Managing GenAI Risk: The Supply Chain

JULY 15, 2023

Ben Ellerby, **aleios**

Generative AI (GenAI) has captured the imagination of millions in just a few months since the release of high-profile applications like ChatGPT, DALL-E and Midjourney. After being stunned by the seemingly magical interaction with these early applications, many of us started to see the real-world uses in our work.

My entire engineering team leverages Copilot (the AI pair programmer from GitHub powered by a GPT model), and we've built several productivity tools on top of our existing knowledge banks. Our clients have also been quick to react and ask for help in leveraging this technology. Right now, many use cases are for obvious customer support and information retrieval, but creative/knowledge work is already in the sight of many.

Almost as heavily publicised as the benefits of such technologies are the regulatory and privacy concerns. Even OpenAI, the company behind ChatGPT, has received attention from regulators across the globe. Other companies, including Goldman Sachs and JP Morgan, have also been quick to ban ChatGPT due to data security and intellectual property concerns. The US Federal Trade Commission (FTC) is now demanding disclosure on the data used to train ChatGPT's underlying models in reaction to false accusations made by ChatGPT and alleged copyright infringement.

The core technology behind GenAI, Large Language Models (LLMs), are not inherently risky but require such a large volume of data and learning that can lead to issues when not handled well. When LLMs are trained on petabytes of public data, it's not inconceivable for intellectual property to be picked up, leading to complex questions about the ownership of generated materials.

Moreover, these models are further trained on their interactions, which means that the content of "prompts" (questions asked by the user) can be "learned" by the model, potentially leading to leaks. For example, if I were working at a marketing firm and asked a free public GPT application for advice on how to handle the secret launch of a new soda by a large beverage provider, it's possible that an executive at a direct competitor could be asking the same application for new product ideas. As people begin to use GenAI applications to automatically take notes during all their business meetings, the situation becomes concerning.

Most of the work I've been involved in with GenAI has been in helping clients understand and mitigate risks. Simply put, risks exist throughout the supply chain of the GenAI application, from model training and foundational model selection through to prompt engineering, context, hosting, fine tuning and integration. By understanding, mitigating, and documenting these risks, companies can leverage the power of GenAI while avoiding legal, societal, and reputational risks.

Foundational Model Training / Selection

LLMs are trained using deep learning techniques on extremely large datasets, much larger than those typically used to train previous deep learning models. The source of the data, as well as the licensing and guarantees on this data, are crucial in avoiding unintentional infringement of IP protections.

For most companies, the use of GenAI will involve the use of a third-party Foundational Models (FMs). Due to the size and scale of the parameters involved in these large FMs, it is almost impossible to reference back to sources, leading to a grey area when it comes to defending against IP claims. To reuse the simple "garbage in ⇒ garbage out" analogy from the ML world, "IP in ⇒ IP risk out". The best way to mitigate this risk is to conduct due diligence on third-party FMs in use, or indeed, to establish a rigorous process for data validation if you are building a specialized FMs.

Apart from the IP risks, when training your own FM there is also a data security risk. For example, if you are a large financial institution that leverages an internal knowledge bank, you may decide to train a GPT-powered assistant to make your staff more efficient. The large corpus of data from your knowledge bank and CRM could be used as part of the training data. Then, an analyst may ask a query from a client and get access to data from or about another client, with potentially sensitive information. This would bypass the traditional "role-based access controls" that your other systems and tools have in place. Knowing the data's provenance, purpose, and permissions required to access it in the current organization is crucial to ensuring that a leaky model, even if only used internally, is not created.

Fine Tuning

LLMs can be customized for a particular use case through a process called "Fine Tuning". In this approach, a copy of a trained model with known weights is taken and then trained again with a use-case-specific data set, starting from the initial weights. This process can be made more efficient by training a subset of weights in an approach called "Parameter Efficient Tuning".

However, this process presents an opportunity for the model to be contaminated with IP-infringing data sets or sensitive information. Fortunately, the volume of data used for fine tuning is typically orders of magnitude lower than that used to train the FM. This means that auditing is simpler, and mistakes are less likely. Nonetheless, it is essential to have strong processes in place regarding the provenance, purpose, and permission level of data used for fine tuning, as well as to educate teams and clearly document any training for audit and accountability purposes.

Hosting

The location where a model resides during training and inference is crucial, especially when companies customize foundational models using high-value intellectual property. Without secure hosting, the intellectual property and sensitive information used to train the model could be at risk. Additionally, the prompts and outputs may contain sensitive information, and the roundtrip from the application layer to the model and back needs to be secure and compliant with data privacy laws. It is also important to have control over where models are trained and hosted to ensure compliance with data sovereignty laws.

Mapping out the end-to-end flow of application prompts to outputs, the locations and security responsibilities of any outsourced components is key to understand your holistic hosting picture. Cloud providers have been quick to make hosting options available with regional segregation and security controls.

Application

When providing users with access to the outputs of LLMs, it is crucial to manage the risk of hallucination. An LLM may confidently and plausibly respond with an answer, even if it is false, leading to defamation lawsuits if individuals are falsely accused of crimes with fake links to non-existent news articles. This creates a risk of defamation and of false information impacting critical business processes.

While researchers are working on creating new FM that are less susceptible to hallucinations, it is important to educate end users and incorporate UI features that remind users of the potential for false information.

In addition, safeguards should be implemented to protect against malicious use of public-facing applications, such as the creation of fake news or inappropriate content (profanity, violence, hate speech, etc.). New FM with built-in protection, such as Amazon's Titan FMs, are also emerging.

Context Embeddings

Vector embeddings enable LLMs to search for similar data. Vector embeddings are generated from application data and stored as vectors inside a database. A user's query can then be converted to an embedding, and a similarity search can be performed, providing a prompt to the LLM with formatting and further information. However, if sensitive data such as another user's order history or address is included in the vector embeddings, there is a risk of data leaking.

Additionally, context can be given with a particular query. The context window is the total number of tokens that can be used with a prompt, and this window continues to grow. As users begin to upload large documents with their queries, such as sales figures and long PDF reports, the end-to-end round-trip for these uploads needs to be understood and secured. Data retention and encryption policies must also be enforced. There is a risk of mixing contexts and queries if not implemented correctly, so clear ring-fencing of context and auditing of applied context is critical.

Self Learning

As people rush to try out the latest public GPT-powered tools, very few are taking the time to read the terms and conditions of their use. For example, ChatGPT's free version states that it "...improves by further training on the conversations people have with it, unless you choose to disable training."

This creates a risk that sensitive data given by users in prompts may be used in training as people start to have conversations around their work. Even users who are aware of this risk can make mistakes as the conversation becomes as natural as talking to a co-worker, and their guard can slip. ChatGPT is currently working on a "ChatGPT Business" version of their immensely popular application that will not use end users' data to train models by default.

Organizations need to train their staff to understand these risks. While some have taken a heavy-handed approach to blocking these tools on their enterprise firewalls, it is clear that GPT is here to stay and will be embedded in many applications. Organizations need to learn how to work with these tools, but that of course needs to happen in a safe environment. In addition, making any use of user prompts for self learning clear to users for any applications organisations build or publish is a key aspect of responsible AI usage.

Additional Risks

While the previous sections cover the clear IP and data security risks associated with widespread use of GenAI, there are also existing AI risks and emergent attack vectors to consider. One such risk is Prompt Injection attacks, where a prompt is crafted to bypass safeguards or data protections, enabling behavior not desired by the application or model creator. For public applications, these risks are typically minor. However, as applications are trained on more sensitive data, the risks become more significant.

While not unique to GenAI, a fundamental risk when leveraging LLMs is the implicit bias of the data used. GenAI has great power to produce content, creating the potential to amplify biases implicit or directly contained in training data. As discussed, FMs require a high amount of training data, which may include data scraped from the internet, risking the inclusion of damaging viewpoints, content, and fake news. As FMs

begin to be used in applications, government, and business strategy, it is essential to have transparency on the provenance of the data used, as well as safeguards (e.g., via prompt engineering).

Conclusion

In conclusion, while GenAI has the potential to revolutionize many aspects of our lives, it also presents significant risks that must be carefully managed. These risks include IP and data security concerns, the potential for malicious use, and implicit bias. It's essential to understand and mitigate these risks throughout the entire supply chain of GenAI applications to avoid legal, societal, and reputational risks. Transparency & accountability across the entire supply chain are key focuses in achieving this goal.

Ben Ellerby

Ben Ellerby is the Founder of **aleios** and a lead organiser of the GenAI Days. In 2020, AWS named him a Serverless Hero for his community and open source work. He is the editor of Serverless Transformation publication. Ben co-organizes the Serverless User Group in London, is part of the ServerlessDays organizing team, and regularly speaks at technology conferences around the world. At **aleios and Theodo Group**, he helps startups disrupt and assists large organizations in remaining competitive by leveraging cutting edge technology to solve real world problems. He advises several open-source Generative AI projects, as well as companies and NGOs.

aleios

With help from Aleios, Theodo & SICARA