Delivering safety in GenAI : our top concerns in 2024

globalapptesting.com
marketing@globalapptesting.com

1

JAILBREAKING

What is jailbreaking?

Achieving unauthorized abilities in an AI system by crafting prompts that bypass or break the intended constraints and filters; “jail breaking” the AI out from its original behaviour.

Consequences of jailbreaking

- “Screenshot attacks” – broken bots get screenshotted and shared on social media in brand and reputation attacks
- Legal & financial consequences – courts have upheld that AI bots can speak on behalf of the company when they have offered discounts.

Tools and best practices to prevent jailbreaking

- **Guardrails** – software by Guardrails AI – which helps protect from malicious user input as well as catching potentially embarrassing output generated by gen AI bot that could hurt us.
- **Moderation API** – via OpenAI – an API which identifies which inputs and outputs violate an AI agents’ content rules, to help third-party developers test the system.
- **Leveraging jailbreaking repositories** – keep on top of the techniques that hackers and bad security agents will use by looking at repositories where up-to-date prompt techniques are tried and scanned. You can also try this game to have a go hacking an LLM in a safe environment.
- **Self-reminders** – a simple technique which you can use for free – after each user prompt, add “reminder” to the gen AI model explaining what is their purpose and they shouldn’t trust fully the user who may try to hack them.

2

PROMPT INJECTION

What is prompt injection?

Prompt injection relates to the term “SQL injection”. It’s a means of using the prompt input box to put a malicious text, that when connected with the original prompt, may lead to the system changing its behaviour.

Consequences of prompt injection

- Data leak - attackers may obtain your original prompt or documents used to train model through RAG - all of these can contain confidential corporate information.

Tools and best practices to prevent prompt injection

- **Good input and output filters** – Manually filter input and output from undesired content, or detect anomalies to prevent dangerous prompts from execution.
- **External LLM evaluation** – Use a dedicated, separated LLM which scans the user prompt and evaluate if it’s malicious or not.
- **The “sandwich defense”** – It’s a technique of wrapping user input in pre-input and post-input and ideally the middle part so it will be easier to separate it from the rest, reducing changes it will be interpreted as a command.
- **Human-in-the-loop** – allow users to flag problematic outputs for review and route cases which require judgement to humans; or escalate very sensitive queries to humans with prompt anomaly detection.

Delivering quality in GenAI : our top concerns in 2024

globalapptesting.com
marketing@globalapptesting.com

3

HALLUCINATION

What is hallucination?

False but credible content generated by AI. GenAI can retrieve information but it can also look to replicate patterns in text, images, or numbers. Therefore there is a risk for some GenAI engines that they will "say something reasonable" – i.e. they will say something which is a plausible answer, but is not correct. That may include facts from the world but even mathematical equations.

Consequences of hallucination in GenAI

- Loss of trust – the consequences of ending clients false information are more far-reaching, but nearly always engender a loss of trust.

Tools and best practices your team can use

- **Retrieval-augmented-generation (RAG)** – giving your GenAI a database or filebase of approved facts to "retrieve" information from, particularly where it refers to sensitive subjects, can help ground your AI output in facts.
- **Chain of thoughts** – requiring your AI to show a "chain-of-thought"; (a step-by-step reasoning process in a human readable format) proves to provide more accurate results.
- **Prompt engineering training** – better prompts generate better responses and you can encourage your users and employees to prompt more effectively by teaching them. There are free, open source knowledge bases available such as this one teaching good practices to increase prompt accuracy.
- **Human-in-the-loop –** again, identifying very sensitive or important cases and escalating them to a human supervisor could help to avoid hallucinations where it really counts.

4

LACK OF DETERMINISM

What do we mean by non-deterministic?

AI is non-deterministic by design, meaning that most of its responses are unique. We've seen this before – you may know that 15% of all Google searches have never been seen before – but a text AI can generate more and more varied answers than a page rank system. In many situations, this is a desired behaviour, but in others, it may not be welcomed, especially if surprises or different output may cause the system to be less reliable.

Consequences of non-determinism in GenAI

- A more statistical approach to quality – testing in a range of environments can give a clear picture, but no test methodology can give an absolute clean bill of health
- Mission-critical systems should not require GenAI in their workflow. The temptation to use GenAI in mission-critical systems might be a source of total system failure in the medium term.

Tools and best practices your team can use

- **Reduce ambiguity** – Every ambiguity you introduce may be multiplied by gen AI randomness and non-deterministic design. Avoid situations like "possibly" empty or null inputs, too many options to pick from or open outputs, if the model output serves as input for another part of the system.
- **External LLM evaluation** – Use a dedicated, separated LLM which scans the user prompt and evaluate if it's malicious or not.
- **Consider a reduced temperature –** "Temperature" is one of the model configuration options which may help you reduce bot "creativeness" and randomness. Depending on your use case, this may be desired or not.

globalapptesting.com
marketing@globalapptesting.com

Why are model updates important?

Model updates are the moment where regression tests are the most appropriate. Whether you're reliant on model updates from an underlying solution or you've build the solution yourself from scratch or you're looking to benchmark your solution with an alternative model vendor, it's good to test for the major issues before going ahead.

What your team should consider in a model update

- **PromptFoo** – Promptfoo is an automated test tool which you can use to create simple unit tests for your prompts accuracy.
- **LangChain evaluators** – A langchain evaluator is a component that assesses the quality or correctness of outputs generated by a language model within the Langchain framework. It provides a method to systematically evaluate an LLM's responses.
- **Manual testing with global app testing** – taking on a statistical approach for UX and quality means that you should run manual tests on pre-live GenAI engines with a range of people, countries, geographies, to ensure that your outcome is consistent with expectations everywhere that you're live. We can help with that.

Global App Testing

Identify local AI output issues with global crowdtesting

- **Assess offensive or biassed output** from specified user groups – crowd testing can help you via Global App Testing with testers in 190+ countries and territories.
- **Benchmark your GenAI output against competitors** with like-for-like prompt comparison across a range of inputs via our competitor benchmarking reports.

GenAI quality survey

Win $250 by filling in our survey on GenAI quality

Head to the survey

Thoughts on compliance in GenAI

Forthcoming compliance issues in GenAi

- The EU AI Act is already a law in place which governs AI The AI Act will focus on Model inventory, risk evaluation, and ensuring transparency.
- Many US Laws are on their way, including an AI Bill Of RIghts

Tools and best practices your team can use

- An AI governance and compliance checklist by OWASP (checklist starts on p14)
- A compliance checklist by AI Guardian
- EU compliance checklist by EU Artificial Intelligence Act

Talk to us at hello@globalapptesting.com