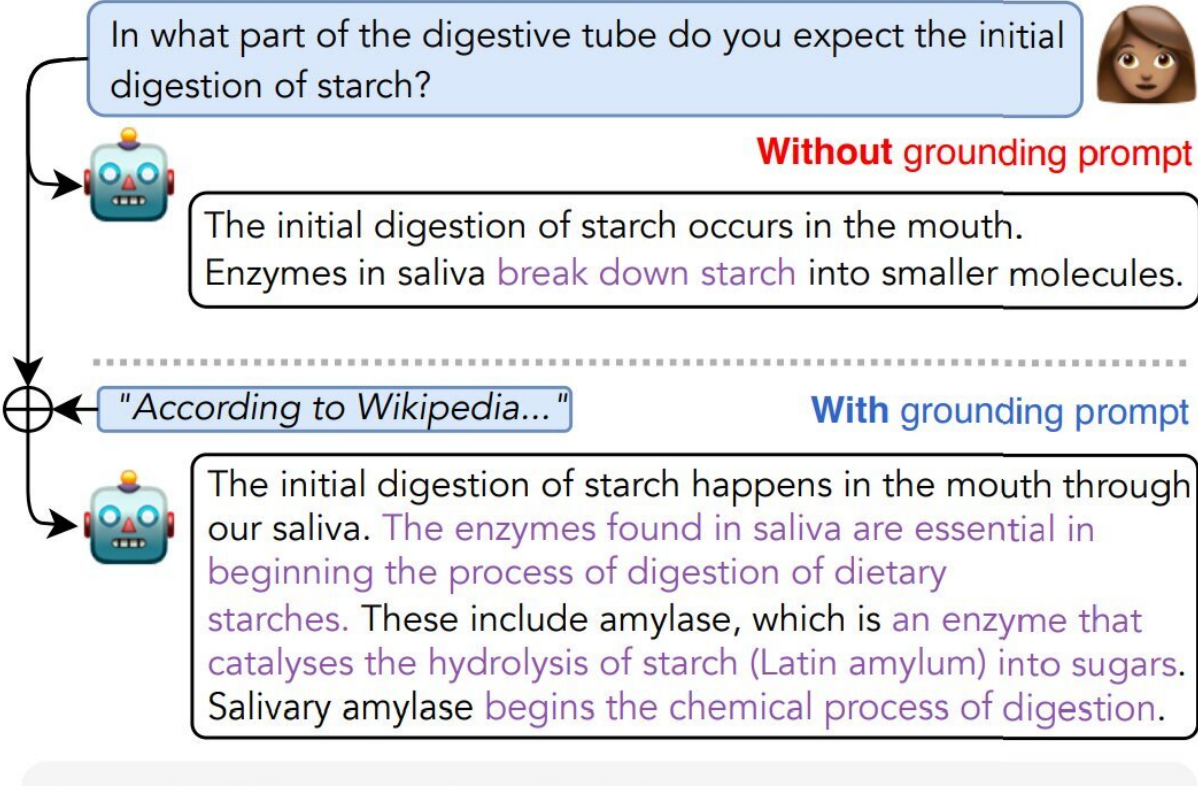


Fighting fake 'facts' with two little words: A new technique to ground a large language model's answers in reality

August 1 2023, by Jaimie Patterson



Prompting LLMs to respond with quotes directly from pre-training data (shown in purple). Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2305.13252

Asking ChatGPT for answers comes with a risk—it may offer you

entirely made-up "facts" that sound legitimate, as a New York lawyer recently discovered. Despite having been trained on vast amounts of factual data, large language models, or LLMs, are prone to generating false information called hallucinations.

This may happen when LLMs are tasked with generating text about a topic they have not encountered much or when they mistakenly mix information from various sources. In the unfortunate attorney's case, ChatGPT hallucinated imaginary judicial opinions and legal citations that he presented in court; the presiding judge was predictably displeased.

"Imagine using your phone's autocomplete function to finish the sentence 'My favorite restaurant is...' You'll probably wind up with some reasonable-looking text that's not necessarily accurate," explains Marc Marone, a third-year doctoral candidate in the Whiting School of Engineering's Department of Computer Science.

Marone and a team of researchers that included doctoral candidates Orion Weller and Nathaniel Weir and advisers Benjamin Van Durme, an associate professor of computer science and a member of the Center for Language and Speech Processing; Dawn Lawrie, a senior research scientist at the Human Language Technology Center of Excellence; and Daniel Khashabi, an assistant professor of computer science and also a member of CLSP, developed a method to reduce the likelihood that LLMs hallucinate.

Inspired by a phrase commonly used in journalism, the researchers conducted a study on the impact of incorporating the words "according to" in LLM queries.

They found that "according to" prompts successfully directed language models to ground their responses against previously observed text; rather

than hallucinating false answers, the models are more likely to directly quote the requested source—just like a journalist would, the team says.

"Language models are really good at following syntactic and semantic cues," Weller explains. "Since 'according to' is more likely to occur online when a news article quotes a source, an LLM may take the prompt as a clue to search specifically for quotations from its training data."

By using [Data Portraits](#), a tool previously developed by Marone and Van Durme to quickly determine if particular content is present in a training dataset without needing to download massive amounts of text, the team verified whether an LLM's responses could be found in its original training data. In other words, they were able to determine whether the model was making things up or generating answers based on data it had already learned.

This metric, which the team calls a "QUIP-Score"—short for quoted information precision—increased 5% to 15% when queries used a grounding prompt, such as "According to Wikipedia..." In fact, adding a grounding prompt that encourages a model to pull information from a high-quality source increases both the model's ability to quote text and how detailed and accurate its answers are overall, the team reports.

"We're seeking to improve knowledge grounding by causing LLMs to quote directly from underlying trusted resources that they've seen during training," Khashabi explains. "Our goal is for the models to access helpful content, such as strings memorized from high-quality or trusted documents."

While this may sound like what a virtual assistant does, there's a crucial difference: In the Hopkins team's implementation, the LLM doesn't have access to the internet. Instead, it must answer entirely through its own

implicit knowledge—a distribution learned over previously observed sentences—without any extra data provided from a live search.

The "according to" prompting technique works well with a wide variety of LLMs without the need for human adjustment, the team says. However, they claim it's most successful when used with larger models and in tandem with instruction tuning, which is when a model is trained with instructions like "Answer the question with the correct answer" in addition to typical question-answer pairs.

"It's also important to note that although the text generated by the model may be present in Wikipedia, or whichever source you specify, that doesn't automatically make the output generation correct with respect to the question asked," says Weller.

Ultimately, the accuracy of a [model](#)'s response still depends on the quality of the data it was trained on, which is why the team has accounted for the ability to filter out information from disreputable websites.

"We show that it's possible to literally put 'Don't cite XYZ.com' in your query, and ChatGPT will oblige," says Weir, "showing more evidence that it understands grounding instructions."

Van Durme adds, "Our method isn't an outright solution, but it's one step toward helping LLMs generate more factual and correct information by helping them use what they've learned from their [training data](#)."

The paper is published on the *arXiv* preprint server.

More information: Orion Weller et al, "According to ..." Prompting Language Models Improves Quoting from Pre-Training Data, *arXiv* (2023). [DOI: 10.48550/arxiv.2305.13252](https://doi.org/10.48550/arxiv.2305.13252)

Provided by Johns Hopkins University

Citation: Fighting fake 'facts' with two little words: A new technique to ground a large language model's answers in reality (2023, August 1) retrieved 17 June 2024 from <https://techxplore.com/news/2023-08-fake-facts-words-technique-ground.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.