White paper

Why We Need Trusted AI

Mitigating Risks and Maximizing Rewards of Generative AI Models

David Loshin
President, Knowledge Integrity, Inc.

teradata.

Version 1.0

Table of Contents

Introduction

The hype of generative AI has raised the profile of the power and potential of artificial intelligence across the spectrum from the boardroom to consumers. And that hype is rapidly transitioning to hope for organizational acceptance of Large Language Models (LLMs) and other AI and machine learning applications. Organizations are evaluating the business needs and potential for substantial benefits in many new use cases that can be derived from adopting technologies like generative AI, especially as they acknowledge the continuing need to master predictive AI into their business processes and corresponding applications.

The expanding breadth of the AI market adds to the challenges of successful adoption. There is a wide variety of products emerging, each with its own benefits, drawbacks, and constraints on adoption, deployment, and use. Different players (including ones with significant investments in developing generative AI technologies) are collaborating and entering into agreements of different shapes and forms that ultimately impact the choices organizations make regarding which solutions they choose. And accompanying the enthusiasm for adoption is a need to evaluate the total cost of operations of adopting emergent algorithmic approaches such as generative AI and LLMs.

These organizations are now piloting LLMs and generative AI techniques to assess their business potential and understand the best approaches for deployment to ensure that it helps create value. In many instances, there is a steep, but achievable, learning curve to understand how these emergent AI environments and models are trained, fine-tuned, and produce results. Yet implementers need to be aware of potential risks that can lead to raised expectations about the degree of trustworthiness of the outputs and responses. In particular, one area that requires additional oversight involves managing risks that could occur abruptly or emerge over time as adoption and reliance on generative AI continues to grow.

Generative AI is emblematic of the broader need for trust in the results and outcomes derived from AI, and in this paper we will delve into some of those risks and explore the types of impacts ignoring those risks might lead to. The need for trust in AI systems is at a crucial tipping point, as organizations must adopt criteria and strategies to ground AI deployments with the appropriate values to ensure trust and transparency.

The risks of adopting generative AI technologies

If the benefits of adopting generative AI seem obvious, the risks may be somewhat obscured. NIST (The National Institute of Standards & Technology) has published an Artificial Intelligence Risk Management Framework along with a Generative Artificial Intelligence Profile that discusses organizational risks that are exacerbated by generative AI. The framework addresses risks that are relevant for commercial businesses, such as the production of erroneous content ("hallucinations"), biases integrated into the models, automation bias, opportunities for fraud, information security and protection, the need for protection of intellectual property, sustainability, and sustainability, among others.[1]

It is valuable to consider these risks as hurdles to overcome, not barriers to success. As we review these risks, take into account some ways that people, transparency, and value creation help to mitigate these risks. For example, **building the right data environment** can help control enterprise-wide data quality, which contributes to reducing hallucinations and identifying and eliminating bias. **Leveraging the new AI ecosystem** not only helps the organization navigate the complexity of the AI market, it allows you to take advantage of integrated technologies to protect against exposure of sensitive data or mistaken use of others' intellectual property. Instituting human accountability and governance within a performance-oriented platform environment allows you to leverage your employees' AI skills and talent, enabling the organization to **make AI innovation work.**

1. NIST AI 600-11, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," accessed 2024-05-09 via https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf

Hallucinations

The NIST Generative AI Profile defines an AI confabulation , which is also referred to as hallucination, as "a phenomenon in which generative AI systems generate and confidently present erroneous or false content to meet the programmed objective of fulfilling a user's prompt." LLMs are probabilistic in predicting reasonable responses to provided prompts, and these responses are based on the data used in training the models. When there are flaws, biases, or insufficient information in that training data, it would not be unexpected that chatbots based on LLMs built using those faulty assumptions would confidently respond with statements that are not true.

For example, in 2023, a plaintiff's attorney used ChatGPT to craft a motion for a legal brief for his client's lawsuit. The lawyer produced a 10-page brief that cited more than six court decisions supporting their client's argument, but the defendant's lawyers were unable to locate those cases. It turned out that the cases were not real - apparently, the generative AI system had created references to "invented" cases that did not exist.[2]

Because the LLMs are unable to distinguish between truthful responses and ones that are completely fictitious, an ingenuous AI user may not be aware when hallucinations have been generated. This can pose a significant risk when the AI systems are being used for critical decision-making scenarios, such as financial applications, medical analyses and diagnoses, transportation applications, or security.

Integrated bias

Even with the most pristine data, the precision and accuracy of results produced by a sophisticated AI model remain at the mercy of the selection of inputs used for its training. This is particularly critical when groups or entities are either overrepresented or underrepresented in the training data. The outcome is that the biases in the training data become "baked into," into the AI model, leading to integrated prejudices against ethnicities, people of different cultures, gender, or sexual orientation

There have been some notorious AI failures due to integrated bias, such as:

- In a 2023 study published by the University of East Anglia, a team of researchers revealed that despite assurances of impartiality, a "significant and systematic left-wing bias" was found in ChatGPT's responses.
- When both ChatGPT-3.5 and GPT-4 were asked about how to calculate eGFR (Estimated Glomerular Filtration Rate, a measure associated with kidney function), the models had runs that "tried to justify race-based medicine with false assertions about Black people having different muscle mass and therefore higher creatinine levels."
- Biased Image Generation: Image generators such as Stable Diffusion and DALL-E were shown to imbue their generated images with lingering cultural biases. For example, when prompted for a portrait of "a person at social services," the generated images tended to be people of color, while a prompt for "a productive person" produced images of white males.

Using AI models trained using data that was lacking in diversity not only complicates diagnosing and treating illness among underrepresented groups, it will also affect the development of drugs that are only effective on a subset of the population.

Integrated AI bias is particularly dangerous when AI models are used in support of healthcare.

2. Benjamin Weiser and Nate Schweber, "The ChatGPT Lawyer Explains Himself," The New York Times, June 8, 2023, accessed 2024-06-18 via https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

Automation bias

As organizations become increasingly reliant on automated reporting, analytics, and AI systems to inform and influence business decisions, people run the risk of falling prey to automation bias. Automation bias is the result of overconfidence in automated systems and is the predisposition to prefer to agree with or believe in recommendations from automated decision-making systems while discounting contradictory information that is observable without the help of an automated system, even if that contradictory information is correct.

This issue becomes much more acute with AI technologies like LLMs. People tend to adjust their actions based on their perception of the level of risk, letting down their guard when using systems with human-like chatbot capabilities that provide a veneer of humanity over the computational system that lend a greater level of credence to generated outputs. This becomes even more concerning when using "black box" AI models generate outputs with no explainability about how those outputs were produced and when there are no guardrails in place to verify the accuracy of the results.

Automation bias comes into play when using generative AI systems that have integrated bias resulting from training data with overrepresented groups. For example clinicians making healthcare treatment decisions based on systems using biased AI models may disregard symptoms of individuals from underrepresented communities even though those symptoms might suggest alternative diagnoses requiring different healthcare protocols.

Failure to acknowledge the dangers of exposing any type of sensitive information can run the risk of regulatory noncompliance, fines and penalties, as well suffer damage to the organization's reputation.

Exposure of sensitive data

Effectively training an LLM requires a massive-scaled, yet unrestrained data collection accumulated from a variety of sources. This data collection process may include information that could be classified as an individual's personally identifiable information (PII) or information considered to be private data. Data harvested from sources in ways that are inconsistent with their original intent not only may violate individual consent preferences for data use but may also violate data privacy laws governing the appropriate use of a data subject's personal information.

And exposure of PII is not the only vulnerability, especially when an organization's employees use publicly available generative AI tools. When either fine-tuning or prompting these AI tools, employees potentially input other types of corporate sensitive information into LLMs, including employee information, customer information, customer and employee contact data as well as customer account numbers and other banking details, corporate financials, intellectual property, company secrets, username/password login credentials, or other types of material nonpublic information. Once this data is input to the publicly available AI system, that becomes incorporated into the pool of data used to continuously refine the model

Because the outputs produced by Generative AI systems are predicted based on the information contained in the data sources employed for training, any sensitive information used during the training process is subject to exposure when the models are put into use.

This can happen due to several reasons:

- Personally identifiable information (PII) is used without masking or other types of anonymizations as the models are trained, allowing for that PII to be inadvertently leaked when the model is being used.[3]
- There are insufficient access controls exercised at the front-end, allowing for unauthorized access to sensitive data.[4]
- Prompt engineering tactics can be used to continually refine the prompts to elicit or infer sensitive information from the system.[5]

3. Motoki, F., Pinho Neto, V. & Rodrigues, V. More human than human: measuring ChatGPT political bias. Public Choice 198, 3–23 (2024). https://doi.org/10.1007/s11127-023-01097-2
4. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digit Med. 2023 Oct 20;6(1):195. doi: 10.1038/s41746-023-00939-z. PMID: 37864012; PMCID: PMC10589311.
5. Nitasha Tiku, Kevin Schaul, Szu Yu Chen, "These fake images reveal how AI amplifies our worst stereotypes," Washington Post 11/01/2023, accessed 2024-05-13 via https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/.

Copyright violations

Another facet of the massive appetite for data that LLMs require is the level of discrimination the developers employ in choosing the data sets used in training. This raises a potential issue when AI tool developers and generators use copyrighted content and violate the rights of the original artists. For example, in late December 2023, the New York Times sued OpenAI and Microsoft for copyright infringement, contending that "millions of articles published by The Times were used to train automated chatbots that now compete with the news outlet as a source of reliable information." [6]In another example, some published authors sued Meta, contending that much of the material in Meta's training dataset (for their LLaMA LLM) "comes from copyrighted works—including books written by Plaintiffs—that were copied by Meta without consent, without credit, and without compensation."[7]

The issue is that generative AI systems trained using authors' works pose a threat to the artists by allowing the market to be flooded with content that is generated "in the style of" those artists, thereby damaging the profession. The AI companies suggest that using those works are training input constitutes "fair use" of the material. Until the question of whether AI LLMs incorporation of others' intellectual property is considered to be "fair use" is addressed by the courts, a responsible organization should monitor the data sources used as inputs and track whether there are any potential violations of rights.

6. Michael M. Grynbaum, Ryan Mac, "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work," December 27, 2023, accessed 2024-06-18 via https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html
7. The complaint is accessible via https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.1.0_1.pdf

Sustainability issues

The computational needs and energy demands to build, train, and maintain large language models and generative AI systems may raise some eyebrows when considering global issues and the need for moderating energy consumption as part of a sustainability objective. In 2019, "researchers found that creating a generative AI model called BERT with 110 million parameters consumed the energy of a round-trip transcontinental flight for one person… Researchers estimated that creating the much larger GPT-3, which has 175 billion parameters, consumed 1,287 megawatt hours of electricity and generated 552 tons of carbon dioxide equivalent, the equivalent of 123 gasoline-powered passenger vehicles driven for one year."[8]

More to the point: The scale of computational resources needed to support existing and planned AI systems is eye-popping. For example, xAI uses 100,000 liquid colled H100 Nvidia GPUs to train the next version of Grok.[9] Meta seeks to establish computer power equivalent to 600,000 Nvidia H100 GPUs in developing its next generation AI. Cloud service providers like Google and AWS are also ramping up their GPU cluster service offerings, opening the door for their customers to train and launch their own AI applications. If the development and adoption of AI-driven applications continues to grow at current rates, the energy demands will rise dramatically. Data centers drawing their power from non-renewable energy sources will have a significant environmental impact.

8. Kate Saenko, "A Computer Scientist Breaks Down Generative AI's Hefty Carbon Footprint," 05/23/2023, accessed 2024-05-09 via https://www.scientificamerican.com/article/a-computer-scientist-breaks-down-generative-ais-hefty-carbon-footprint/
9. Kate Irwin, "Elon Musk's xAI Powers Up 100K Nvidia GPUs to Train Grok," accessed 2024-08-22 via https://www.pcmag.com/news/elon-musk-xai-powers-up-100k-nvidia-gpus-to-train-grok Michael Kan, "Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs," Accessed 2024-08-22 via https://www.pcmag.com/news/zuckerbergs-meta-is-spending-billions-to-buy-350000-nvidia-h100-gpus

The need for trust in AI

Increasingly sophisticated AI technologies can create tremendous opportunities for those organizations that choose to adopt and incorporate AI systems into their business processes. At the same time, the types of risks discussed in this paper raise questions about the level of trust one can put into the results of those technologies.

But if these issues are exacerbated by ungoverned AI integration, an alternative approach considers the need for trust in AI. A company that rushes to incorporate Generative AI and LLMs into the enterprise without assessing the risks associated with ensuring a level of trust will jeopardize the success of their implementation. Organizations must ensure that AI systems produce desired value-adding outcomes in ways that are trustworthy, that the business remains accountable for the outcomes, and that there is transparency in how the results were produced.

Seek out the right partners who can support the development of appropriate governance frameworks to address vulnerabilities and to mitigate risks. Institute the appropriate data quality controls and oversight to reduce or eliminate the production of unverified or unvetted results that creates an exposure to inadvertent reliance on misinformation. Consider alternatives for model monitoring and governance, which can help to address the lack of transparency about the data used for training that opens the door to biased results, as well as abusive, toxic, disparaging, or stereotyping content. Establish relationships with partners whose guidance can help eliminate the integrated biases that can lead to dangerous recommendations, especially in the healthcare arena. Devising the right strategies to ground AI deployments with the appropriate values to ensure trust and transparency will help your organization understand the best ways to navigate through the risks to derive the best benefits from the new generation of AI technologies.

Consultant Bio

David Loshin is a recognized thought leader and expert consultant in the area of data strategy, information risk, and information innovation. David is a prolific author regarding information management best practices as the author of numerous books and papers, including Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph and The Practitioner's Guide to Data Quality Improvement. David is a frequent invited speaker at conferences, web seminars, and sponsored websites and channels. David is also a Senior Lecturer and Director of External Relations program at the University of Maryland's College of Information Studies

David Loshin
President, Knowledge Integrity, Inc.

What is Trusted AI?

Accelerate your innovation journey with Teradata

Business leaders believe AI is the future, yet the path to that future isn't clear or easy. And trust remains a serious concern. But what does it mean to have Trusted AI?

At Teradata, we believe Trusted AI is the way that people, data, and AI work together — with transparency — to create value. With Teradata's three principles of Trusted AI — people, transparency, and value creation — your organization can address concerns about trust in AI and help inspire a new era of creativity and confidence in decision-making.

Explore Trusted AI

About Teradata

At Teradata, we believe that people thrive when empowered with trusted information. We offer the most complete cloud analytics and data platform for AI. By delivering harmonized data and Trusted AI, we enable more confident decision-making, unlock faster innovation, and drive the sustainable, successful business results organizations need most.
See how at **Teradata.com**

17095 Via Del Campo, San Diego, CA 92127 Teradata.com

teradata.