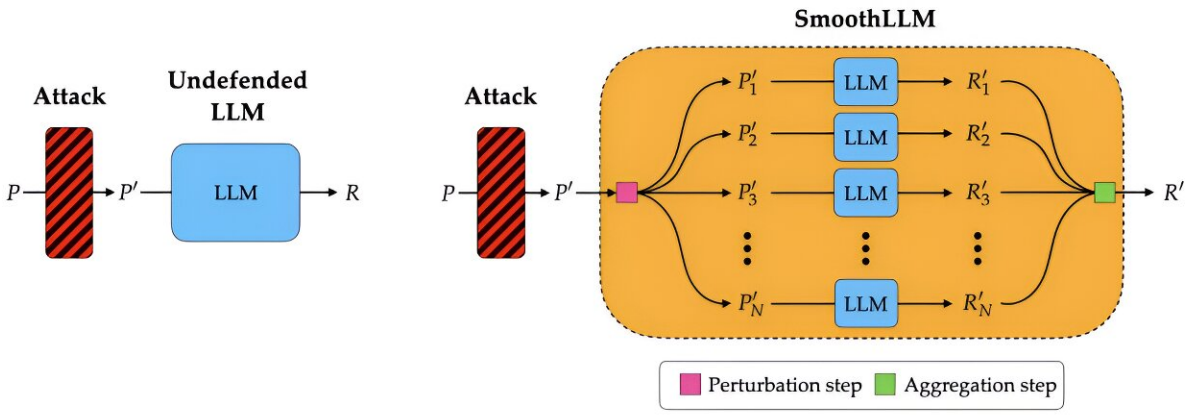
AI can write a wedding toast or summarize a paper, but what happens if it's asked to build a bomb?

November 30 2023, by Nathi Magubane



We introduce SmoothLLM, an algorithm designed to mitigate jailbreaking attacks on LLMs. (Left) An undefended LLM (shown in blue), which takes an attacked prompt P′ as input and returns a response R. (Right) SmoothLLM (shown in yellow) acts as a wrapper around any undefended LLM; our algorithm comprises a perturbation step (shown in pink), where we duplicate and perturb N copies of the input prompt P′, and an aggregation step (shown in green), where we aggregate the outputs returned after passing the perturbed copies into the LLM. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2310.03684

During the past year, large language models (LLMs) have become incredibly adept at generating synthesizing information and producing

humanlike outputs. LLMs are likened to digital librarians, as they have been trained on vast datasets sourced directly from the internet and can therefore generate or summarize text on nearly any topic. As a result, these LLMs have become ubiquitous in such fields as copywriting, software engineering, and entertainment.

However, the body of knowledge and capabilities in LLMs make them attractive targets for malicious actors, and they are highly susceptible to failure modes—often referred to as jailbreaks—that trick these models into generating biased, toxic, or objectionable content.

Jailbreaking an LLM is akin to fooling these digital librarians into revealing information they are programmed to withhold, such as instructions for how to build a bomb, defraud a charity, or reveal private credit card information.

This happens when users manipulate the model's input prompts to bypass ethical or safety guidelines, asking a question in a coded language that the librarian can't help but answer, revealing information it's supposed to keep private.

Alex Robey, a Ph.D. candidate in the School of Engineering and Applied Science, is developing tools to protect LLMs against those who seek to jailbreak these models. He shares insights from his latest research paper, posted to the *arXiv* preprint server, regarding this evolving field, with a particular emphasis on the challenges and solutions surrounding the robustness of LLMs against jailbreaking attacks.

Bad actors co-opting AI

Robey emphasizes the rapid growth and widespread deployment of LLMs in the last year, calling popular LLMs like OPenAI's ChatGPT "one of the most prevalent AI technologies available."

This explosion in popularity has been likened to the advent of the internet, and underscores the transformative nature of LLMs, and the utility of these models spans a broad spectrum of applications into various aspects of daily life, he says. "But what would happen if I were to ask an LLM to help me hurt others? These are things that LLMs are programmed not to do, but people are finding ways jailbreak LLMs."

One example of a jailbreak is the addition of specially chosen characters to an input prompt that results in an LLM generating objectionable text. This is known as a suffix-based attack. Robey explains that, while prompts requesting toxic content are generally blocked by the safety filters implemented on LLMs, adding these kinds of suffixes, which are generally nonsensical bits of text, often bypass these safety guardrails.

"This jail break has received widespread publicity due to its ability to elicit objectionable content from popular LLMs like ChatGPT and Bard," Robey says. "And since its release several months ago, no algorithm has been shown to mitigate the threat this jailbreak poses."

Robey's research lies addresses these vulnerabilities. The proposed defense, which he calls SmoothLLM, involves duplicating and subtly perturbing input prompts to an LLM, with the goal of disrupting the suffix-based attack mechanism. Robey says, "If my prompt is 200 characters long and I change 10 characters, as a human it still retains its semantic content."

While conceptually simple, this method has proven remarkably effective. "For every LLM that we considered, this success rate of the attack dropped below 1% when defended by SmoothLLM," Robey says. "Think of SmoothLLM as a security protocol that scrutinizes each request made to the LLM. It checks for any signs of manipulation or trickery in the input prompts. This is like having a security guard who double-checks each question for hidden meanings before allowing it to

answer."

Aside from mitigating suffix-based jail breaks, Robey explains that one of the most significant challenges in the field of AI safety is monitoring various trade-offs. "Balancing efficiency with robustness is something we need to be mindful of," he says. "We don't want to overengineer a solution that's overly complicated because that will result in significant monetary, computational, and energy-related costs. One key choice in the design of SmoothLLM was to maintain high query efficiency, meaning that our algorithm only uses a few low-cost queries to the LLM to detect potential jail breaks."

Future directions in AI safety

Looking ahead, Robey emphasizes the importance of AI safety and the ongoing battle against new forms of jailbreaking. "There are many other jailbreaks that have been proposed more recently. For instance, attacks that use social engineering—rather than suffix-based attacks—to convince a language model to output objectionable content are of notable concern," he says. "This evolving threat landscape necessitates continuous refinement and adaptation of defense strategies."

Robey also speaks to the broader implications of AI safety, stressing the need for comprehensive policies and practices. "Ensuring the safe deployment of AI technologies is crucial," he says. "We need to develop policies and practices that address the continually evolving space of threats to LLMs."

Drawing an analogy with evolutionary biology, Robey views adversarial attacks as critical to the development of more robust AI systems. "Just like organisms adapt to environmental pressures, AI systems can evolve to resist adversarial attacks," he says. By embracing this evolutionary approach, Robey's work will contribute to the development of AI

systems that are not only resistant to current threats but are also adaptable to future challenges.

More information: Alexander Robey et al, SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks, *arXiv* (2023). DOI: 10.48550/arxiv.2310.03684

Provided by University of Pennsylvania