

In GPT Should We Trust?

Advice for interacting with generative language models

By Nicole DeCario and Peter Clark



By now, you've likely heard of or interacted with the phenomenon known as [ChatGPT](#). Its impressive capabilities have launched artificial intelligence into conversations at dinner parties, schools, major conferences, and on news channels across the globe. More recently, [Bing's new chatbot](#), built on the same technology as ChatGPT, has been making headlines as a "cheerful but erratic reference librarian" and a "moody, manic-depressive teenager" as summarized in [Kevin Roose's highly cited NYTimes piece](#). And last week, [GPT-4 debuted](#) as a multi-modal model that can notably process both text and images. Much has already been written about this game-changing technology; what is missing from popular conversations is a focus on [the importance of improving AI literacy](#) to support our understanding and responsible use of these new tools. Our goal in this article is to provide some clarity on the technology that powers the latest bots and offer practical advice for users to consider when interacting with these new, emergent companions.

Chatbots have been around for decades. [ELIZA](#), one of the first chatbots, was created in the 1960s, and users engaged with it as though it were a sympathetic human (the "[ELIZA effect](#)"). In 2014, the chatbot [Eugene Goostman](#) passed a simplified version of the famed [Turing Test](#) by cleverly selecting pre-written responses, similar to how a person might fake their way through a conversation they didn't really understand. Has anything changed with the launch of the latest bots? In one sense, no: large language models (LLMs), the type of technology that powers chatbots, simply predict the best response to a user's prompt. But in another, critical sense, something quite dramatic has changed. Recent systems appear to have learned that to *really* make good predictions, some internal "model" or understanding of the topic is needed, and these systems appear to build these (primitive) understandings internally. While this ability is still limited (and under debate), the emergence of a new level of "understanding" by machines is remarkable. It is also the cause of the recent frenzy, both for its potential benefits (e.g., improved search, summarizing long articles, personalized education, helping write software) and potential

risks (e.g., hallucinating false information, misleading users that it has feelings, providing bad suggestions).

New technology will continue to emerge, and we will knowingly and unknowingly interact with it. Chatbots, for example, already appear as customer service agents and healthcare appointment schedulers. They will become more ubiquitous as companies seek to leverage this improved technology.

So what can users do to remain aware and engage in a responsible way? Here are some tips:

1. Fact-check answers. A known limitation of current LLMs is that they can “[hallucinate](#)”, providing answers to questions that are simply wrong. This is particularly pernicious as they can present hallucinations with compelling authority (e.g., [AI Homework](#)).
2. Confirm citations. Some LLMs with access to the internet attempt to support their answers with citations back to source articles. Again, this appears authoritative, but always check that the cited source exists and that it does in fact support the answer.
3. Check the chain of thought. Current systems can recite a “chain of thought” arguing from what it knows to a conclusion, but sometimes that chain of reasoning can be wrong in subtle ways. For example (abbreviated), when asked how fund managers can influence share prices, ChatGPT [suggested](#): “*If the market price is high, the fund manager can remove shares, reducing supply and thus push the price down.*” This sounds eminently plausible, except the result of removing shares would actually be to push the price *up* (law of supply and demand), not down.
4. Avoid [anthropomorphization](#). Attributing human-like qualities to AI systems is [well-documented](#). Anthropomorphizing AI can lead to unrealistic expectations, perpetuating biases, and over- or under-estimating AI’s capabilities. Therefore, it is important to treat AI as what it is: a tool designed to perform specific tasks and not a conscious being.
5. Ignore emotional language. Remember that chatbots do not have feelings; they cannot be happy, angry, or sad. But they can sometimes convincingly fake such emotions, but a chatbot does not have a mind of its own — it has no internal motivations, desires, or intentions.

Despite the incredible new capabilities of LLMs, they are also fallible in unpredictable ways and do not follow the norms of intellectual evolution that humans go through. As a result, it will take time for us to learn their strengths and weaknesses and how to use them to their full potential while heeding the pitfalls many others have warned about.

To sum it up, we asked ChatGPT “should I trust ChatGPT?”, and this was its reply:

As an AI language model, I strive to provide accurate and helpful information to the best of my abilities based on the data and knowledge available to me. However, it’s important to note that I am not infallible, and there may be limitations and biases in the information and responses I provide. Therefore, it’s up to you to use your own judgment and critical thinking skills to evaluate the information I provide and decide whether or not to trust it. If you have any doubts or concerns, it’s always a good idea to do additional research and consult with experts in the relevant field.