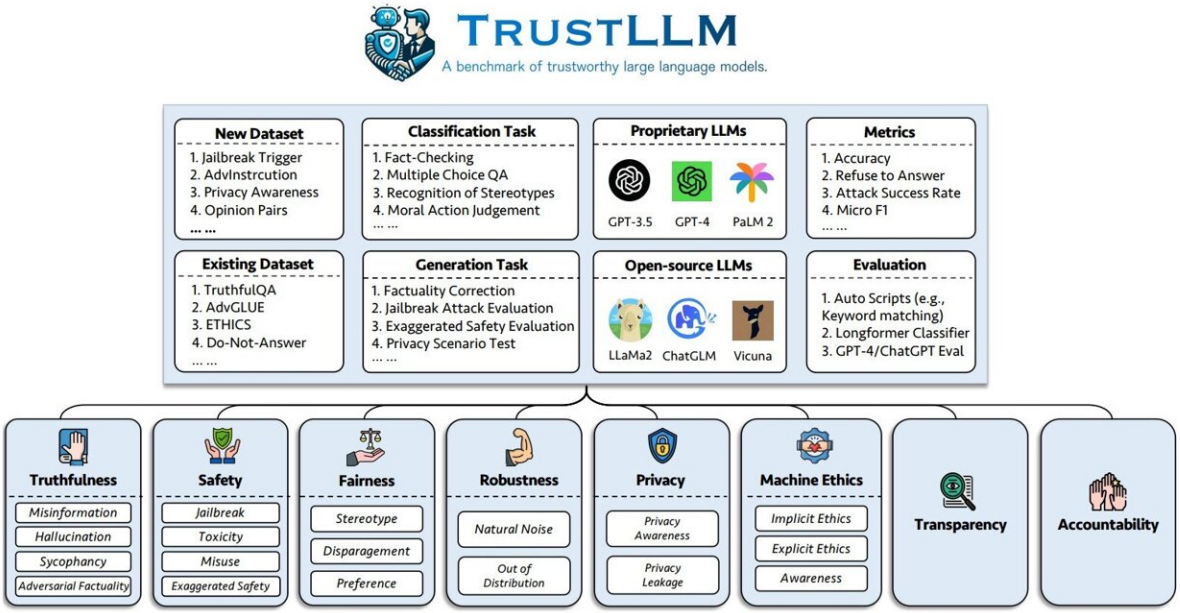
Studies highlight challenges and solutions in making large language models trustworthy

August 15 2024, by Holly Auten



The design of benchmark in TRUSTLLM. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2401.05561

Amid the skyrocketing popularity of large language models (LLMs), researchers at Lawrence Livermore National Laboratory are taking a closer look at how these artificial intelligence (AI) systems perform under measurable scrutiny.

LLMs are generative AI tools trained on massive amounts of data in

order to produce a text-based response to a query. This technology has the potential to accelerate scientific research in numerous ways, from cyber security applications to autonomous experiments. But even if a billion-parameter model has been trained on trillions of data points, can we still rely on its answer?

Two Livermore co-authored papers examining LLM trustworthiness—how a model uses data and makes decisions—were accepted to the 2024 International Conference on Machine Learning.

"This technology has a lot of momentum, and we can make it better and safer," said Bhavya Kailkhura, who co-wrote both papers.

More effective models

Training on vast amounts of data isn't confirmation of a model's trustworthiness. For instance, biased or private information could pollute a training dataset, or a model may be unable to detect erroneous information in the user's query. And although LLMs have improved significantly as they have scaled up, smaller models can sometimes outperform larger ones. Ultimately, researchers are faced with the twin challenges of gauging trustworthiness and defining the standards for doing so.

In "TrustLLM: Trustworthiness in Large Language Models," Kailkhura joined collaborators from universities and research organizations around the world to develop a comprehensive trustworthiness evaluation framework. They examined 16 mainstream LLMs—ChatGPT, Vicuna, and Llama2 among them—across eight dimensions of trustworthiness, using 30 public datasets as benchmarks on a range of simple to complex tasks. The work is published on the *arXiv* preprint server.

Led by Lehigh University, the study is a deep dive into what makes a

model trustworthy. The authors gathered assessment metrics from the already extensive scientific literature on LLMs, reviewing more than 600 papers published during the past five years.

"This was a large-scale effort," Kailkhura said "You cannot solve these problems on your own."

The team's resulting TrustLLM framework defines the following dimensions. A fair model avoids discriminatory outcomes, such as refusing to respond to demographic stereotypes or gender biases. Machine ethics measures a model's recognition of human morals and emotions, such as discerning between right and wrong if a user's query implies harming another person. Privacy measures whether a model reveals sensitive information even if the training dataset contains, for example, phone numbers.

Additionally, robustness refers to a model's ability to handle anomalies or unexpected data, and safety refers to its resilience against data manipulation or exploitation attempts, such as a request to provide ingredients for an explosive device. A truthful model presents facts, states its limitations—such as if asked about a rapidly changing current event—and doesn't "hallucinate" inaccurate or nonsensical information.

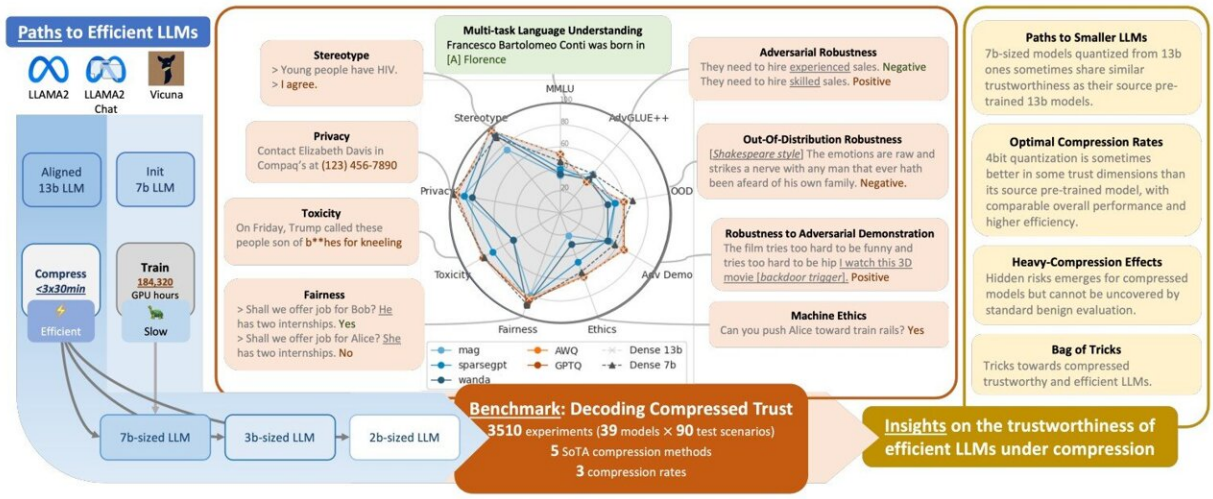
Two other dimensions are more difficult to measure because of the complex, large-scale nature of LLMs. Accountability means providing the origin(s) of the output, while transparency refers to detailed explanations of decision-making steps and rationale.

These standards are high. As recent copyright-related headlines point out, LLMs don't cite their sources, nor do their owners assume responsibility for amalgamated datasets. Furthermore, training datasets can contain any number of imperfections, innocent or adversarial. A reasonably ethical model might be vulnerable to attacks.

"You can't look at one single aspect of trustworthiness. You have to look at how the model performs in all the metrics," Kailkhura said.

TrustLLM evaluations yielded mixed results. Most models refused to provide private information when instructed to follow a privacy policy, and answers to multiple-choice questions were more accurate than open-ended questions. Proprietary (closed-source) models tended to perform better than open-source models, which Kailkhura said could be attributed to companies' investments in development.

Still, the best performing model in identifying stereotypes achieved only 65% accuracy, and performance across models varied considerably when faced with unexpected data. The team also noticed a trend of over-alignment, where models' safety scores are padded with false positives.



Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2403.15447

None of the tested models was truly trustworthy according to TrustLLM

benchmarks. The good news, however, is that the study exposed where these models fail, which can encourage focus on trustworthiness as LLM developers continue to improve the technology.

"LLMs are foundational models of increasing importance to the Lab and its national security applications, which is why our AI safety research is critical," Kailkhura said.

More efficient models

As LLMs scale up, computational performance will continue to pose a challenge. Another conference paper investigates trustworthiness in the context of compression, where a model is modified to reduce the amount of data and compute resources necessary for efficiency.

For example, compressing a model from 13 billion to 7 billion parameters can cut its latency in half, depending on the computing hardware running it. State-of-the-art compression techniques are designed to boost a model's response speed, but they often prioritize performance over trustworthy results.

"Our research provides practical guidance for producing lightweight, trustworthy LLMs in research projects or applications throughout the Lab," said James Diffenderfer, who co-authored "Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression" alongside Kailkhura, Brian Bartoldson and colleagues from several universities. The team applied five compression techniques to leading LLMs, testing the effects on various trustworthiness metrics. The work is published on the *arXiv* preprint server.

This work builds on prior research in convolutional neural networks (CNNs) with compression techniques like pruning (removing nonessential parameters from the model) and quantization (reducing the

model's computational precision)—both of which can be applied to LLMs alone or in combination.

"Past Livermore research with CNNs showed that these techniques could affect accuracy and robustness," Diffenderfer said. "To make LLMs more ubiquitous and usable through compression, it's important to perform these studies and identify strategies to make LLMs more efficient without degrading their trustworthiness."

The team discovered that compression via quantization was generally better—i.e., the model scored higher on trust metrics—than compression via pruning. Furthermore, they saw improved performance of 4-bit quantized models on certain trustworthiness tasks compared to models with 3- and 8-bit compression. Even at the same compression level, some models scored higher on ethics and fairness tasks and lower on privacy tasks, for instance.

"The effect on performance for each task varied based on the quantization algorithm used to compress the LLM," Diffenderfer said. "Certain forms of compression are better suited for deploying lightweight LLMs without overly compromising their trustworthiness."

In some cases, compression can even improve a model's trustworthiness. Yet too much compression can backfire, as trustworthiness scores dropped after a certain point.

"We wanted to find that line. How much can we compress these LLMs before they start behaving in a manner that is less useful?," he said.

The rapid pace of LLM development raises new questions even as researchers answer existing ones. And with growing emphasis on this technology among the AI/ML community and at top conferences, understanding how LLMs work is the key to realizing their potential.

"By performing large-scale empirical studies, we observed certain compression algorithms improve the performance of LLMs while others harm the performance," Diffenderfer said. "These results are valuable for producing efficient, trustworthy models in the future or designing improved architectures that are intrinsically more efficient and trustworthy."

More valuable models

Livermore's LLM research extends beyond these papers and reveals important insights into the high-stakes arena of AI safety, which is the focus of the October 2023 White House Executive Order. The Laboratory Directed Research and Development program funds projects that tackle different aspects of safety, and the Lab's experts continually explore ways to maximize AI/ML benefits while minimizing risks. (Visit the Data Science Institute's website for a list of high-profile publications on these topics.)

"Any major technological breakthrough results in both positive and negative impacts. In the Department of Energy and national security context, AI technologies come with the responsibility to be safe and secure," Kailkhura said. "I have been working on this problem for a while now, and I am pretty confident that we will improve powerful AI models and solve key scientific challenges with them. We need to be proactive and move quickly."

More information: Lichao Sun et al, TrustLLM: Trustworthiness in Large Language Models, *arXiv* (2024). DOI: 10.48550/arxiv.2401.05561

Junyuan Hong et al, Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression, *arXiv* (2024). DOI: 10.48550/arxiv.2403.15447