LLM SECURITY CONCERNS SHINE A LIGHT ON EXISTING DATA VULNERABILITIES

Authors

Michael Papadopoulos, Nicholas Johnson,
Michael Eiden, Philippe Monnot,
Foivos Christoulakis, and Greg Smith

In the rapidly evolving landscape of artificial intelligence (AI), large language models (LLMs) have emerged as a powerful tool, capable of generating human-like text responses, creating conversational interactions, and transforming the way we perceive and interact with technology.

Like all powerful tools, LLMs come with a set of security concerns. This article delves into those concerns, emphasizing that although LLMs certainly present novel security threats, the fundamental concerns, protections, and remedies remain similar to existing, well-understood information security challenges. In fact, characteristics of LLMs and their associated data pipelines allow more sophisticated and proportional security interventions, potentially leading to a better equilibrium between protection and benefit.

The first point to understand is that LLMs, by their nature, can only divulge information they were exposed to during their training phase. Thus, if an LLM reveals sensitive or private information, it's not because the model is inherently insecure — it's because it was given access to this information during its training. This highlights that the root of the problem is improper data access and management. Consequently, the focus should be on ensuring that data used to train these models is carefully curated and managed in order to prevent any potential downstream data leaks.

However, managing the training data is just one part of the equation. Even with the best data management practices, an LLM might still generate inappropriate or harmful content based on the patterns it learned during training. This is where the implementation of an LLM module, coupled with strategic prompt engineering, can serve as a robust, layered security mechanism.

Prompt engineering involves carefully crafting the prompts that are given to the LLM to guide it toward generating the desired output. By scrutinizing both the inputs (user prompts) and the outputs of the LLM, we can establish a multitiered safety environment that can effectively mitigate security risks. For instance, an LLM module can be designed to reject certain types of prompts that are likely to lead to harmful outputs, and it can be programmed to filter out any potentially harmful content from the LLM's responses.

MANAGING THE TRAINING DATA IS JUST ONE PART OF THE EQUATION

This approach to security doesn't just protect against the known risks associated with LLMs, it provides a framework for identifying and mitigating new risks as they emerge. It's a dynamic, adaptable approach that can evolve alongside the LLMs. Indeed, the pace of innovation within the LLM and wider language-processing domain ensures that any security approach not based on continuous sensing, analyzing, adapting, and iterating is doomed to failure.

It's important to be aware of the security concerns associated with LLMs, but it's equally important to understand that these concerns are new manifestations of existing security threats and thus manageable. With proper data handling and innovative security strategies, we can harness the full potential of these powerful AI tools without compromising safety or security.

TOP 10 LLM SECURITY CONCERNS

The exponential integration of LLMs within organizations holds the promise of seamless automation and enhanced efficiencies. However, with these advancements come unique security challenges.

Our research and use in the field have yielded a top 10 list of vulnerabilities that pose either new threat vectors or new context for typical vulnerabilities to be exploited or manipulated in an LLM context (see Figure 1):

1. Prompt injection
2. Insecure output handling
3. Training data poisoning
4. Model denial of service
5. Supply chain vulnerabilities
6. Data leakage/sensitive information disclosure
7. Insecure plug-in design
8. Excessive agency
9. Overreliance
10. Model theft

PROMPT INJECTION

The age-old tactic of manipulating systems through cunning inputs finds its way to LLMs. Attackers craftily modify the prompts fed into the model, leading to unintended actions. There are two primary avenues for these attacks: (1) direct injections involve overriding the system prompts and (2) indirect ones alter the inputs from external sources. These can compromise the integrity of the LLM's response and, subsequently, the systems relying on it.

To remediate prompt injection attacks, users must validate and sanitize all inputs before they're processed. Simultaneously, they should maintain a white list of accepted commands to aid in filtering out malicious inputs. Regular monitoring and logging of prompts become vital to detect and address unusual patterns swiftly, and it's beneficial to limit the amount of user-defined input that an LLM can process. Finally, introducing a system of regular user feedback can help fine-tune the model's responsiveness to malicious prompts.

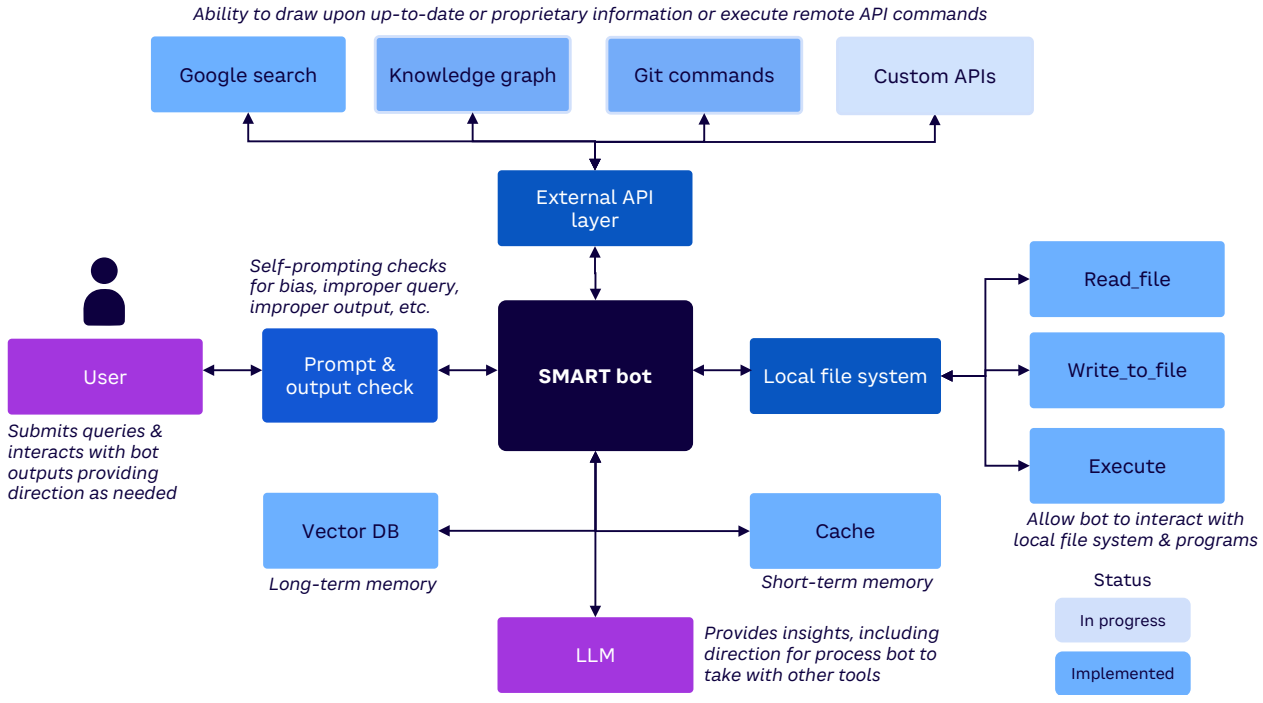

Figure 1. LLM application architecture: remediation of most common vulnerability issues (source: Arthur D. Little)

INSECURE OUTPUT HANDLING

LLMs can produce a wide variety of outputs. Accepting these without proper verification opens the gates for multiple threats, including XSS (cross-site scripting), CSRF (cross-site request forgery), and SSRF (server-side request forgery). Moreover, privilege escalation or remote code execution becomes feasible, posing an enormous security risk to the back-end systems that treat the output as safe.

Proactive measures like output sanitization, strict validation, and monitoring should be established to prevent privilege escalation and remote code execution. This will ensure the consistent security of responses generated by LLMs.

TRAINING DATA POISONING

Training data is the backbone of any LLM. However, when this data is compromised or injected with malicious intent, the resultant LLM can exhibit vulnerabilities or biases. This can weaken the model's security, overall effectiveness, and even ethical behavior.

To remediate training data–poisoning attacks in LLMs, it's crucial to prioritize the integrity of the training data by sourcing it exclusively from reputable sources and meticulously validating its quality. Applying rigorous data-sanitization and preprocessing techniques is essential to weed out potential vulnerabilities or biases inherent in the data. It's also beneficial to conduct periodic reviews and audits of the LLM's training data and its fine-tuning processes. Finally, incorporating monitoring and alerting systems can be invaluable in identifying any unusual behavior or performance anomalies, further bolstering the model's security.

MODEL DENIAL OF SERVICE

The resource-intensive nature of LLMs makes them susceptible to denial-of-service attacks. Perpetrators can introduce resource-heavy operations, overburdening an LLM, causing either service degradation or unexpectedly high operational costs.

To counteract these attacks, it's essential to implement rate-limiting measures and monitor user inputs for resource-heavy operations. By managing the workload and detecting unusual spikes in resource usage, organizations can maintain optimal LLM performance and prevent excessive operational costs.

SUPPLY CHAIN VULNERABILITIES

The lifecycle of LLM applications involves data sets, pretrained models, plug-ins, and more. Introducing vulnerabilities at any of these stages can compromise the entire model, making it an attractive target for attackers. To secure the LLM application lifecycle, conduct regular audits of all components. Employing stringent validation and vetting processes during integration will safeguard the model, reducing its susceptibility to external threats.

TRAINING DATA IS THE BACKBONE OF ANY LLM

DATA LEAKAGE/SENSITIVE INFORMATION DISCLOSURE

LLMs, while sophisticated, may unintentionally leak confidential information through their responses. This can lead to unauthorized data access, breaches, and severe privacy violations. Organizations must stress data sanitization and user policies to circumvent such exposures. We have found that using a secondary LLM to test the outputs for sensitive information is an excellent way to help ensure security.

INSECURE PLUG-IN DESIGN

LLMs often incorporate plug-ins to enhance functionality. However, if these plug-ins have insecure input mechanisms or flawed access controls, they become glaring vulnerabilities. Exploiting them might result in grave consequences, including remote code execution.

To mitigate vulnerabilities in LLM plug-ins, ensure rigorous vetting before integration. Prioritize plug-ins with robust input validation and stringent access controls. Regular security audits of plug-ins can also help detect and rectify potential weak points, preventing potential exploits.

EXCESSIVE AGENCY

Assigning excessive permissions, functionality, or autonomy to LLMs can spell disaster. Such models can autonomously make decisions, potentially leading to significant unintended consequences. This issue emphasizes the need for setting boundaries for LLM-based systems.

To safeguard against overpowered LLMs, it's imperative to implement a permissions framework, limiting the LLM's functionality and autonomy. Regularly review and adjust these permissions to strike a balance between operational efficiency and control to ensure LLMs function within defined boundaries.

OVERRELIANCE

Reliance on LLMs without human oversight is a treacherous path. Such "blind reliance" can lead to misinformation, legal conundrums, and a host of security vulnerabilities, mainly if the LLM churns out incorrect or inappropriate content.

To counteract this risk, introduce human oversight in critical decision-making processes. Establishing a hybrid system, in which human experts review and validate LLM outputs, can reduce misinformation risks, address potential legal issues, and bolster overall security against inappropriate content generation.

MODEL THEFT

Proprietary LLMs are of immense value. Unauthorized access or exfiltration can cause substantial economic losses, erode competitive advantages, and even expose sensitive information. Ensuring stringent security protocols is paramount to prevent such incidents.

To protect proprietary LLMs, deploy multilayered security measures, including encryption, access controls, and regular audits. By closely monitoring system activity and restricting unauthorized access, organizations can safeguard their valuable assets, preserving both competitive advantage and data confidentiality.

The era of LLMs is transformative, heralding countless possibilities. However, navigating this landscape requires organizations to be acutely aware of the inherent security challenges. Addressing these concerns head-on will ensure a future where LLMs can be harnessed safely and efficiently.

HOW LLMs CAN IMPROVE SECURITY

Rather than introducing wholly unprecedented threats into society, LLMs highlight and stress test existing vulnerabilities in how organizations govern data, manage access, and configure systems. With care and responsibility, we can respond to their revelations by engineering solutions that make technology usage more secure and ethical overall.

Specific ways responsible LLM adoption can improve security include:

- **Red team penetration testing.** Use LLMs to model criminal hacking and fraud to harden defenses.
- **Automated vulnerability scanning.** Leverage LLM conversational ability to identify flaws in public-facing chat interfaces.
- **Anomaly detection.** Monitor corporate system logs with LLMs fine-tuned to flag unusual internal events as possible attacks.
- **Safety analysis.** Stress test new features through automated conversational exploration of potential abuses.
- **Product-security reviews.** Use LLMs as a team member when designing new products to probe attack possibilities in simulated conversations.
- **Threat intelligence.** Continuously train LLMs on emerging attack data to profile bad actors and model potential techniques.
- **Forensic reconstruction.** Assist investigations of past incidents by using LLMs to speculate about criminal conversations and motives.
- **Security policy analysis.** Check that policies adequately address LLM-relevant risks revealed through conversational probing.
- **Security training.** Use LLM-generated attack scenarios and incidents to build staff defensive skills.
- **Bug bounties.** Expand scope of bounty programs to include misuse cases identified through simulated LLM hacking.

With careful design and effective oversight, LLMs can be an ally rather than a liability in securing organizations against modern technological threats. Their partially open nature invites probing for weaknesses in a controlled setting.

> **THE ERA OF LLMs IS TRANSFORMATIVE, HERALDING COUNTLESS POSSIBILITIES**

LLMs present a further opportunity to improve an organization's information security capability. The practical application of LLMs to business challenges requires creating sophisticated, multistage, software-driven data pipelines. As these pipelines start to become prevalent, an opportunity to design with more effective security protocols is presented.

Various security postures can be applied at different points in the pipeline. For instance, a permissive security posture that allows an LLM to generate the best possible response can be followed by a more restrictive security filter that automatically checks the output for potential data leakage.

If we accept that LLM security problems are new manifestations of existing information security challenges (and that human behavior is the biggest cause of security breaches), then automated multistage processes with carefully constructed security gateways can provide a powerful new tool in the toolkit.

CONCLUSION

LLMs, such as GPT-4, represent a breakthrough in language-capable AI, but commentary casting their risks as wholly unprecedented is overstated. A closer look reveals that concerns around their potential for data exposure and security issues/bias largely echo existing vulnerabilities, often exacerbated by poor underlying security and data governance practices.

Rather than engaging in an ultimately futile battle to ban promising AI innovations, the responsible path is to address underlying root causes. The route to achieving this is well understood but often poorly implemented, requiring organizations to take a systematic and pragmatic approach to security, including better aligning access controls, tightening monitoring, enhancing information literacy, and ensuring effective oversight. LLMs can even assist in this by stress testing systems and uncovering policy gaps through exploratory conversation.

The emerging technology does not intrinsically undermine safety — it shines a light on long-standing cracks that ought to be sealed and has the potential to enhance security.

About the authors

Michael Papadopoulos is a Cutter Expert, Chief Architect of Arthur D. Little's (ADL's) UK Digital Problem Solving practice, and a member of ADL's AMP open consulting network. He is passionate about designing the right solutions using smart-stitching approaches, even when elegance and architectural purity are overshadowed by practicality. Mr. Papadopoulos leads the scaling of multidisciplinary organizations by focusing on continuous improvement, establishing quality standards, and following solid software engineering practices. He mentors team members, leaders, and managers along the way. Mr. Papadopoulos is a strong advocate of the DevOps culture and Agile principles and has demonstrated experience in solving problems in challenging global environments. Coming from a development background, he remains highly technical, with hands-on involvement in code review, design, architecture, and operations. Mr. Papadopoulos has more than 15 years' experience in technology and digital consulting and has worked in a variety of sectors, including telecom, gaming, energy, and media. He can be reached at experts@cutter.com.

Nicholas Johnson is a Partner with ADL's Digital Problem Solving practice, based in London. He focuses on how emerging digital technologies can be harnessed to drive transformation of both the business and its internal technology function. Mr. Johnson believes that the technology patterns/approaches used in the past are no longer appropriate to today's challenges, and that businesses must adopt new approaches. With over 20 years in technology consulting, he has worked across a wide range of sectors, including telco, media, transport, logistics, gaming, and energy, having considerable experience delivering global digital transformation initiatives, often from initial idea conception through full product launch. Mr. Johnson also has extensive experience in managing large-scale Agile projects and proof of concepts. He earned a bachelor of science degree from Cardiff University, UK, and a master of science degree in computer science, with distinction, from the University of Bath, UK. He can be reached at experts@cutter.com.

Michael Eiden is a Cutter Expert, Partner and Global Head of AI & ML at ADL, and a member of ADL's AMP open consulting network. Dr. Eiden is an expert in machine learning (ML) and artificial intelligence (AI) with more than 15 years' experience across different industrial sectors. He has designed, implemented, and productionized ML/AI solutions for applications in medical diagnostics, pharma, biodefense, and consumer electronics. Dr. Eiden brings along deep expertise in applying supervised, unsupervised, as well as reinforcement ML methodologies to a very diverse set of complex problem types. He has worked in various global technology hubs, such as Heidelberg (Germany), Cambridge (UK), and Silicon Valley (US), with clients ranging from small and medium-sized enterprises to globally active organizations. Dr. Eiden earned a doctorate in bioinformatics. He can be reached at experts@cutter.com.

Philippe Monnot is a Data Scientist with ADL's Digital Problem Solving practice, and a member of ADL's AMP open consulting network. He's passionate about solving complex challenges that impact people's livelihood through the use of data, statistics, and ML. Mr. Monnot enjoys developing accessible solutions that customers will adopt through effective data storytelling and explainable AI. Before joining ADL, he worked in R&D, where he used ML to implement smart, scalable manufacturing processes to manufacture sustainable composite structures for the aerospace and oil and gas industries. He can be reached at experts@cutter.com.

Foivos Christoulakis is a Solutions Architect with ADL's Digital Problem Solving practice and a member of ADL's AMP open consulting network. He is a passionate cloud architect who has designed and implemented numerous solutions currently in production in global-scale organizations. Mr. Christoulakis helps organizations grow by focusing on high engineering standards and following solid software engineering practices. He continues to be a strong advocate of DevOps and Agile principles and showcases both skill sets mentoring and being a servant leader for multiple DevOps and architecture teams. Mr. Christoulakis has more than 10 years' experience in cloud architectures across many business verticals, including telecoms, entertainment, and software development. He can be reached at experts@cutter.com.

Greg Smith is Managing Partner of ADL. He founded and co-leads ADL's Digital Problem Solving practice and is a member of ADL's Executive Committee, where he has responsibility for ADL's global innovation strategy. His work focuses on business strategy in the context of digital transformation as well as the application of disruptive information technologies in solving intractable business problems in major enterprises. Recently, Mr. Smith has been focusing on digital operating models for established businesses, including the changes required in technology, culture, IT functions, business technology interactions, organizational design, and governance, and how these can combine to enhance customer and service experience. He holds a bachelor of science degree in biological sciences from the University of Leicester, UK, and finds that after 30 years of dormancy within his professional life, the underlying concepts of biology are becoming increasingly valuable at unlocking business problems and articulating solutions — especially where reductive, engineering-based approaches need to be replaced with whole-system, evolutionary thinking. He can be reached at experts@cutter.com.