Technology Innovation Institute

Zero-Trust Security, Resilience, and Safety in Foundational AI Models: A Comprehensive Framework for Trustworthy AI

Whitepaper

2025

tii.ae

Executive summary

Recent innovations in generative AI, particularly large language models, have demonstrated impressive utility across various sectors, including coding, summarization, assistants, AI agents, and autonomous systems. By 2025, generative AI models will power over 50% of enterprise applications[1]

However, their large size, opaqueness, and tendency to hallucinate raise significant concerns regarding new security vulnerabilities, resilience, and safety. This white paper explores how the application of zero-trust principles, established for security, could be a beacon of hope—improving the safety and resilience of Large Language Models (LLMs) and supporting more trustworthy AI.

Introduction

Researchers have studied machine-learning techniques for decades in their long-term effort to develop ever-more-capable AI systems. Though they made great strides, the most-capable AIs continue to require multiple, combined machine-learning processes and remained suitable only for structured data.

The discovery of the attention mechanism in 2017 [2] ushered in a new wave of innovation—allowing development of more capable large language models (LLMs) at scale. The attention mechanism is a key part of transformer neural networks. It works like a spotlight, focusing the most relevant parts of the input data, much as humans pay attention to key details in a conversation. These newer models can automatically discover relationships between things across unstructured data modalities, including text, code, sound, images, protein sequences, and machine data streams. Excitement about LLMs and other large-scale foundation models has driven trillions of dollars in investment into startups, infrastructure, research, and data management.

However, this progress comes with significant challenges. For example, LLMs and their supporting infrastructure have much larger attack surfaces than traditional AI models. They are big and complex, which means they can break in unexpected ways. They are also prone to hallucinations—nonsensical results spawned by plausible errors in the modeling process. As a result, these LLMs can pose significant risks of harm to people, infrastructure, and society—more than sufficient reason for concern and urgent research.

1Ball, S. (2023, December 18). LLM use-by data - who's keeping the data up to date? ERP Today. https://erp.today/llm-use-by-data/ Accessed: Mar. 11, 2025
2Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Introduction

Large AI vendors (like OpenAI, Anthropic, Cohere, DeepSeek and others) are making commendable progress on these issues. The Ais' inner workings are black boxes, however, which can make it difficult to probe them, secure their boundaries, recover from problems, and improve safety in new use cases. Over the last couple of years, several vendors have begun releasing more-open models. These offer more insight into the underlying system but may lack some of the proprietary models' security, guardrails, and governance.

In the long run, truly open-source approaches could make it easier to erect better safety guardrails, improve security processes, and make AI systems more resilient. Extending existing zero-trust security principles across the LLM lifecycle could help immensely.

It should be noted that trustworthy AIs, particularly LLMs, are complex. Researchers have mounted multiple efforts to develop frameworks for characterizing fundamental components.

For example, the International Telecommunications Union (ITU) has suggested protecting data by feeding these systems.[3] Deloitte has suggested another framework that privileges tools and processes for privacy, transparency, impartiality, responsibility, accountability, robustness & reliability, and safety & security.[4]

From a technical perspective, the US National Institute of Standards and Technology (NIST) suggests that the essential building blocks should include modules for validity & reliability, safety, security & resiliency, accountability & transparency, privacy, and fairness.[5] These are all good places to start as the industry considers better tools and processes for all of the components of more trustworthy LLMs and their supporting infrastructure. The US White House also recently published a National Security Memorandum to ensure that the ITU will lead international consensus-building and AI governance to develop safe, secure, and trustworthy systems. [6]

[3] "Trustworthy AI," Wikipedia. Jun. 01, 2024. Accessed: Mar. 11, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Trustworthy_AI
[4] "Trustworthy Artificial Intelligence (AI)™," Deloitte United States. Accessed: Mar. 11, 2025. [Online].
Available: https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html
[5] "Trustworthy and Responsible AI," NIST, Jul. 2022, Accessed: Mar. 11, 2025. [Online]. Available: https://www.nist.gov/trustworthy-and-responsible-ai
[6] White House, "FACT SHEET: Biden-Harris Administration Outlines Coordinated Approach to Harness Power of AI for U.S. National Security," The White House. Accessed: Mar. 11, 2025. [
Online]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2024/10/24/fact-sheet-biden-harris-administration-outlines-coordinated-approach-to-harness-power-of-ai-for-u-s-national-security/

Extending zero-trust security to safety and resilience

Zero-trust principles were initially conceived to protect cloud systems. Protecting current and future networks, data, and apps is at least as important as armoring the cloud. Ensuring this safety requires extending established trust boundaries to protect against many more kinds of harm. Further, ensuring that systems remain resilient requires stepping up system supervision, troubleshooting, and maintenance via continuous monitoring, more responsive infrastructure (LLMs as code), and continuous-integration, continuous deployment (CI/CD) processes.

Here are some examples of how the core principles could be extended to improve the security, safety, and resilience of LLMs and other foundation models:

1. Trust No One (Mutual Authentication): Using authentication mechanisms to guarantee safe input from trusted sources and between agents using secure tokens and certificates.

2. Fail Safely and Securely: Implement safeguards so that errors or hallucinations prompt the models to shut down safely without exposing sensitive data or causing harm.

3. Complete Mediation (Check Every Access): Validate inputs for queries and API calls to models, or between agents, to prevent data poisoning or prompt-injection attacks.

4. Rule of Least Privilege: Limit the permissions granted to individual models, or communications between multiple models, to prevent privilege creep.

5. Separation of Duty: Use distinct modules for input processing, model execution, and output, and find ways to separate complex tasks into specialized agents to minimize the impact of compromise of any one agent.

6. Least Common Mechanism: Isolate functions such as encryption and validation from LLMs and use dedicated agents for specialized tasks that can be secured independently.

7. Secure the Weakest Link: Identify vulnerable components for validating input, identifying relevance, or calling external APIs, and enhance these with additional checks or security protections.

8. Defense in Depth: Implement multiple layers of security, safety, and hallucination detection to protect models and limit their impacts on other systems or users.

9. Simplicity: Simplify architectures for deploying LLMs and managing interactions across multiple LLMs to reduce the risks of misconfiguration and of complex interactions that could create security or safety gaps or make it harder to troubleshoot issues in production.


Project Timeline
Research
Develop

Extending zero-trust security to safety and resilience

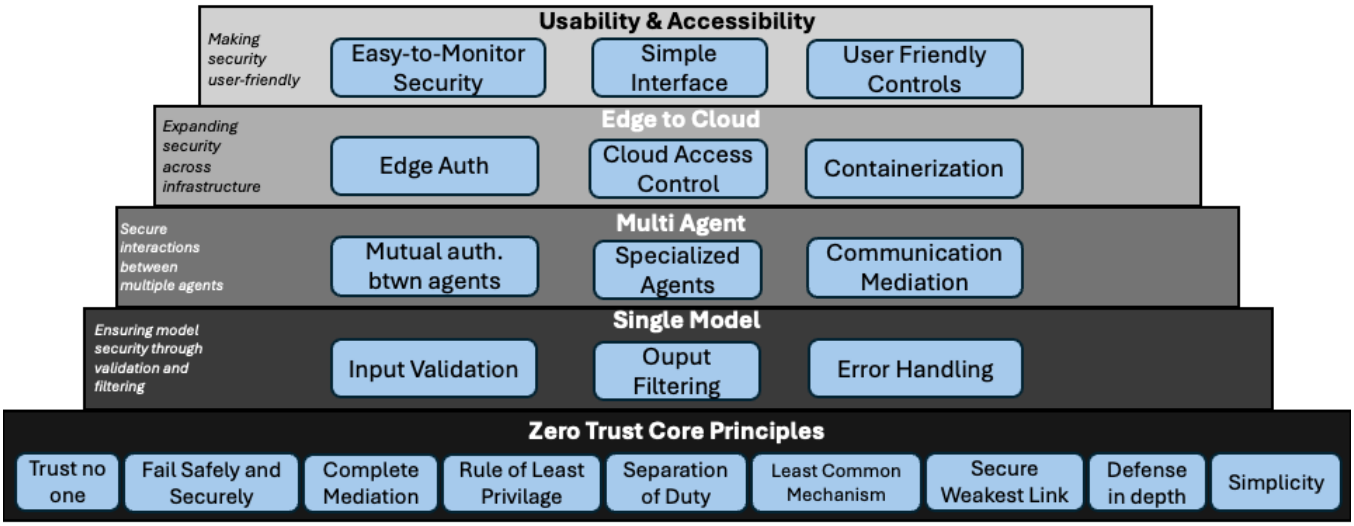


Figure 1: A Zero-trust Security Framework for AI Systems: This layered approach ensures security from foundational principles to usability, covering model validation, multi-agent interactions, edge-to-cloud security, and user-friendly controls. By enforcing zero-trust principles, this framework strengthens resilience against cyber threats while maintaining accessibility.

Recent progress in foundation models, particularly LLMs, has led to far-more-capable AI systems and introduced many new problems relating to their trustworthiness.[7] A comparison of traditional and foundational models (Table 1), shows that LLMs, in particular, have benefited from new processes for making sense of text, audio, video, or raw sensor data. LLMs have traditionally been trained on static data, which must be manually updated through an expensive retraining process. Today, however, innovations in reinforcement learning can enable these systems to learn after deployment, raising new privacy and data protection issues.

[7]Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223 1.2 (2023).

Extending zero-trust security to safety and resilience

Table 1: Comparison of Traditional AI Models and. Foundation Models: This table highlights key differences in architecture, training, scalability, flexibility, explainability, security, and deployment. While traditional AI models are task-specific and interpretable, foundation models like LLMs offer generalization and scalability at the cost of higher computational demands and security challenges.

Aspect	Traditional AI Models	Foundation Models (e.g., LLMs)
Architecture	- Typically use simpler architectures (e.g., decision trees, SVMs, basic neural networks). - Designed for specific tasks (e.g., classification, regression).	- Use advanced architectures like **transformers** with **attention mechanisms**. - Designed for **general-purpose tasks** (e.g., text generation, summarization).
Training Data	- Trained on **small, curated datasets** specific to a task. - Data is often **structured** (e.g., tabular data).	- Trained on **massive, diverse datasets** (e.g., text, code, images, audio). - Data is **unstructured** and often scraped from the web.
Training Process	- Requires **task-specific training**. - Training is **less computationally intensive**.	- Uses **self-supervised learning** on large datasets. - Training is **extremely resource-intensive** (e.g., requires GPUs/TPUs and large-scale infrastructure).
Scalability	- Limited scalability; models are **task-specific** and do not generalize well.	- Highly scalable; models can **generalize across tasks** and domains (e.g., GPT-4 can handle text, code, and images).
Flexibility	- **Inflexible**; models are designed for specific use cases and cannot adapt quickly.	- **Highly flexible**; can be fine-tuned or prompted for various tasks (e.g., translation, summarization, coding).

Extending zero-trust security to safety and resilience

Explainability	- Easier to interpret due to simpler architectures (e.g., decision trees).	- **Black-box nature** makes explainability challenging; it requires tools like XAI (Explainable AI).
Security Challenges	- Vulnerabilities are **localized** to specific tasks (e.g., adversarial attacks on image classifiers).	- **Broader attack surface** due to general-purpose nature (e.g., prompt injection, data poisoning). - **Hallucinations** can lead to misleading or harmful outputs.
Bias and Fairness	- Bias is **limited to the training data** for a specific task.	- Bias can be **amplified** due to large, diverse datasets (e.g., toxic content, stereotypes). - Requires **guardrails** to mitigate bias.
Deployment Complexity	- Easier to deploy and monitor due to smaller size and task-specific nature.	- **Complex deployment** due to model size and resource requirements. - Requires **continuous monitoring** for safety and security.
Use Cases	- Used for **specific applications** (e.g., fraud detection, sentiment analysis).	- Used for **general-purpose applications** (e.g., chatbots, coding assistants, autonomous systems).

The discovery of the attention mechanism using transformers in 2017 dramatically changed the development and capabilities of AI systems. Aspects of this include 1) new methods for discovering correlations in large unstructured data sets; 2) automating processes for learning from extremely large, unstructured data sets; 3) new methods for synthetic data generation; 4) innovations in user-experience design; and 5) multimodal approaches for correlating relationships across different types of data.

Extending zero-trust security to safety and resilience

Unstructured data revolution

Before recent innovations in transformers, generative adversarial networks (GANs), and diffusion models, considerable human effort was required to organize unstructured information like text, code, images, audio, and raw sensor feeds into formats suitable for AI and ML training. The new GenAI algorithms, however, can automatically capture essential correlations. For example, the seminal Vaswani, et al. paper on transformers suggested that "attention is all you need" to build a more competent translator. This allowed transformers to automatedly map words, entities, and concepts into vector embeddings directly, rather than the human hand-coding required for earlier approaches like Word2Vec[8] and GloVe[9]

Automating scale

Early work on GenAI algorithms focused on simple tasks like language translation or generating realistic-looking faces or strings of digits. OpenAI's seminal insight was that applying these new algorithms to extremely large data sets could lead to chatbots, coding assistants, and copilots.

Synthetic data generation

Existing GenAI models tend to be too large, too computationally intensive, and too slow to support large numbers of different tasks. Today, these applications play supporting roles in generating synthetic data sets or simulations that speed development of smaller and faster ML models in areas such as fault, malware, and intrusion detection to improve task performance, security, safety, and resilience.[10]

User experience

Large language models can also summarize complex information for different user audiences and expand simple prompts into appropriate commands applicable to a wider range of systems.
For example, we, the Technology Innovation Institute (TII), have been developing a natural language interface that allows humans to verbally create complex control programs for a swarm of robots, for such applications as monitoring the perimeter of an event.[11] Other research in the field explores how vision language models could improve UIs for drones, robots, and other autonomous systems.

[8]word2vec. (n.d.). TensorFlow. https://www.tensorflow.org/text/tutorials/word2vec, Accessed: Mar. 11, 2025
[9]Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
[10]Figueira, Alvaro, and Bruno Vaz. "Survey on synthetic data generation, evaluation methods and GANs." Mathematics 10.15 (2022): 2733.
[11]M. Andreoni, W. T. Lunardi, G. Lawton and S. Thakkar, "Enhancing Autonomous System Security and Resilience with Generative AI: A Comprehensive Survey," in IEEE Access, vol. 12, 2024).

Extending zero-trust security to safety and resilience

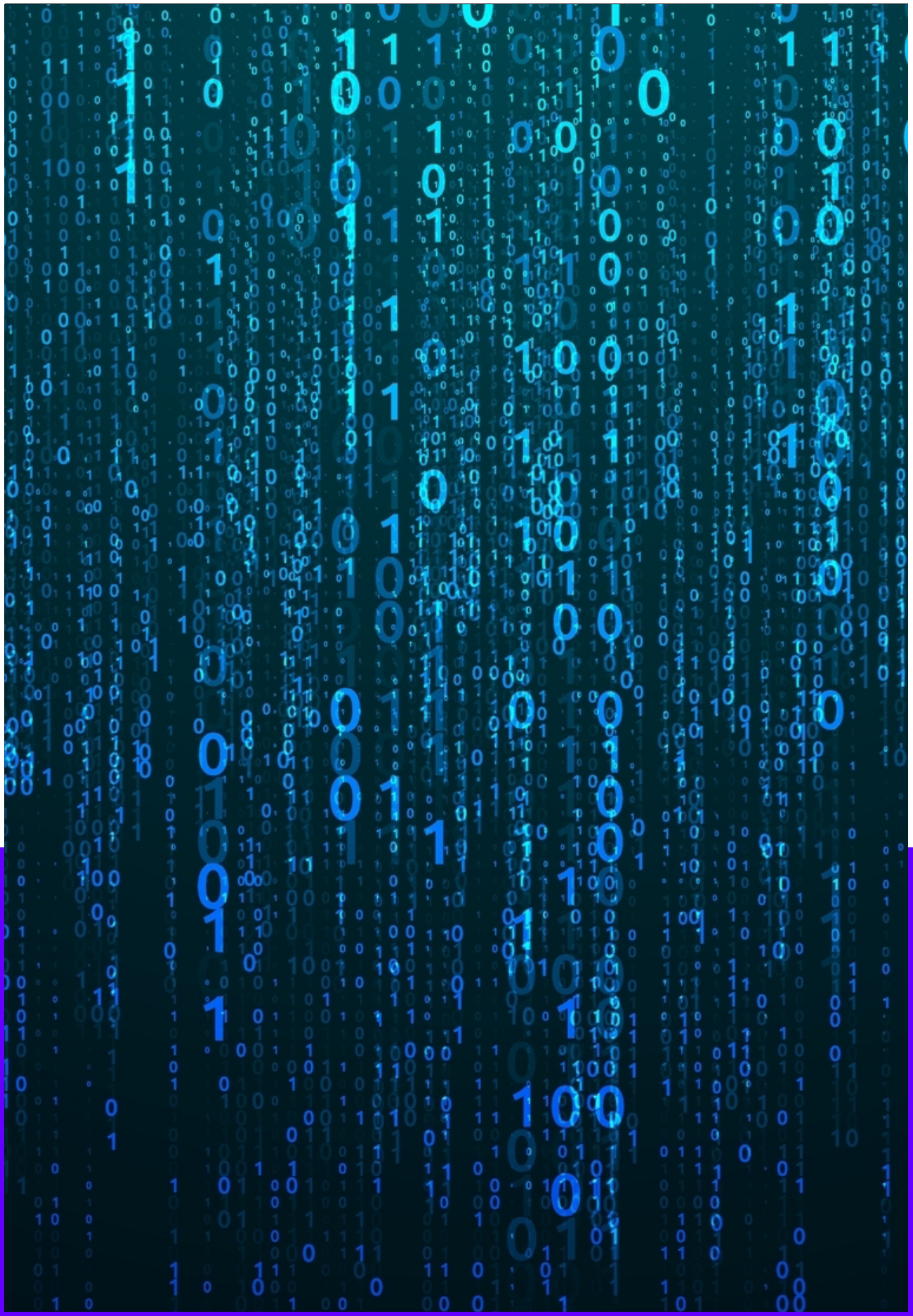
Improving precision and accuracy

AI hallucinations are a growing concern. LLMs tend to hallucinate confidently, particularly when assessing edge cases or describing things that are not well-represented in their training data. Some coping mechanisms for boosting accuracy and reducing hallucinations have included 1) fine-tuning LLMs for specific use cases; 2) priming LLMs with the most relevant subset of data using retrieval augmented generation (RAG);[12] 3) using GraphRAG[13] to prime them with a knowledge graph that represents the relationship between entities in the data; and 4) using special-purpose transformers or LLMs to decompose unstructured

Multimodal integration

The first generation of GenAI algorithms were trained on single data modalities—all text or all images, for example. Research has yielded valuable new approaches algorithm-training, such as transformers, that operate on multiple modalities of data such as text, audio, video, sensor data, or robot instructions.[14] Training a new language model from scratch requires considerable time and computing resources, so initial R&D focused on fusing new data modalities into existing LLMs. Recent progress has focused on combining modalities at training time, which can better represent correlations across modalities in the vector embedding space. For example, voice-chat assistants, notably OpenAI's GPT-4o, can learn the rhythms, cadence, and prosody of human speech rather than just the text words and their audio equivalent.

[12]Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." arXiv preprint arXiv:2312.10997 2 (2023).
[13]Han, Haoyu, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar et al. "Retrieval-augmented generation with graphs (graphrag)." arXiv preprint arXiv:2501.00309 (2024).
[14]Wu, Jiayang, et al. "Multimodal large language models: A survey." 2023 IEEE International Conference on Big Data (BigData). IEEE, 2023.



New Trust challenges

The substantial size and inherent complexities of Large Language Models (LLMs) and their supporting infrastructure introduce new challenges that complicate security, resilience, and safety. Trustworthy AI systems based on LLMs, especially autonomous systems that rely on these models, must address several critical issues. These include identifying and minimizing gaps in the models themselves, refining the development and deployment processes, establishing robust quality assurance and testing protocols to ensure trustworthiness for specific use cases, continuous monitoring in production to detect emerging issues, and creating resilient recovery mechanisms to address newly discovered security or safety concerns.

Moreover, the challenges extend beyond the well-documented problems of LLMs and Multimodal Large Language Models (MLLMs), such as hallucinations or other known limitations.[15] For example, we often download models from public repositories without fully understanding how they were trained or their underlying behavior. These models function as "black boxes," making it difficult to predict their reliability, biases, or potential risks. While transparency frameworks like the Foundation Model Transparency Index (FMTI)[16] aim to address these issues, they often fall short—focusing narrowly on documentation and corporate disclosures, while overlooking critical aspects like interpretability, ethical considerations, and the role of open-source communities. This lack of transparency complicates efforts to build secure and trustworthy AI systems.

Considerations of security, safety, and resilience intersect across the AI landscape. For example, safety requires identifying and mitigating the impacts of adversarial attacks. It can also include security and encryption techniques to minimize data leakage that might empower an adversary, or an actor who might misappropriate data. Fundamental to system security and robustness are resilience mechanisms—means of recovering from deliberate attacks, LLM hallucinations, bias, data leakage, and other problems. For example, researchers have demonstrated how easily large language model (LLM) chatbots can be manipulated into bypassing their security measures. Simply asking the chatbot to assume the role of a grandmother and hinting at feeling tired could make the chatbot tell a bedtime story, sidestepping its alignment and security configurations.[17] In this experiment, the authors highlight the susceptibility of AI systems to straightforward social engineering tactics, emphasizing the need for more robust security measures in AI deployments.

The zero-trust methodology originally conceived for security can also be applied to: 1) establish a chain of identity, authorization, and access management; 2) minimize the impact of any issues discovered in the LLMs themselves, or the tools used around them in production— i.e., retrieval augmented generation, chain of thought, and fine-tuning; 3) and facilitate post-deployment recovery from newly discovered problems, operational challenges, and security attacks.

[15]Mou, Yutao, Shikun Zhang, and Wei Ye. "SG-Bench: Evaluating LLM Safety Generalization Across Diverse Tasks and Prompt Types." Advances in Neural Information Processing Systems 37 (2024): 123032-123054.
[16]Foundation Model Transparency Index. (Standford.). https://crfm.stanford.edu/fmti/May-2024/index.html Accessed: Mar. 11, 2025.
[17]Eran Shimony and Shai Dvash. Operation Grandma: A Tale of LLM Chatbot Vulnerability. https://www.cyberark.com/resources/threat-research-blog/operation-grandma-a-tale-of-llm-chatbot-vulnerability, Accessed: Mar. 11, 2025

New Trust challenges

Here are some of the ways these concepts show up with LLMs :

Safety: the condition of being protected from or unlikely to cause danger, risk, or injury.
In autonomous things, safety could mean protecting humans and infrastructure from harm. With LLMs, the concept broadens: could the AI offer dangerous advice such as mixing bleach and ammonia for cleaning, eating poisonous mushrooms, or recommending actions likely to cause harm. Safety can also mean protecting sensitive data that might be misused by bad or legitimate actors whose goals and policies differ from the data subject.

Resilience: positive adaptation, or the ability to maintain or regain health despite experiencing adversity.
In autonomous systems, resilience can mean compensating for failure and maintaining operations (or at least safely returning home) following accidental damage disruptions, challenging environmental conditions, or cyberattacks. With LLMs, resilience can also include safely recovering from hallucinations, data-poisoning attacks, bias in training data, and misuse.

Security: methods, tools, and personnel used to defend an organization's digital and physical assets.
For autonomous systems, security can mean protecting the development and deployment environment, safeguarding the applications in production, minimizing data leakage, securing communications, and protecting IP from reverse engineering.

[18]Shi, Dan, et al. "Large language model safety: A holistic survey." arXiv preprint arXiv:2412.17686 (2024).

New security vulnerabilities

There are many ways that hackers can compromise the security, safety and resilience of LLMs. Some relate to the LLMs' characteristics, while others are more esoteric and build on other attack vectors and vulnerabilities in their supporting infrastructure. Securing systems against these threats requires extending traditional Development, Security, and Operations (DevSecOps) approaches to continuously improve the security of applications to Machine Learning, Security, and Operations (MLSecOps) approaches for continuous security LLM models.

The Open Worldwide Application Security Project (OWASP) has compiled a list of the top 10 LLM security threats reported in the wild.[19] These include:

1. Prompt injection: Hackers craft a prompt that causes an unwitting LLM to execute their intentions. This can allow malefactors to extract sensitive information or influence decision-making processes.

2. Insecure output handling: Outputs passed to other systems can give attackers indirect access to additional functionality. This can enable privilege escalation or remote code execution on backend systems.

3. Training data poisoning: Attackers find ways to introduce poisoned data into an LLM's training data in ways that compromise its effectiveness, security, or ethical behavior. Hackers might plant this on websites used to train LLMs or exploit backdoor access to the training data repository. Such interference might allow an attacker to induce an LLM to generate biased or harmful outputs.

4. Model denial of service: Antagonists interact with the LLM in a way that devours resources, degrades performance, and increases operating costs. This can include not just increasing request volume but also amplifying the impact of requests via specially crafted prompts that take advantage of the LLMs' context limitations.

5. Supply chain vulnerabilities: Attackers find ways to insert malicious code or data that affects training data, model integrity, and deployment platforms. This can allow them to bias outcomes, breach security, or cause complete system failure. Attack scenarios include compromising Python libraries and registries, malicious plug-ins, poisoning models and public data sets.

6. Sensitive information disclosure: Attackers find a way to reveal sensitive information through an LLM's output, allowing the aggressors to access sensitive data and intellectual property, violate privacy, and launch other security breaches.

[19] OWASP Top 10: LLM & Generative AI Security Risks. OWASP Top 10 for LLM & Generative AI Security. https://genai.owasp.org/ Accessed: Mar. 11, 2025

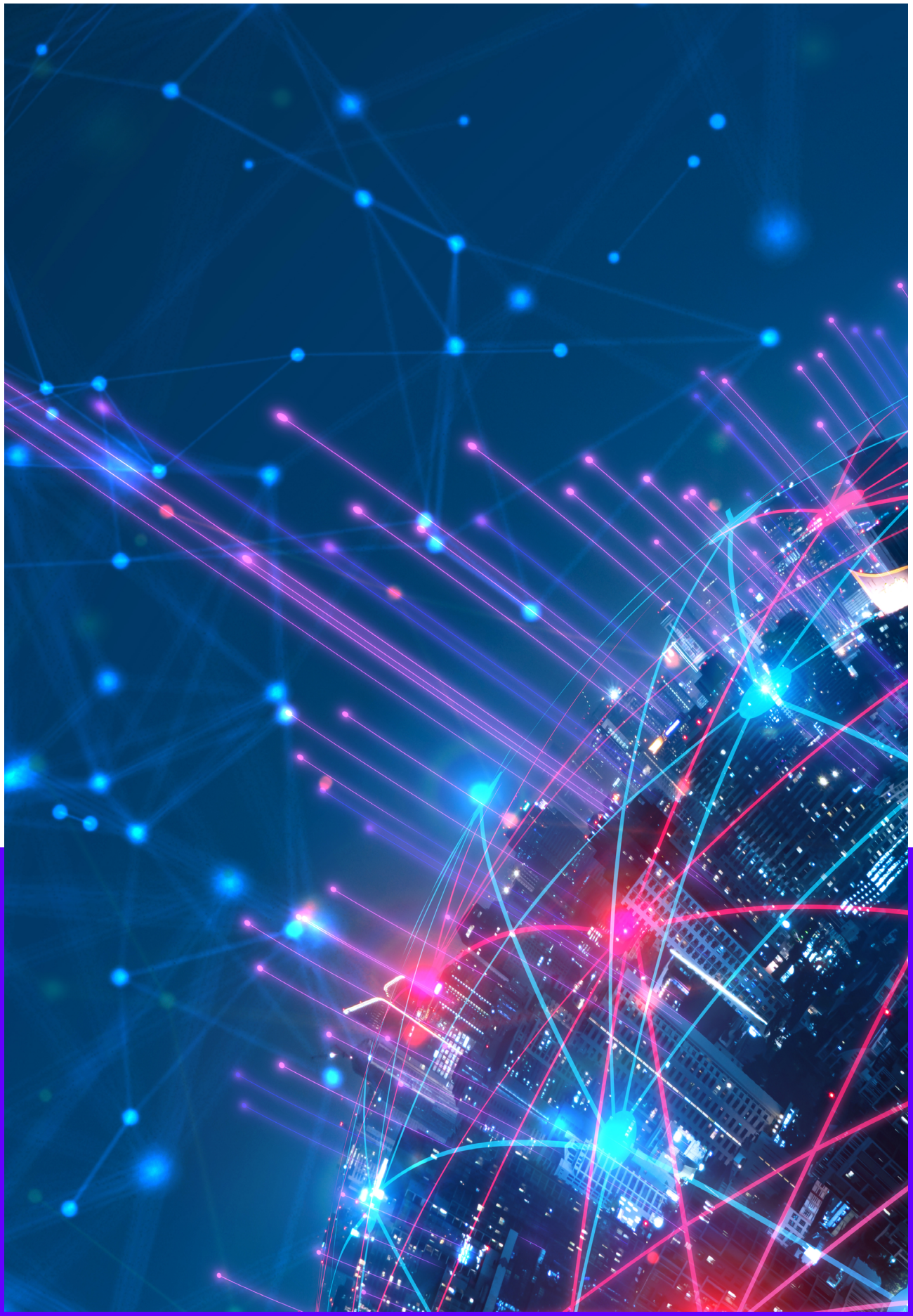
New security vulnerabilities

7. Insecure plugin design: Attackers compromise LLM plugins, typically Representational State Transfer (REST) APIs that are called during user interaction. This could allow hackers to activate a range of undesired behaviors, such as remote code execution.

8. Excessive agency: Weaknesses in identity and authorization mechanisms allow attackers to do damage with ambiguous LLM output. These undesirables include hallucinations, prompt injection, malicious plugins, poorly engineered prompt mechanisms, or poorly performing models. These let attackers exploit excessive functionality, permissions, or autonomy, harming the system's confidentiality, integrity, availability, and supporting data.

9. Overreliance: An unchecked LLM authoritatively hallucinates inaccurate or unsafe output, causing security breaches, miscommunication, legal issues, and reputational damage.

10. Model theft: Hackers find a way to compromise, physically steal, or copy the weights and parameters of an LLM to create a functional equivalent. This can cause economic damage, erosion of competitive advantage, or access to sensitive data within the model.



Safety issues

The main safety focus of zero-trust principles is to minimize the effects of hallucinations and reduce the blast radius of harmful decisions or actions taken on inaccurate or biased generated data.
Quality control is a growing challenge for LLM builders. The foundation models are often trained on unlabeled data that can reflect offensive and toxic behavior and unwanted biases. Humans are usually enrolled to help evaluate these models. They may be exposed to a raw stream of toxic content that can leave them traumatized. Cyberattackers may also find ways to poison training data and compromise the LLMs' safety.

Responsible AI vendors devote considerable effort to help filter out these types of safety issues. To protect their AIs during infancy, vendors like IBM and Meta have recently developed open-source guardrail models to help identify and rectify safety issues that might arise in training LLMs.

For example, IBM's Granite Guardian[20] model can identify safety issues related to social bias, hate speech, toxicity, violence, sexual content, unethical behavior, and jailbreaking. The sentinel can ameliorate some safety issues buried across the weights and interconnections represented as millions or billions of parameters in an LLM.

Similarly, Meta's LlamaGuard[21] incorporates a safety-risk taxonomy for various conversational use cases to moderate LLM model inputs and outputs, protecting against high-risk or policy-violating content.[22]
The tool also defends against adversarial inputs and jailbreaking attempts.

Resilience issues

LLMs also have underlying needs for explainability, ongoing monitoring, and CI/CD (continuous integration/continuous deployment or delivery) processes to identify and fix safety and security issues during production. Providing these for LLMs may be more challenging than it has been for earlier AIs, thanks to LLMs' large size and the extensive ecosystem of tools required to support them. This is where MLSecOps approaches can help to ensure a chain of trust across all artifacts and necessary infrastructure to support LLMs in production. Notably, the machine learning community has collectively advanced the MLSecOps framework,[23] which formalizes best practices in AI-specific security and promotes their adoption throughout the AI and machine learning lifecycle.

[20] IBM-granite/granite-guardian. (Oct. 30, 2024). Jupyter Notebook. IBM Granite. Accessed: Mar. 11, 2025. [Online]. Available: https://github.com/ibm-granite/granite-guardian
[21] Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations | Research - AI at Meta. (2023, December 7). https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/ Accessed Mar. 11, 2025
[22] H. Inan et al., "Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations," Dec. 07, 2023, arXiv: arXiv:2312.06674. doi: 10.48550/arXiv.2312.06674
[23] MLSECOPS Community. https://mlsecops.com/ Accessed Mar. 11, 2025.

Resilience issues

For example, one team of researchers at A*Star has identified five common types of disruptions in LLMs, from causes that can include:[24]

- Automatic Speech Recognition (ASR) errors,
- Optical Character Recognition (OCR) errors,
- grammatical mistakes,
- typographical errors, and
- distractive content.

One promising approach they suggested was a re-pass strategy, which distinguishes instructions from noise before they are forwarded to the LLM for processing. The group also found, however, that correcting noisy instructions poses significant challenges of its own.

Another group of researchers, The Royal Society and Humane Intelligence, has also explored ways of improving resilience to filter out scientific disinformation. They evaluated how red-teaming practice, initially conceived for cybersecurity testing, could also strengthen LLM-related processes.[25]

They investigated jailbreaking methods—techniques for crafting prompts that bypass safety features to generate harmful text and code. They found guardrail approaches that were effective in preventing the spread of disinformation. They also found, however, that participants could successfully break those guardrails and generate specific scientific disinformation.

LLMs often struggle to communicate the nuance of scientific debates and uncertainties. They are particularly good at authoritatively mimicking colloquial and scientific forms of communication, even when the information and argument are wrong. The Royal Society researchers found this particularly concerning. The group also identified the many ways that an LLM can draw equally from peer-reviewed scientific articles and unconfirmed pseudoscience and public relations material without qualifying the sources' validity or importance. Finally, the researchers found that LLM guardrail models, like Llama 2[26] can tamp down or shut down common pseudoscience distribution patterns.

[24]B. Wang, C. Wei, Z. Liu, G. Lin, and N. F. Chen, "Resilience of Large Language Models for Noisy Instructions," Oct. 03, 2024, arXiv: arXiv:2404.09754. doi: 10.48550/arXiv.2404.09754.
[25]Red teaming large language models (LLMs) for resilience to scientific disinformation | Royal Society." Accessed: Mar. 11, 2025. [Online]. Available: https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/
[26]Inan, Hakan, et al. "Llama guard: LLM-based input-output safeguard for human-ai conversations." arXiv preprint arXiv:2312.06674 (2023).

Resilience issues

TII has been developing its own alternative framework for improving the resilience of LLMs in autonomous system scenarios (see the diagram below).

Essential components of this framework consider and cover several aspects of machine-learning security across multiple domains, including:

• **Model Tampering and Integrity:** During the training phase, it's essential to ensure that the models are not tampered with or altered maliciously.

• **Data Assurance:** Humans can make errors and add faulty or corrupted data to training.

• **Robustness to Adversarial Attacks:** Models must be trained to be robust against adversarial attacks, where small, intentional perturbations to input data can mislead the model into making incorrect decisions.

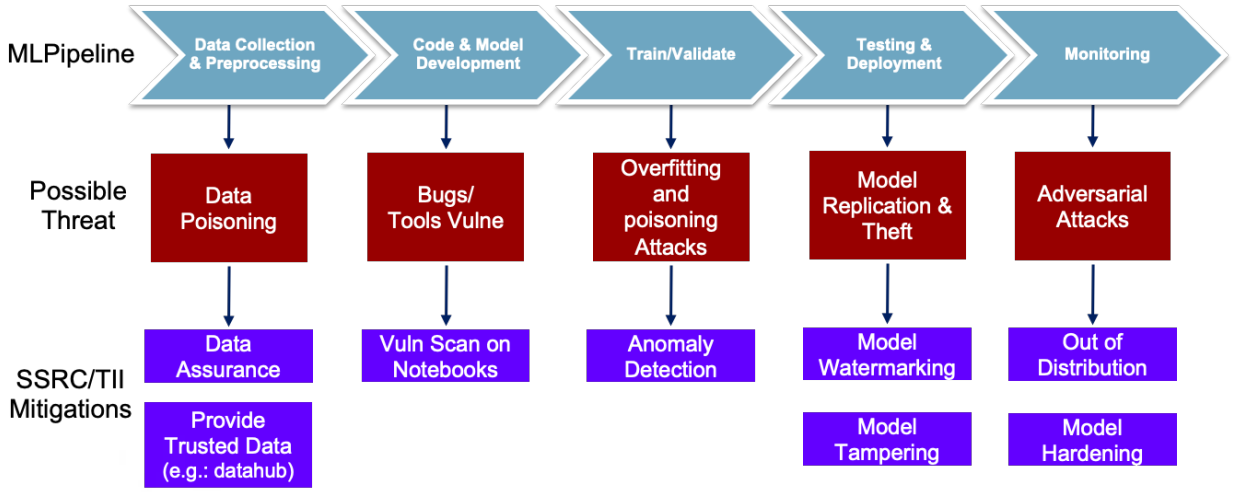
• **Secure Deployment Environments:** Once training is complete, the model deployment must be secured against unauthorized access and attacks. This includes secure provisioning of the model onto hardware, protecting the model from being copied or altered, and ensuring that the operational environment keeps the model's integrity intact.

• **Compliance and Traceability:** Maintaining compliance with industry standards and regulations is necessary throughout the training and deployment phases. This includes logging and monitoring all actions for auditability and traceability, which is essential not just for regulatory compliance but also for diagnosing and responding to incidents.

• **AI Safety:** Identifies unusual inputs or small changes from training data to protect the model from errors or attacks. This ensures that AI systems are reliable, ethical, and do not cause harm. It also focuses on preventing unintended actions, biases, and risks.

Resilience issues

Figure 2: SecMLOps[27] Pipeline: A Secure Machine Learning Operations Framework. This pipeline highlights key security threats at each stage of the ML lifecycle, from data collection to monitoring and corresponding mitigation strategies. By integrating security best practices such as data assurance, anomaly detection, model watermarking, and adversarial robustness, SecMLOps enhances the resilience of ML systems against attacks and vulnerabilities.



Trustworthy processes

Considerations around security, safety and resilience are required to adapt zero-trust principles from their initial focus on network and application security to LLMs. Their primary purpose remains improving the security and safety, now expanded tolls and their infrastructures and to LLMs' impacts on decision-making, recommendations, and actions. At the same time, zero-trust implementations consider resilience, to ensure security and safety even under adversarial attacks, new environments, or changes in infrastructure or models.

[27] Zhang, Xinrui, and Jason Jaskolka. "Conceptualizing the secure machine learning operations (secmlops) paradigm." 2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS). IEEE, 2022.

Trustworthy processes

Here are just a few of the varied factors that system architects should consider:

• **Distribution shift:** LLM performance may suffer when confronted with data that differs from its training set or in edge use cases.

• **Robustness to errors:** LLM systems may include smaller, specially trained LLMs, connected via prompt chaining or linked through a mixture of expert approaches. In these situations, however, minor errors in each component can accumulate to push complex tasks off-course.

• **Sharing adversarial prompts:** LLM builders should create and share libraries of adversarial prompts against which to test new prompts for malign intent.

• **Distributed learning:** LLM trainers should use decentralized processes, like federated learning and homomorphic encryption, to protect sensitive data while improving AI algorithms.

• **Monitoring:** AI architects should apply a variety of traditional and AI-powered tools to identify new vulnerabilities in LLM tools and infrastructure.

• **Community engagement:** LLM communities should offer bug bounty programs and support security working groups to enhance infrastructure.

Hallucinations

Hallucinations create special issues for developing trustworthy AI. Many AI and machine learning systems can suffer eroded accuracy and reliability, particularly when analyzing edge cases. LLMs may generate particularly persuasive hallucinations, thanks to their ability to produce output that is highly confident and realistic-seeming—whether or not it is accurate or misleading.[28] This can threaten system or human safety if the LLM recommends harmful actions or misleads people or other systems involved in making decisions.

Prominent examples include lawyers who have cited fake cases in legal proceedings and researchers misled into referencing non-existent publications in peer-reviewed papers. One airline lost a lawsuit after its chat service provided inaccurate advice on bereavement fare compensation.

Perhaps more troubling, LLMs have also been found to introduce inaccurate details in transcribing physicians' verbal notes into text. This could lead to catastrophic safety issues for patients, if other doctors rely on the hallucinations for later medical care. There are also examples of navigational AIs in autonomous systems that have mistaken one kind of object for another, or incorrectly interpreted or projected object behavior,[29] leading to accidents and death. Although these systems may not have used LLMs directly, the failures point toward some of the more serious safety issues that might arise as LLMs are adopted more broadly for developing and operating autonomous systems.

The growing adoption of agentic AI systems (which combine multiple smaller processes) can further compromise AI systems' security, safety, and resilience[30] Here, the concern is that hallucinations in different components can compound, to magnify effects across a much larger process.

[28]Mündler, Niels, et al. "Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation." arXiv preprint arXiv:2305.15852 (2023).
[29]Boudette, Neal E. "'It Happened So Fast': Inside a Fatal Tesla Autopilot Accident." International New York Times (2021): NA-NA.
[30]Liu, Bang, et al. "Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems." arXiv preprint arXiv:2504.01990 (2025).

Hallucinations

It should be noted that many types of hallucinations need to be considered and addressed. Some examples include:

- **Nonsensical response:** The model gives a response that makes no sense in the given context.
- **Creative liberties:** The model includes facts or information that is not reality-based.
- **Bias:** The model introduces biases owing to the training data.

Several benchmarks have been developed to quantify the hallucination rate. Combined with other hallucination mitigation techniques, these can be used to measure and improve the accuracy of LLMs and their results. Researchers are also developing hallucination benchmarking suites, such as the TRUE benchmark,[31] which considers accuracy across eleven measures.

Researchers are developing several strategies to help mitigate hallucinations, including:

Limiting context: Tools for retrieval augmented generation (RAG), particularly when combined with data organized into graphs, can limit the scope of information to focus on the context of the LLM.

Fine-tuning: Refining an LLM using new training data relevant to a particular domain.

Episodic memory: A newer fine-tuning technique for updating LLMs more efficiently and effectively.

Guardrails: Special purpose guardrails LLMs, such as IBM's Granite Guardian and Google's t5-11b-ANLI, can double-check results to block answers or refine prompts when hallucinations are detected.

Hallucinations manifest in many ways, and there are many ways to manage them. Consequently, it is important to develop modular frameworks to refine hallucination-detection and mitigation strategies, tailoring them for the task and acceptable risk level. For example, IBM is incorporating Granite Guardian into its watsonx.governance platform.[32] Cisco has proposed PolygraphLLM for creating generic building blocks that will make it easier to double-check for hallucinations using multiple approaches—while taking advantage of the most relevant benchmarks.[33] Also, NVIDIA offers NeMO-Guardrails for adding programmable guardrails to LLM-based conversational applications.[34]

[31] O. Honovich et al., "TRUE: Re-evaluating Factual Consistency Evaluation," May 03, 2022, arXiv: arXiv:2204.04991. Accessed: Mar. 11, 2025. [Online]. Available: http://arxiv.org/abs/2204.04991
[32] "IBM watsonx.governance." Accessed: Mar. 11, 2025. [Online]. Available: https://www.ibm.com/products/watsonx-governance
[33] cisco-open/polygraphLLM. (Oct. 31, 2024). Python. Cisco. Accessed: Mar. 11, 2025. [Online]. Available: https://github.com/cisco-open/polygraphLLM
[34] NVIDIA/NeMo-Guardrails. (Nov. 02, 2024). Python. NVIDIA Corporation. Accessed: Mar. 11, 2025. [Online]. Available: https://github.com/NVIDIA/NeMo-Guardrails

Spectrum of openness

Developers creating new AI systems on top of the coming generation of LLMs need to balance the increased capabilities of proprietary models against the security and safety risks of less visible interpretable open-source models. The open-source community is still trying to translate traditional open-source concepts to LLMs and other foundation models.

In traditional open source, a company or collective crafted and shared software code under one of many well-understood open-source licensing schemes. One concern is that vendors have also introduced various new licensing models for "somewhat-open" LLMs, licenses that diverge significantly from traditional agreements.

Another concern is that LLMs are immensely complicated. They are often shared as collections of model weights across millions or billions of parameters or features.

In addition, training data can be equally important in understanding potential biases, safety risks, and other factors that may affect an LLM in production.

The Open-Source Initiative (OSI), tasked with defining what "open-source" means, has defined open-source AI systems as publicly available applications grant users the freedoms to:[35]

- **Use** the system for any purpose and without having to ask for permission.
- **Study** how the system works and inspects its components.
- **Modify** the system for any purpose, including modifying it to change its output.
- **Share** the system for others to use, with or without modifications, for any purpose.

[35] "The Open Source AI Definition – 1.0," Open Source Initiative. Accessed: Mar. 11, 2025. [Online]. Available: https://opensource.org/ai/open-source-ai-definition

Spectrum of openness

The OSI specifies a preferred process for making modifications, which should also include all the following:

• **Data Information:** Sufficiently detailed information about the data used to train the system that a skilled person can build a substantially equivalent system. Data Information shall be made available under OSI-approved terms.
o In particular, the description must include: (1) the complete description of all data used for training, including (if used) unshareable data, disclosing the provenance of the data, its scope and characteristics, how the data was obtained and selected, the labeling procedures, and data processing and filtering methodologies; (2) a listing of all publicly available training data and where to obtain it; and (3) a listing of all training data obtainable from third parties and where to obtain it, including data provided for a fee.

• **Code:** The complete source code used to train and run the system. The Code shall represent the full specification of how the data was processed and filtered, and how the training was done. Code shall be made available under OSI-approved licenses.
o For example, if used, this must include code used for processing and filtering data, code used for training, including arguments and settings used, validation and testing, supporting libraries like tokenizers and hyperparameters search code, inference code, and model architecture.

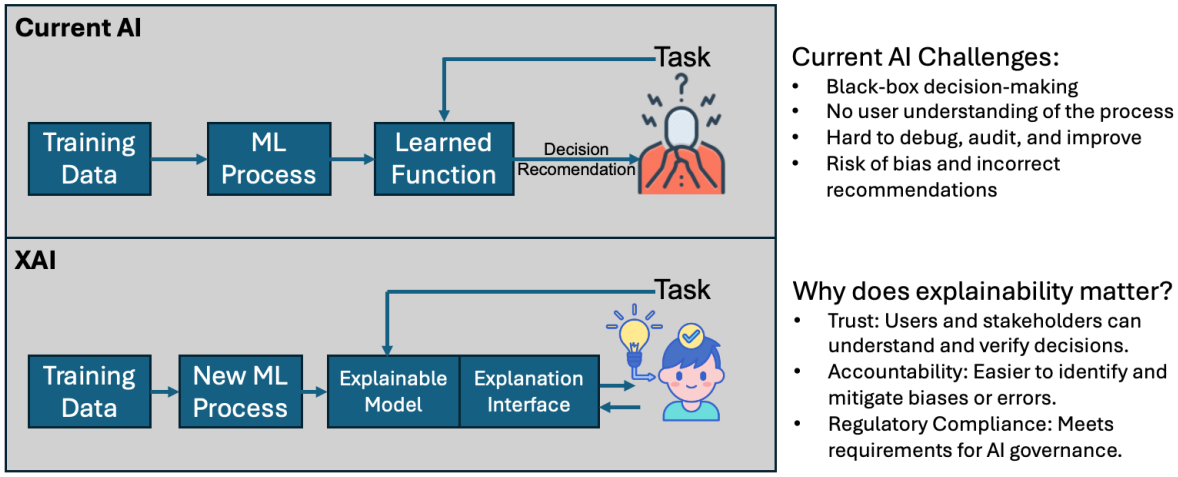
• **Parameters:** The model parameters, such as weights or other configuration settings. Parameters shall be made available under OSI-approved terms.
o For example, this might include checkpoints from key intermediate stages of training and the final optimizer state.


AI
AX-234r 09551P

Explainable artificial intelligence (XAI)

Explainable AI (XAI) provides users and others with information that allows them to understand how a given model generates responses or makes decisions.[36] It's a key element of the broader concept of AI transparency, which can help developers, users, and regulators understand how an AI system works.
A growing concern is that the sizable footprints of LLMs can confound efforts to understand how or why they make a particular decision that could impact security or safety. Figure 3 shows the comparison between traditional AI and Explainable AI.

Figure 3: Explainable AI (XAI) vs. Traditional AI: Traditional AI models provide decisions without transparency, leaving users uncertain about their reasoning. XAI introduces explainability mechanisms, allowing users to understand model decisions through interpretable explanations, fostering trust, accountability, and improved decision-making.



XAI is thus essential for building more resilient and trustworthy AI systems. Although XAI can help illuminate the workings of the models, it is also important to consider how the underlying training data can affect an LLM's security, safety, and resilience. It is further essential to consider ways to improve the governance of unstructured data used to enhance results using RAG and fine-tuning processes.

This is one area where more open approaches can improve the transparency, visibility, and explainability of LLM outputs. For example, today's proprietary LLMs usually surface few details in the data used to train their models. There are many reasons for this: respect for the data-owners' intellectual property, or fear of pushback from content creators (and others) for using data pulled from the web. Regulators are still unclear whether training AIs on copyrighted data constitutes IP infringement.
There has been some progress in improving tools for understanding factors—weights, features, neural network architectures—might spawn inaccurate results. These tools worked reasonably well for relatively simple AI models—e.g., for classifying images or data. But the remedies have been much harder to scale up for LLMs with hundreds of millions or billions of parameters.

[36] Das, Arun, and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey." arXiv preprint arXiv:2006.11371 (2020).

Explainable artificial intelligence (XAI)

There has been some progress in improving tools for understanding factors—weights, features, neural network architectures—might spawn inaccurate results. These tools worked reasonably well for relatively simple AI models—e.g., for classifying images or data. But the remedies have been much harder to scale up for LLMs with hundreds of millions or billions of parameters.

One survey of explainability and interpretability tools suggests breaking them down into the following categories:[37]

• Input attribution: These look at how inputs affect LLM performance.

• Component importance analysis: These consider how circuit discovery and causal interventions can balance the complexity of various intervention methods across various LLMs.

• Model internal visualization: These help visualize model weights and activations in LLMs.

Another survey, from researchers at Imperial College London, suggests sorting explainability techniques into seven broad categories:[38]

• **Model editing:** Tools to help developers refine models.

• **Enhancing model performance:** Tools to help improve long-text and in-context learning.

• **Controllable generation:** Tools to reduce hallucination and improve alignment.

• **Mechanistic interpretability:** Tools that help understand vocabulary, improve causality tracing, and discover neural network circuits.

• **Probing-based methods:** These are tools for probing knowledge and representations in LLMs.

• **Dissecting transformer blocks:** Tool for analyzing neural network sub-layers.

• **Feature attribution analysis:** Tool for understanding how slight prompt changes affect performance.

[37] J. Ferrando, G. Sarti, A. Bisazza, and M. R. Costa-jussà, "A Primer on the Inner Workings of Transformer-based Language Models," Oct. 13, 2024, arXiv: arXiv:2405.00208. doi: 10.48550/arXiv.2405.00208.
[38] H. Luo and L. Specia, "From Understanding to Utilization: A Survey on Explainability for Large Language Models," Feb. 22, 2024, arXiv: arXiv:2401.12874. doi: 10.48550/arXiv.2401.12874.

The importance of open-source collaboration

Today, many large AI vendors are developing comprehensive frameworks to improve the security, safety, and resilience of LLM-based solutions built on their platforms. Some vendors, such as IBM and Google, create hybrid models with many open-source components, such as TensorFlow, Hugging Face Transformers, among others, but require some proprietary elements to assemble a complete solution.

For example, IBM has open-sourced its capable Granite 3.0 LLMs and complementary Granite Guardrails models to improve LLM safety, theirs and others'. However, to take advantage of these tools, enterprises need to adopt IBM's Watsonx.AI governance platform to support the complete responsible-AI infrastructure.

Meanwhile, Google has developed its Secure AI Framework (SAIF) that includes 1) expanding strong security foundations into the AI ecosystem; 2) extending detection and response by integrating AI systems into the organization's security monitoring, ensuring timely identification and mitigation of AI-specific threats; 3) automating defenses to keep pace with existing and new threats; 4) harmonizing platform-level controls to ensure consistent security; 5) adapting controls to adjust mitigations and create a faster feedback loop for AI deployment, and 6) contextualizing AI system risk in surrounding business processes.[39]

In 2025, the Open Source Initiative (OSI) released the industry's Open Source AI Definition,[40] establishing clear guidelines for open-source artificial intelligence. This definition builds on the principles of traditional open-source software. It goes on to adapt them to the unique challenges of AI, including code, data, models, and training processes. As noted above, OSI's definition specifies four key freedoms (to use, to study, to modify, and to share for any purpose). OSI also emphasizes the importance of transparency in training data, model weights, and code, ensuring users can access the components necessary to understand, replicate, and improve AI systems. The initiative seeks to foster collaboration, innovation, and trust in AI development by providing a standardized framework for open-source AI while addressing ethical use, bias, and accountability concerns. This effort marks a significant step toward creating a more open and equitable AI ecosystem.

In the long run, improving LLMs' security, safety, resilience, and other foundation models will require similar approaches incorporating more open components. This will require a concise architecture that can absorb the best-of-breed open models most suited for a context, which can be changed or updated as needed.

[39] "Google's AI Security Framework – Google Safety Centre." Accessed: Mar. 11, 2025. [Online]. Available: https://safety.google/cybersecurity-advancements/saif/
[40] The open source initiative announces the release of the industry's first open-source AI definition. (2024, October 28). Open Source Initiative. https://opensource.org/blog/the-open-source-initiative-announces-the-release-of-the-industrys-first-open-source-ai-definition Accessed: Mar. 11, 2025

International cooperation and standards

Today's AI landscape includes dozens of new frameworks for building more trustworthy systems on top of LLMs and other foundation models. Offerings come from many national AI safety organizations. Internationally, the United Nations is rallying around the AI for Good framework. OWASP is exploring LLM security issues. The International Telecommunications Union (ITU) is exploring some data-sharing issues. And the ISO has proposed its ISO/IEC 23894 AI Risk Guidance standards.[41]

There are also numerous efforts to develop broader standards, tools, and best practices for responsible and safer AI more, with such programs as the US NIST AI RISK Management Generative AI Profile.[42] At least nine countries have also agreed to collaborate on an international network of AI safety institutes, modeled after existing programs in the US and UK.[43]

Moreover, other organizations are also doing work to improve AI safety more broadly, efforts that could be applied to LLMs. These include the AI Now Institute; the Berkman Klein Center for Internet & Society at Harvard University; the Institute for Technology, Ethics and Culture; and the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

These are all works in progress. The AI community has shared many security vulnerabilities and has discovered and explored new safety issues. Yet AI developers are still discovering how these highly complex systems can be compromised and exploited to create new safety issues that must be addressed swiftly. This is a particularly pressing where the applications include autonomous systems like drones and more agentic processes that could affect millions of people.

[41] "ISO/IEC 23894:2023(en), Information technology — Artificial intelligence — Guidance on risk management." Accessed: Mar. 11, 2025. [Online]. Available: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:23894:ed-1:v1:en
[42] "AI Risk Management Framework," NIST, Jul. 2021, Accessed: Mar. 11, 2025 [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework
[43] "U.S. Secretary of Commerce Gina Raimondo Releases Strategic Vision on AI Safety, announces Plan for Global Cooperation Among AI Safety Institutes," U.S. Department of Commerce. Accessed: Mar. 11, 2025

International cooperation and standards

Table 2: Key Global AI Governance Initiatives: This table highlights major efforts in AI risk management, security, ethical deployment, and safety evaluation. From international frameworks like the UN AI for Good and ISO/IEC 23894 to specialized security and safety initiatives like OWASP LLM Top 10 and the UK AI Safety Institute, these initiatives contribute to responsible AI development and deployment.

Initiative	Focus Area	Key Contribution	Source
UN AI for Good	Sustainable development	Promotes ethical AI use for global challenges like climate change and healthcare.	https://aiforgood.itu.int
ISO/IEC 23894	Risk management	Provides guidelines for managing AI risks across industries.	https://www.iso.org/standard/77304.html
OWASP LLM Top 10	Security vulnerabilities	Identifies and mitigates top security risks in LLMs (e.g., prompt injection).	https://owasp.org/www-project-top-10-for-large-language-model-applications
US NIST AI RMF	Risk management	Offers a framework for managing AI risks in US organizations.	https://www.nist.gov/itl/ai-risk-management-framework
UK AI Security Institute	AI safety evaluation	Evaluate and mitigate risks in advanced AI systems.	https://www.aisi.gov.uk/
Center for AI Safety (CAIS)	AI safety research	Conducts technical and conceptual research to reduce societal-scale AI risks	https://safe.ai/
AI Now Institute	Ethical AI	Explores the societal impacts of AI and advocates for responsible deployment.	https://ainowinstitute.org/

International cooperation and standards

Stanford HAI	Human-centered AI	Focuses on designing AI systems that prioritize human well-being and safety	https://hai.stanford.edu
International Network of AI Safety Institutes	Global collaboration	Facilitates knowledge sharing and coordinated research on AI safety among member countries.	https://www.nist.gov/document/international-network-ai-safety-institutes-mission-statement
Future of Life Institute	Existential risk	Advocates for safe and beneficial AI; influential in global policy and governance conversations.	https://futureoflife.org/

Zero-trust in practice

Here are some best practices that must be considered to align LLM deployments with zero-trust principles and architectures:

Data security and privacy are foundational. Only essential data should be collected (data minimization), and it must be protected through anonymization techniques such as differential privacy. All sensitive information should be encrypted at rest and in transit, with strict access controls based on role-based and least-privilege principles. Compliance with regulations like the EU's GDPR and the U.S. HIPAA is also essential to safeguard privacy.

Ensuring model training integrity requires secure, isolated environments for development, strong supply chain security through third-party audits, and transparent data provenance that tracks input sources and transformations. These measures reduce the risk of untrusted code or corrupted datasets compromising the training pipeline.

International cooperation and standards

To enhance resilience against attacks, models should undergo adversarial training to withstand crafted inputs. Continuous anomaly monitoring through intrusion detection systems, along with infrastructure redundancy and failover mechanisms, further harden systems against threats.

Safety mechanisms should include clearly defined protocols for emergencies, comprehensive risk management practices such as failure-mode and effects analysis (FMEA), and robust disaster recovery plans with secure, tested backups.

Verification and validation are also key. Models must be stress-tested under extreme conditions, and explainable AI techniques should be used to ensure transparency in decision-making. Where feasible, formal verification methods can provide mathematical guarantees of safety. Secure deployment pipelines must also be in place to control what reaches production.

Implementing a zero-trust architecture involves enforcing microsegmentation to establish multiple security boundaries, securing APIs with encryption and rate limits, and enabling continuous monitoring to rapidly detect and respond to incidents.

Finally, several LLM-specific vulnerabilities must be addressed. Data poisoning can be mitigated through strict validation and cleaning procedures. Techniques like regularization help defend against model inversion, while prompt injection risks require careful input sanitization. To counter hallucinations and harmful content, organizations should incorporate fact-checking, content filtering, and alignment techniques. Emergent behaviors, unexpected capabilities in large models, require adaptive monitoring and continuous evaluation as models evolve.

International cooperation and standards

Figure 4: Core Zero-Trust Practices for Securing Foundational Models Deployments. This layered architecture illustrates essential security domains—ranging from data privacy and training integrity to safety and model validation—necessary to implement zero-trust

principles across the Model lifecycle

Conclusion

The rapid advancement of large language models (LLMs) and foundational AI systems has unlocked transformative capabilities while introducing significant security, safety, and resilience challenges.
This whitepaper demonstrates how extending zero-trust principles—such as mutual authentication, least-privilege, and continuous monitoring—to the AI lifecycle can address these challenges.
By integrating zero-trust methodologies with techniques like retrieval-augmented generation (RAG), fine-tuning, and explainable AI (XAI), we can mitigate such risks as data poisoning, prompt injection, and hallucinations—and so pave the way for more trustworthy AI systems. At SSRC/TII (the Secure Systems Research Center of the Technology Innovation Institute), we are actively working on projects to enhance the security and safety of foundational models, contributing to the growing body of research that underscores the need for a consolidated framework to address these critical issues.

While initiatives from IBM's Granite Guardian, Google's Secure AI Framework (SAIF), OWASP and others have made progress, the complexity of LLMs demands ongoing innovation, collaboration, and standardization. Open-source approaches, international cooperation, and community-driven efforts are essential to ensure that future AI systems are secure, ethical, and resilient. Looking ahead, we must prioritize research into robust hallucination detection, adversarial training, and real-time monitoring to address emerging threats. Policymakers, industry leaders, and researchers must work together to harmonize global standards and foster a culture of shared responsibility.

The principles of zero-trust provide a robust foundation for building AI systems that are not only powerful but also trustworthy. By fostering transparency, accountability, and continuous improvement, we can ensure that AI technologies serve as a force for good, empowering humanity while minimizing risks. The road ahead is challenging, but with concerted effort and collaboration, we can pave the way for a future where AI systems are as trustworthy as they are transformative. The time to act is now. Let's build a future where AI works for everyone safely and securely.

Dr. Shreekant Thakkar
Chief Researcher
shreekant.thakkar@tii.ae

Dr. Martin Andreoni
Principal Researcher
martin.andreoni@tii.ae

Secure Systems Research Center (SSRC) – Technology Innovation Institute (TII)