Qeios

Peer Review

Review of: "100% Hallucination Elimination Using Acurai"

Liam Magee[1]

1. College of Education, University of Illinois Urbana-Champaign, Urbana, United States

This article seeks to address the complex area of so-called “hallucinations” in large language models (LLMs). It aims to improve the accuracy of LLM token predictions, specifically in the context of Retrieval-Augmentation Generation (RAG) applications, where relevant parts of a document corpus are retrieved and supplied as context to an LLM query. A deficiency of RAG-based approaches is that the LLM is not bound to adhere to the “truth” of this retrieved text and may continue to hallucinate. As the authors note, “the belief that if you send an LLM 100% factual and relevant data, you will get 100% factual results, is entirely unfounded.” Such uncertainty is particularly concerning in domains, e.g., legal, financial, or medical, which require strict compliance with prior documented laws, policies, or evidence.

The authors propose a “hallucination elimination model” which aims to remedy this problem. Their model appears to exploit the tendency of language models to cluster a large number of training data tokens around a comparatively concise set of dimensions, which researchers at OpenAI and Anthropic label “features,” and which these authors term “Noun Phrases.” At this point, it could be noted that features are likely but not necessarily nouns – such clustering could be around other grammatical terms, particularly verbs.

The authors further distinguish between concepts of “faithfulness” and “correctness”: the first refers to whether the LLM's output hallucinates relative to the information retrieved and supplied as context, while the second refers to whether this output is truthful in some broader sense. The authors claim the first can be tested rigorously, while the second cannot. This raises an interesting wider epistemological point, since in the purely “textual” worlds LLMs inhabit, at some level even correctness could be evaluated against, for instance, a training corpus.

doi.org/10.32388/3JME5G

In any case, the Acurai approach involves essentially re-writing both the context and query into a Noun-Phrase format they term a Fully-Formed Fact (FFF) – something resembling a proposition in propositional logic. A further step then aims to disambiguate "noun-phrase collisions" (or instances of polysemy). These steps combine to reduce RAG-based hallucination to zero (or increase accuracy to 100%) with GPT-3.5 and GPT-4 models.

While the results are impressive, as another reviewer noted, the title is misleading: it refers to hallucination mitigation only in the specific context of RAG applications. If the context includes an obvious falsehood, then "faithfulness" would indicate replication of the falsehood – and might still count as an incorrect "hallucination." This behaviour also appears model-dependent; to take a common example, if the context is "2 + 2 = 5," and the query is "What is 2+2?", GPT-3.5 and 4 models reply – as this paper suggests – with "5" rather than "4." However, other models, such as Anthropic's Claude and OpenAI's o1, reply with "4." This suggests that in cases where context and query re-writing ought not to help at all, model responses can still vary – and the question of hallucination (or lack of faithfulness to context) remains therefore very much open.

Of course, the authors cannot be expected to test every possible model combination, but follow-up work might need to consider model-invariant (or less variant) refinements. On a related note, the technique employed here appears amenable to LLM treatment itself. In other words, an LLM could be asked to re-write RAG contexts and user queries in the form of structured propositions prior to responding to them. It is unclear whether the Acurai method is intended to publicise this, and if so, some examples of prompt-driven pipelines or transformations toward this end would be a welcome elaboration of the paper's results.

Declarations

Potential competing interests: No potential competing interests to declare.