Generative AI Guardrails

What is Generative AI?

Generative AI tools are a class of artificial intelligence tools that are designed to create content, including but not limited to text, images, audio, video, and even software code. An example of these is ChatGPT, which is an AI-based large language model (LLM) designed to be more conversational than traditional AI tools. It is trained on enormous amounts of publicly available information, which allows it to provide outputs that closely resemble how humans interact.

There are various potential benefits associated with using these tools. They excel at analyzing and understanding existing content to create new content, such as summaries, lesson plans, basic policies, software code, and acting as chat bots and personal assistants. For example, a legal professional could use a generative AI tool to draft and/or interpret contracts or filings, while a marketing professional could use one to draft the initial text and images for a new campaign.

Issues Associated with Generative AI

There are several issues and concerns associated with the use of generative AI tools that you should be aware of before using them[1]:

1. **Fabricated and inaccurate answers.** ChatGPT in particular is prone to "hallucinations," where it makes up facts, non-existent court decisions, non-existent URLs, etc. When ChatGPT is challenged on the accuracy of these facts, sometimes it corrects them, while other times, it insists that they are accurate. As OpenAI itself notes, "ChatGPT sometimes writes plausible sounding but incorrect or nonsensical answers.[2]"
2. **Data privacy and confidentiality.** Any information entered into a generative AI tool may become part of its training dataset.
3. **Model and output bias.** Sometimes this is because the content used to train the model is biased; other times, biased results can be the output of how the model interprets data and how it itself is configured.
4. **IP and copyright risks.** Content generated from LLMs may be subject to the third-party intellectual property rights of the data used to train the LLM. Software code generated by LLMs, for example, may infringe the copyright of the author that developed the code that trained the LLM. Separately, code developed using generative AI could also be subject to disclosure requirements if based on open-source software. Courts are also starting to

[1] https://www.gartner.com/en/newsroom/press-releases/2023-05-18-gartner-identifies-six-chatgpt-risks-legal-and-compliance-must-evaluate
[2] https://openai.com/blog/chatgpt

review how transformative a work must be for it not to be simply derivative. There is also some concern as to whether AI-generated content can be copyrighted at all or whether the LLM adequately transfers its rights, if any, to the user.

5. **Cyber fraud risks.** This includes things like prompt injection attacks, where malicious prompts can be entered to write malware or trick an AI-based personal assistant into taking damaging actions. Be extra vigilant in spotting Deep Fakes in social engineering attacks. Voice and video deepfakes are being utilized to perpetuate unauthorized access and fraudulent activity.
6. **Consumer protection risks.** These relate to the need to disclose the use of AI, such as chatbots, to consumers.
7. **LLM can be limited by the data it is trained on.** For the open-source version of ChatGPT in particular, its LLM does not have anything in it after September 2021, so it either won't know about later developments, or it will make things up.

- **Authenticity.** Content developed by generative AI tools can be so realistic that it is difficult to determine its accuracy – for example, many companies are targeted for phishing schemes involving senior management needing emergency money wired to them. These are generally easy to confirm/refute; it will be significantly harder if a scammer calls the company using the CEO's AI-cloned voice.
- **Poor quality results.** Content generated by generative AI is often generic and abstract. The quality can be improved by effective use of prompts and careful editing, but both require substantial human involvement.
- **Model limitations.** It is well-documented that many generative AI tools have substantial issues with simple math, e.g., how many fingers are on a standard human hand.

Guidance for Using Generative AI

Generative AI tools can be used as a resource, but users are responsible for the AI-generated content that they rely on, and **care should be taken to protect the Company from liability and reputational risk.**

- Do not blindly rely on anything created by generative AI. Ensure a human subject matter expert reviews the AI contribution. This is especially important for code, court cases, URLs, and other factual assertions.[3] You are responsible for the errors introduced by generative AI.
- Unless you have licensed a generative AI tool that ensures confidentiality, assume that everything input into a generative AI tool will be used to train that tool and potentially used in a response to another user. Treat any information you post into

[3] See, e.g., https://www.businessinsider.com/lawyer-apologizes-for-using-chatgpt-for-an-affidavit-2023-5

these tools as if you were posting it on a public social media site. Do not use Company sensitive or proprietary information with generative AI tools – that is, do not copy Company or personal information into generative AI tools.

- Similarly, do not use Company intellectual property with generative AI tools. Company intellectual property includes, but is not limited to, custom software, algorithms, etc.
- Discount Tire software development teams should adhere to the OWASP AI Security & Privacy guide when designing, creating, or testing in-house developed AI-enabled applications.
- Online generative AI tools should not be accessed without "Oversharing" cybersecurity tools to prevent data leakage and provide network auditing tools.

Conclusion

The bottom line is that generative AI tools are just tools like any other, and ultimately Discount Tire owns and is responsible for any business content created using these tools. Generative AI tools are definitely inviting and can be easy; we need to make sure we use them safely as well.

If You Have Questions:

Contact the ILM team at ilm@discounttire.com.