses significant challenges to Large Language Models (LLMs) used for open-domain question answering. LLMs often struggle with the **inherent uncertainties** of human communication, leading to misinterpretations, miscommunications, **hallucinations**, and **biased** responses. This significantly weakens their ability to be used for tasks like fact-checking, question answering, feature extraction, and sentiment analysis. Using **open-domain question answering** as a test case, we compare off-the-shelf and few-shot LLM performance, focusing on measuring the impact of explicit disambiguation strategies.

Problem Statement



Type	Example
Event references (39%)	What season does meredith and derek get married in grey's anatomy? Q: In what season do Meredith and Derek get informally married in Grey's Anatomy? Q: In what season do Meredith and Derek get legally married in Grey's Anatomy?
Properties (27%)	How many episode in seven deadly sins season 2? Q: How many episodes were there in season 2, not including the OVA episode? / A: 25 Q: How many episodes were there in season 2, including the OVA episode? / A: 26
Entity references (23%)	How many sacks does clay matthews have in his career? Q: How many sacks does Clay Matthews Jr. have in his career? / A: 69.5 Q: How many sacks does Clay Matthews III have in his career? / A: 91.5
Answer types (16%)	Who sings the song what a beautiful name it is? Q: Which group sings the song what a beautiful name it is? / A: Hillsong Live Q: Who is the lead singer of the song what a beautiful name it is? / A: Brooke Ligertwood
Time-dependency (13%)	When does the new family guy season come out? Q: When does family guy season 16 come out? / A: October 1, 2017 Q: When does family guy season 15 come out? / A: September 25, 2016 Q: When does family guy season 14 come out? / A: September 27, 2015

Recent work [1] has demonstrated LLMs struggle to understand ambiguous text, because fail to 'understand' the context resulting in improper responses, or hallucinating factually wrong responses with high confidence [2]. To investigate its effect, this work answers:

1. How do off-the-shelf LLMs perform on ambiguous questions with zero-shot?
2. Can fine-tuning or training-based disambiguation improve performance?

References

[1] A. Tamkin, K. Handa, A. Shrestha, and N. Goodman, "Task ambiguity in humans and language models," arXiv preprint arXiv:2212.10711, 2022.

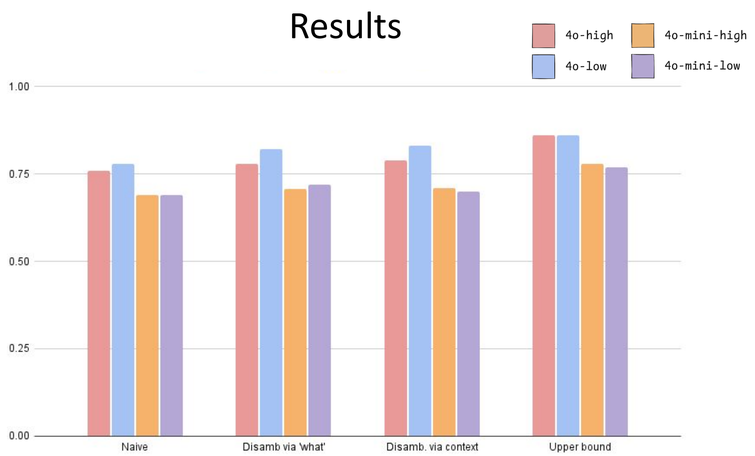
[2] Y. Zhang et al., "Siren's song in the AI ocean: A survey on hallucination in large language models," 2023. Available: https://arxiv.org/abs/2309.01219

[3] A. Salinas and F. Morstatter, "The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance," arXiv preprint arXiv:2401.03729, 2024.

Question Disambiguation Model

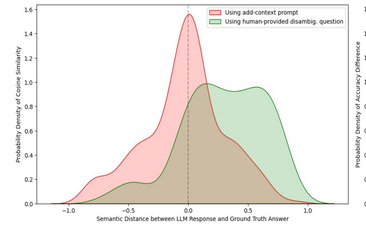
We used two prompt-level disambiguation methods - (1) Rephrasing a question using 'what', and (2) Add Context to the Ambiguous Question. Since LLMs have vast amounts of knowledge due to extensive pre-training and instruction tuning, we use that to find and return relevant information about the ambiguous question. We then append the original ambiguous question at the end of this information blob and pass it back to the LLM to obtain the final answer. We use the publicly available AmbigQA Dataset with 14,042 ambiguous questions to experiment on LLM performance using our methods.

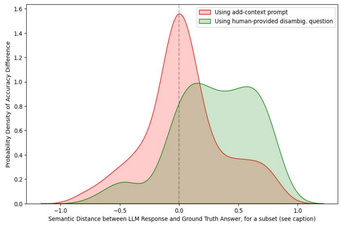
Results



Comparison of Cosine Similarity between the Ground Truth Answer and Model Generated Output for GPT 4o and 4o-mini, tested under both high and low temperatures.

What does this tell us?





Using simple disambiguating prompts improves performance over the default setting, implying that simple prompt-based, training-free approaches are useful in improving LLM performance for ambiguous queries. Out of the two disambiguating methods explored, we see that disambiguating by adding context performs better for large, as well as small sized LLMs.

Additionally, our results imply that lowering the temperature of a model may not provide any benefits in LLM performance for answering ambiguous questions.

Therefore, Contextual enrichment has promising ability to enhance disambiguation accuracy, but it is often inaccurate because the LLM adds or hallucinates irrelevant context *(as seen in plot 1, which is skewed left)*. However, the accuracy of the LLM increases while running it on a subset of human-provided disambiguated questions *(since plot 2 is skewed right)*. This shows that LLMs are able to better understand certain social cues to correctly disambiguate the provided question in cases where the human annotators were able to disambiguate them as well.

Future Work

In future work, we plan to fine-tune the LLM for accurate context-enhancement. Specifically, we will take the contextually enriched information blob and fine-tune the model to generate a disambiguated question that is as close as possible to human-provided disambiguated question to maximize accuracy for question-disambiguation based strategies, which will improve the accuracy for questions outside of AmbigQA dataset

FURI

Ira A. Fulton Schools of Engineering
Arizona State University