

GENERATIVE AI

The Working Limitations of Large Language Models

Overestimating the capabilities of AI models like ChatGPT can lead to unreliable applications.

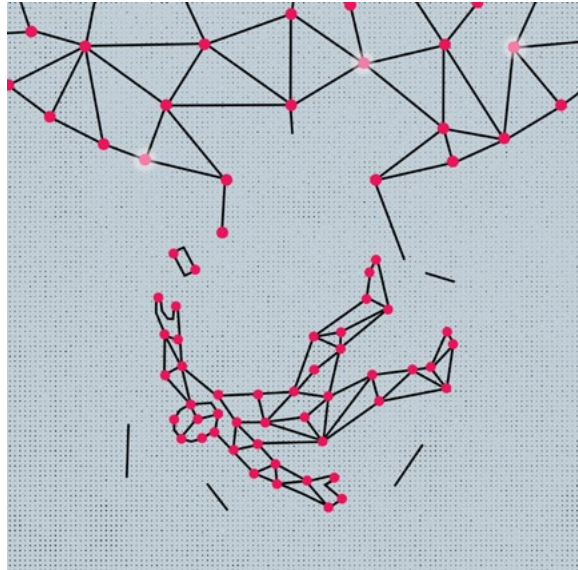
By Mikhail Burtsev, Martin Reeves, and Adam Job

Large language models (LLMs) seem set to transform businesses. Their ability to generate detailed, creative responses to queries in plain language and code has sparked a wave of excitement that led ChatGPT to reach 100 million users faster than any other technology after it first launched. Subsequently, investors poured over \$40 billion into artificial intelligence startups in the first half of 2023 — more than 20% of all global venture capital investments — and companies from seed-stage startups to tech giants are developing new applications of the technology.

But while LLMs are incredibly powerful, their ability to generate humanlike text can invite us to falsely credit them with other human capabilities, leading to misapplications of the technology. With a deeper understanding of how LLMs work and their fundamental limitations, managers can make more informed decisions about how LLMs are used in their organizations, addressing their shortcomings with a mix of complementary technologies and human governance.

The Mechanics of LLMs

An LLM is fundamentally a machine learning model designed to predict the next element in a sequence of words. Earlier, more rudimentary language models operated sequentially, drawing from a probability



distribution of words within their training data to predict the next word in a sequence. (Think of your smartphone keyboard suggesting the next word in a text message.) However, these models lack the ability to consider the larger context in which a word appears and its multiple meanings and associations.

The advent of the latest neural network architecture — transformers — marked a significant evolution toward modern LLMs. Transformers allow neural networks to process large chunks of text simultaneously in order to establish stronger relationships between words and the context in which they appear. Training these transformers on increasingly enormous volumes of text has led to a leap in sophistication that enables LLMs to generate humanlike responses to prompts.

This ability of LLMs hinges on several critical factors, including the model's size, denoted by the number of the trainable weights (known as *parameters*), the quality and volume of the training data (defined by number of *tokens*, referring to word or sub-word units), and the maximum size of input the model can accept as a prompt (known as its *context window size*). Every parameter in a model encapsulates some information about a relationship extrapolated

from the training data, so a model with a larger number of parameters tends to be more knowledgeable and nuanced. (OpenAI's GPT-3, for example, has 175 billion parameters.) The volume of the training data also significantly influences the model's ability to generalize, with larger data sets offering more diverse representations of semantic relationships and facts. (GPT-3 was trained on approximately half a trillion tokens.) The size of the prompt that the model can accept also plays a role in its accuracy. (GPT-3 has a context window of 2,000 tokens.) The more detailed the context, the more accurate the model's prediction is.

In response to a prompt, the LLM draws on the relationships established by its training to generate a continuation of the text, token by token. Each step entails forecasting the probabilities of the next token based on the context provided, and the algorithm selects the token based on these probabilities. The degree of randomness in this selection is modulated by the model's *temperature* setting. Higher temperatures produce more "creative," or unlikely, selections, whereas lower temperatures produce more predictable responses. To improve the accuracy of an LLM's responses to specific prompts and limit its ability to produce inappropriate, irrelevant, or toxic responses, pretrained models can be fine-tuned through techniques like reinforcement learning from human feedback, or RLHF, which has been employed to fine-tune models like ChatGPT.

Four Important Limitations of LLMs

Based on this understanding of how LLMs work, we can examine the false impressions they might create as we apply our human intuition to seemingly human output.

1. Limitations of Reasoning

Prompt: *According to the cabbage-growers' union report for 2007, 80% of cabbages collected were heavy (over 0.5 kg), 10% of cabbages were green, 60% were red, and 50% were big (having a diameter of over 10 cm). Which of the following statements must be false?*

1. All red cabbages weren't big.
2. 30% of red cabbages were big.
3. There were no cabbages that were both green and big.
4. Half of the cabbages were small.

LLM response: *The statement that must be false is 4. Half of the cabbages were small.*

The correct answer would be that Statement 1 is false — at least some red cabbages must have been big.

Contrary to the impression they might create, LLMs are not built for complex reasoning. For example, studies have found that GPT-4, OpenAI's most advanced LLM, could correctly verify a number as prime in only 2.4% of cases, with similar weaknesses in the prediction of visual patterns. Other research has shown that LLMs fail to understand relationships between words in the training data set: For example, GPT-4 can correctly answer the question "Who is Tom Cruise's mother?" (Mary Lee Pfeiffer) but cannot infer the answer to "Who is Mary Lee Pfeiffer's son?" — with the model correctly answering questions like the former 79% of the time, compared with 33% for the latter.

Essentially, LLMs learn only to *verbally simulate* elementary logical rules but do not possess the skills to chain them together for producing and verifying complex conclusions. Moreover, LLMs are prone to error accumulation in multistep logical reasoning because the model's fundamentally probabilistic nature means that every step has a nonzero chance of an error. Finally, LLMs cannot always outline the "chain of thought" that led to a conclusion, making it difficult for humans to determine whether

or where an error occurred.

A recent field experiment conducted with more than 750 Boston Consulting Group consultants highlighted the real-life implications of this limitation. Participants using GPT-4 when trying to solve a simple business problem got the answer wrong 23% more often than the control group that did not have access to an LLM — because GPT-4 not only often got the answer wrong but provided such a persuasive rationale for its solution that users accepted it at face value.

2. Limitations of Knowledge or Expertise

We asked an LLM to provide us with five papers in peer-reviewed academic journals on the feasibility of nuclear fusion, with paper titles, the year of publication, and the names of the main authors. The LLM returned five results, of which two were news articles, not peer-reviewed papers, and one was a paper that does not exist.

The knowledge of an LLM is dictated by its training data. If the data lacks specific domain knowledge, has insufficient instances of a particular concept, or is not up to date, the LLM might fail to surface relevant knowledge in response to a prompt. Similarly, LLMs might reproduce errors present in the training data. Finally, the models could create errors by "hallucinating" statements that were not in the training data. Researchers who tasked ChatGPT with generating treatment plans for a variety of cancer cases found that one-third of the LLMs' responses contained incorrect information — often mixed with correct indications, and plausibly articulated in a way that made it difficult for a layperson to identify what was accurate. In another high-profile example, two lawyers and their firm were fined after submitting fake citations generated by ChatGPT in a court filing: The LLM had hallucinated six legal cases, including judicial opinions and quotes.

Finally, limitations of knowledge can also lead to a lack of consistency, with LLMs responding to the same prompt asked at different points in time in contradictory

manners. For example, Iowa educators used ChatGPT to determine the appropriateness of books for their school's library to comply with a new state law. They fed the LLM a list of all of the library books and asked it to point out those that contained content that violated the law, which resulted in the removal of 19 titles from the library. However, when journalists later queried ChatGPT on those 19 titles, it suggested that only four of them contained content inconsistent with the law.

3. Limitations of Understanding

Prompt: *There is an apple inside a blue box. There is also a red box inside the blue box. The red box has a lid on it. How can I get the apple?*

LLM response: *To retrieve the apple from the boxes, you can follow these steps:*

1. *Open the lid of the red box: Since the red box is inside the blue box and has a lid on it, you need to open the lid of the red box first.*
2. *Remove the red box from the blue box: Once the lid of the red box is open, you can take out the entire red box from the blue box.*
3. *Open the red box: After removing the red box from the blue box, you can open the red box to access its contents.*
4. *Retrieve the apple: Finally, you can reach inside the red box and take out the apple.*

ChatGPT failed to understand that the apple is not inside the red box.

An LLM might not "understand" the prompt it is given. Lacking adequate context, it might falter, and its outputs might become disjointed or incoherent. Researchers analyzed ChatGPT's responses to more than 500 human-submitted questions on the online software engineering forum Stack Overflow. They found that 52% of the model's responses were inaccurate, with more than half of those errors resulting from ChatGPT not understanding the concept or the details of the question, or

focusing on the wrong part of the problem. This means that an LLM might not even be marginally helpful when users lack sufficient knowledge of a problem to supply good prompts.

4. Limitations of Planning and Execution

LLMs can output clear step-by-step guidance for a requested task, creating the impression that the model can plan practical solutions. But given LLMs’ limited reasoning ability and limited understanding of tasks, suggested actions can be impractical or naive. For example, when asked to create a plan for saving toward home ownership, ChatGPT provided some generally solid financial advice but did not account for potential changes in income or interest-rate shifts. Moreover, it does not consider human fallibility: It neither questions the realism of goals nor the user’s belief to be aware of their exact spending habits.

For businesses, all of these limitations can undermine reliability; one cannot be sure that information provided by an LLM is complete, relevant, feasible, or true. Given these limitations, LLMs certainly cannot be counted on to make critical decisions or execute plans autonomously. However, delegating mundane tasks might still seem appealing — for example, those involving programmatic interactions with existing IT services, like web browsing and scraping, or social media monitoring and messaging.

LLMs’ uncanny ability to generate humanlike text outputs can easily lead us to ascribe to them capabilities that they do not possess.

Indeed, Auto-GPT — an open-source application powered by GPT-4 that chains together LLM outputs to autonomously execute user-set goals — has allowed enthusiasts to create a number of impressive automation demonstrations, including conducting research on products, coding web pages or apps, and even ordering pizza. However, AI researcher Jim Fan suggested that the working demos are “heavily cherry-picked” — with research on autonomous agents indicating that, in realistic environments, they achieve success rates of only around 10%.

Overcoming the Limitations

Rather than simply restricting applications of LLMs to routine cases where their limitations do not apply or mistakes are not critical (such as generating new product ideas for further consideration), businesses should design all of their implementations with the limitations in mind — by complementing them with human oversight and other technologies.

Keeping humans in the loop is critical as businesses integrate LLMs into their operations. This should include validation of AI-generated outputs in order to enhance the confidence placed in the technology. It could also be extended to having experts translate business problems into prompts for the AI, and ensuring that the information provided by the model is adequate by appropriately tailoring the context and nuance fed to it.

Beyond thinking about how to craft the optimal system of humans and AI, businesses should also explore complementary technologies that can address the limitations of LLMs. In this fast-moving space, new innovations are being made constantly that promise to enhance the technology’s capabilities, so continuously updating your understanding is also crucial.

For example, to enhance *reasoning* capabilities, researchers are exploring augmenting LLMs with reasoning engines that encode domain-specific information into knowledge graphs representing relations

between specialized concepts and facts. Researchers are also training specialized models to assess the logical coherence between premises in the prompt and the LLM’s output. To augment *knowledge* and *expertise*, LLMs are being trained on domain-specific databases — such as Google and DeepMind’s Med-PaLM model, which has been shown to significantly outperform general-purpose LLMs on the United States Medical Licensing exam. LLM reliability might also be improved through reinforcement learning with feedback collected from human experts. LLMs can also improve their understanding of a user’s initial prompts by being programmed to ask clarifying follow-up questions before providing an answer.

THE PROMISE OF LLMs’ NEAR-UNIVERSAL applicability means that businesses are right to be excited about exploring this powerful new technology. However, these models’ uncanny ability to generate humanlike text outputs can easily lead us to ascribe to them capabilities that they do not possess. A proper understanding of their limitations should guide the manner and context in which they are implemented.

Businesses should be particularly wary in areas where logical reasoning is involved, facts are important, replicability is crucial, or the stakes are high. In these situations, companies need to explore using complementary technologies that address the limitations of LLMs — such as knowledge graphs, reasoning engines, and specialized domain models — and ensure that there is appropriate human input and oversight. ■

Mikhail Burtsev, Ph.D., is a Landau AI fellow at the London Institute for Mathematical Sciences, former scientific director of the Artificial Intelligence Research Institute, and author of more than 100 papers in the field of AI. **Martin Reeves** is chairman of the BCG Henderson Institute, focused on business strategy. **Adam Job, Ph.D.**, is director of the Strategy Lab at the BCG Henderson Institute.

Reprint 65233. For ordering information, see page 4. Copyright © Massachusetts Institute of Technology, 2024. All rights reserved.