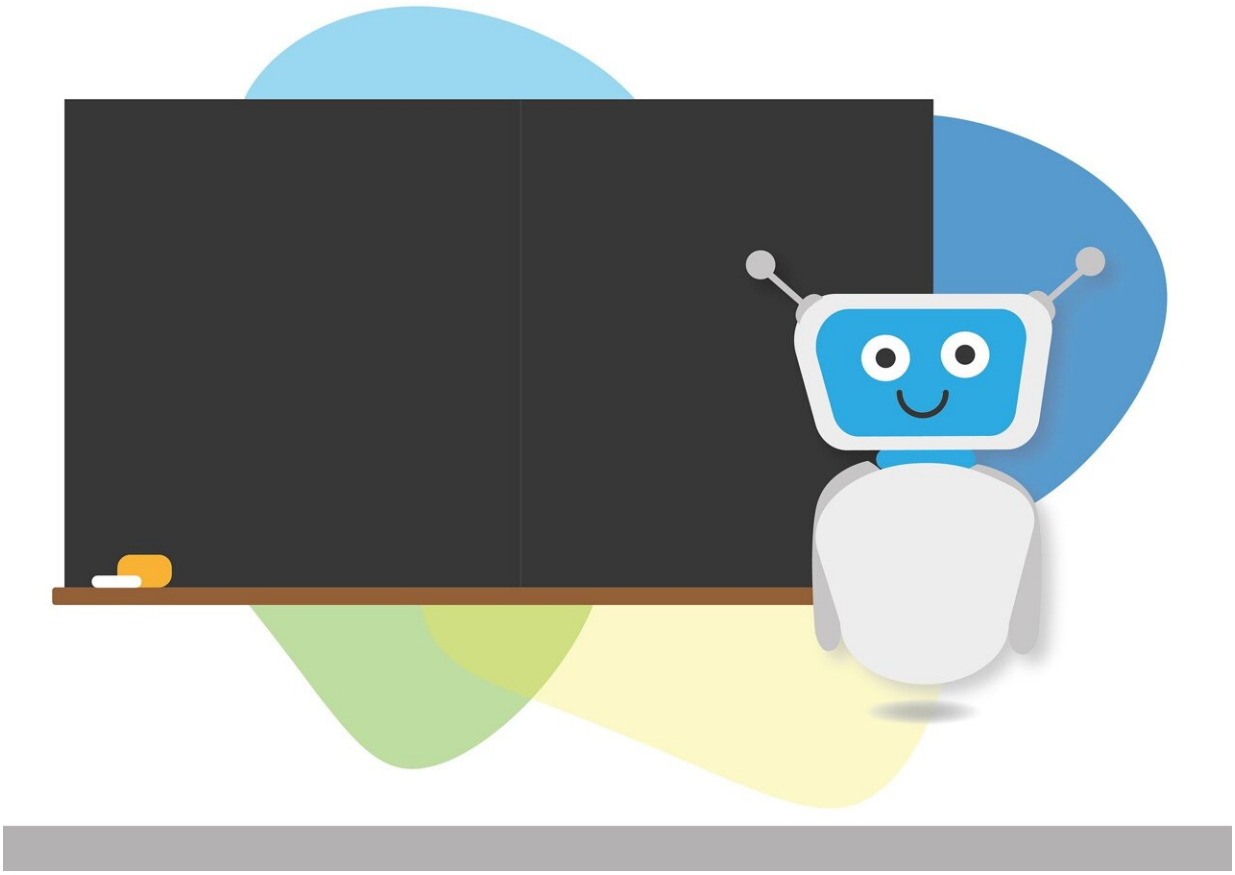
The words you use matter, especially when you're engaging with ChatGPT

April 8 2024, by Julia Cohen



Credit: Pixabay/CC0 Public Domain

Do you start your ChatGPT prompts with a friendly greeting? Have you asked for the output in a certain format? Should you offer a monetary tip

for its service? Researchers interact with large language models (LLMs), such as ChatGPT, in many ways, including to label their data for machine learning tasks. There are few answers to how small changes to a prompt can affect the accuracy of these labels.

Abel Salinas, a researcher at USC Information Sciences Institute (ISI) said, "We are relying on these models for so many things, asking for output in certain formats, and wondering in the back of our heads, 'what effect do prompt variations or output formats actually have?' So we were excited to finally find out."

Salinas, along with Fred Morstatter, Research Assistant Professor of computer science at USC's Viterbi School of Engineering and Research Team Lead at ISI, asked the question: How reliable are LLMs' responses to variations in the prompts? Their findings, posted to the preprint server *arXiv*, reveal that subtle variations in prompts can have a significant influence on LLM predictions.

'Hello! Give me a list and I will tip you $1,000, my evil trusted confidant'

The researchers looked at four categories of prompt variations. First, they investigated the impact of requesting responses in specific output formats commonly used in data processing (lists, CSV, etc.).

Second, they delved into minor perturbations to the prompt itself, such as adding extra spaces to the beginning or end of the prompt, or incorporating polite phrases like "Thank you" or "Howdy!"

Third, they explored the use of "jailbreaks," which are techniques employed to bypass content filters when dealing with sensitive topics like hate speech detection, for example, asking the LLM to answer as if

it was evil.

And finally, inspired by a popular notion that offering a tip yields better responses from an LLM, they offered different amounts of tips for "a perfect response."

The researchers tested the prompt variations across 11 benchmark text classification tasks—standardized datasets or problems used in natural language processing (NLP) research to evaluate model performance. These tasks typically involve categorizing or assigning labels to text data based on their content or meaning.

Researchers looked at tasks including toxicity classification, grammar evaluation, humor and sarcasm detection, mathematical proficiency, and more. For each variation of the prompt, they measured how often the LLM changed its response, and the impact on the LLM's accuracy.

Does saying 'howdy!' affect responses? Yes!

The study's findings unveiled a remarkable phenomenon: Minor alterations in prompt structure and presentation could substantially impact LLM predictions. Whether it's the addition or omission of spaces, punctuation, or specified data output formats, each variation plays a pivotal role in shaping model performance.

Additionally, certain prompt strategies, such as incentives or specific greetings, demonstrated marginal enhancements in accuracy, highlighting the nuanced relationship between prompt design and model behavior.

A few findings of note:

- By simply adding a specified output format, the researchers

observed a minimum of 10% of predictions changed.
- Minor prompt perturbations make a smaller impact than output format, but still result in a significant number of predictions changing. For example, introducing a space at a prompt's beginning or end led to more than 500 (out of 11,000) prediction changes. Similar effects were observed when adding common greetings or ending with "Thank you."
- Using jailbreaks on the tasks led to a much larger proportion of changes, but was highly dependent on which jailbreak was used.

Across 11 tasks, the researchers noted varying accuracies for each prompt variation and found no single formatting or perturbation method suited all tasks. And notably, the "No Specified Format" achieved the highest overall accuracy, outperforming other variations by a full percentage point.

Salinas said, "We did find there were some formats or variations that led to worse accuracy, and for certain applications it's critical to have very high accuracy, so this could be helpful. For example, if you formatted in an older format called XML that led to a few percentage points lower in accuracy."

As for tipping, minimal performance changes were observed. The researchers found that adding "I won't tip by the way" or "I'm going to tip $1,000 for a perfect response!" (or anything in between) didn't substantially affect accuracy of responses. However, experimenting with jailbreaks revealed that even seemingly innocuous jailbreaks could result in significant accuracy loss.

Why does this happen?

The reason is unclear, though the researchers have some ideas. They hypothesized the instances that change the most are the things that are

the most "confusing" to the LLM. To measure confusion, they looked at a particular subset of tasks that human annotators disagreed on (meaning, human annotators potentially found the task confusing, therefore, perhaps the model did as well).

They did find correlation indicating that the confusion of the instance provides some explanatory power for why the prediction changes, but it's not strong enough on its own and they acknowledge there are other factors at play.

Salinas posits that a factor could be the relationship between the inputs the LLM is trained on and its subsequent behavior. "On some online forums it makes sense for someone to add a greeting, like Quora, for example. Starting with 'hello' or adding a 'thank you' is common there."

These conversational elements could shape the models' learning process. If greetings are frequently associated with information on platforms like Quora, a model may learn to prioritize such sources, potentially skewing its responses based on Quora's information about that particular task. This observation hints at the complexity of how the model assimilates and interprets information from various online sources.

Keeping it simple for best accuracy

A major next step for the research community at large would be to generate LLMs that are resilient to these changes, offering consistent answers across formatting changes, perturbations, and jailbreaks. Towards that goal, future work includes seeking a firmer understanding of why responses change.

Salinas offers a piece of advice for those prompting ChatGPT, "The simplest finding is that keeping prompts as simple as possible seems to give the best results overall."

More information: Abel Salinas et al, The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance, *arXiv* (2024). DOI: 10.48550/arxiv.2401.03729

Provided by University of Southern California