Quality Control with LLMs

It's critical to remember that new advances in AI are not only scientific discoveries, they are products with revenue potential. Being able to conduct your own research and separate marketing claims from actual capabilities and applications is an important skill for successfully implementing AI technologies.

Be selective with AI solutions, and try to take a problem-solving approach that considers the past, present, and future of the problem. Remember that AI – and indeed no technology – is a silver bullet for a complicated problem.

Authority bias and LLMs

Outputs from LLMs can always contain factually inaccurate information, and LLMs do not know or understand facts. They perform language tasks with confidence and it's important for humans to be skeptical editors with an awareness of authority bias.

LLMs don't know what they don't know! Unless they've been specifically trained to say "I don't know the answer to this" or "This isn't something I can answer," an LLM will present the best answer it can with no indication that it lacks data.

Security Vulnerability in LLMs

Without proper security measures in place, LLMs are vulnerable to attacks like prompt injections. Additionally, motivated users are generally able to find workarounds to content limits.

Additionally, LLMs are purportedly trained only on publicly-available data. Private, proprietary, or personally-identifying information should always be kept out of an LLM, just as it would be protected when working with any 3rd party source.

Generalization & Compression in LLMs

Generalization is both a core feature and a core challenge of LLMs and other generative AIs because it allows them to produce generated (i.e. original) but also unverified sentences.

For example, an LLMs could finish the sentence “To be or not to be: that is the ____” in many different ways. LLMs are good at combining content, but they can also return answers that aren’t true.

GPT Scope

GPTs are pre-trained on data that cannot be changed or amended, but they can be fine-tuned on local data to better address specific questions. As with all data-related questions, the quality of the data greatly impacts the quality of the outcomes.

Not all questions are well-suited to be answered by an LLM. Using local data when needed, creating a well-worded prompt, and using human judgment to determine when it's appropriate to use generative AI are all important.