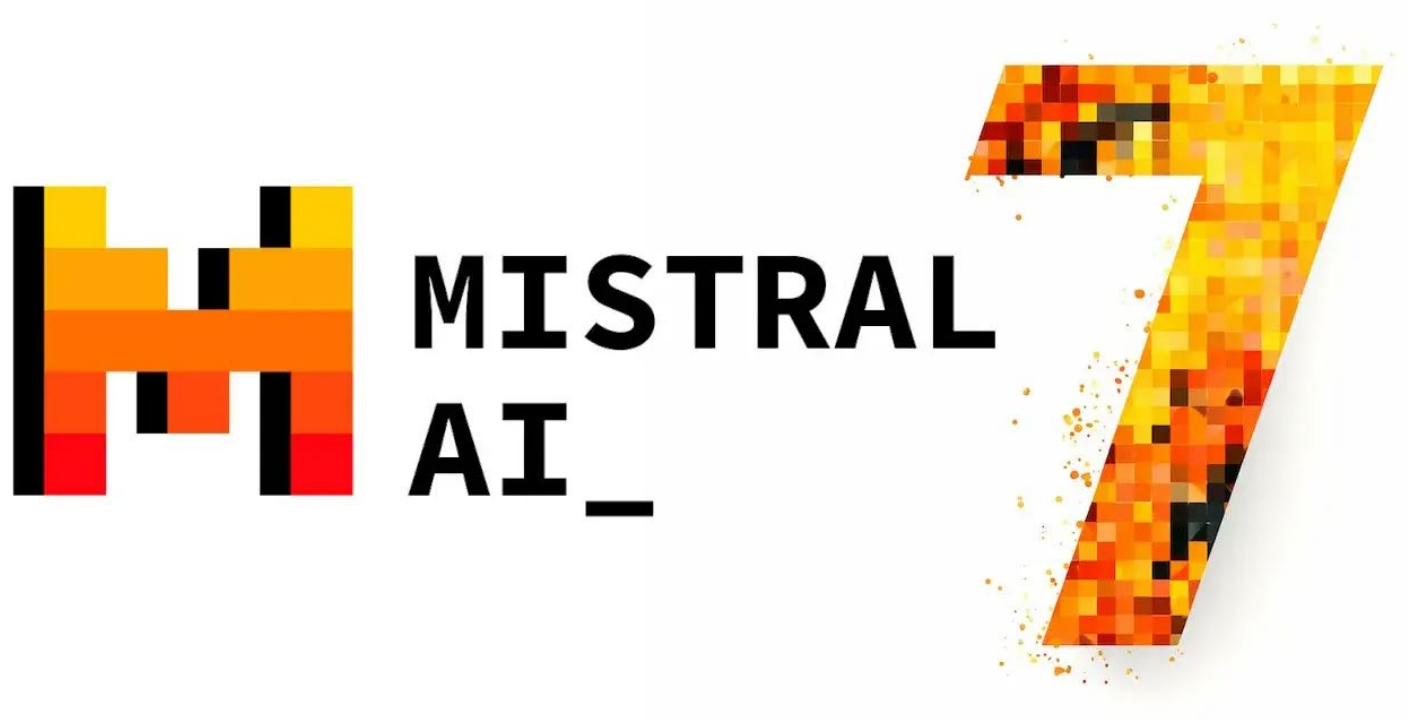


A futuristic robot with a white and grey body is shown from the back, reaching out towards a large, glowing blue digital interface. The interface is composed of a grid of binary code (0s and 1s) and intricate circuitry patterns. The robot's head is turned slightly to the right, and its arms are extended forward. The overall scene is set against a dark blue background with a subtle gradient.

Securing AI

Kate Carruthers
CRESTCon Canberra 2024

**How I
learned that
bad actors
would
embrace AI**



<https://mistral.ai/news/announcing-mistral-7b/>

MEMORY PROBLEMS —

Hacker plants false memories in ChatGPT to steal user data in perpetuity

Emails, documents, and other untrusted content can plant malicious memories.

DAN GOODIN - 9/25/2024, 6:56 AM

Within three months of the rollout, Rehberger **found** that memories could be created and permanently stored through indirect **prompt injection**, an AI exploit that causes an LLM to follow instructions from untrusted content such as emails, blog posts, or documents. The researcher demonstrated how he could trick ChatGPT into believing a targeted user was 102 years old, lived in the Matrix, and insisted Earth was flat and the LLM would incorporate that information to steer all future conversations. These false memories could be planted by storing files in Google Drive or Microsoft OneDrive, uploading images, or browsing a site like Bing—all of which could be created by a malicious attacker.


<https://arstechnica.com/security/2024/09/false-memories-planted-in-chatgpt-give-hacker-persistent-exfiltration-channel/>



AI is moving fast

If you can imagine it there is already someone somewhere researching it

LLMs are getting smaller

 > cs > arXiv:2402.17764

Search...
Help | Adv

Computer Science > Computation and Language

[Submitted on 27 Feb 2024]


The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, Furu Wei

Recent research, such as BitNet, is paving the way for a new era of 1-bit Large Language Models (LLMs). In this work, we introduce a 1-bit LLM variant, namely BitNet b1.58, in which every single parameter (or weight) of the LLM is ternary $\{-1, 0, 1\}$. It matches the full-precision (i.e., FP16 or BF16) Transformer LLM with the same model size and training tokens in terms of both perplexity and end-task performance, while being significantly more cost-effective in terms of latency, memory, throughput, and energy consumption. More profoundly, the 1.58-bit LLM defines a new scaling law and recipe for training new generations of LLMs that are both high-performance and cost-effective. Furthermore, it enables a new computation paradigm and opens the door for designing specific hardware optimized for 1-bit LLMs.

Comments: Work in progress

Subjects: **Computation and Language** (cs.CL); Machine Learning (cs.LG)

Cite as: arXiv:2402.17764 [cs.CL]
(or arXiv:2402.17764v1 [cs.CL] for this version)
<https://doi.org/10.48550/arXiv.2402.17764> 

Submission history

From: Shuming Ma [[view email](#)]

[v1] Tue, 27 Feb 2024 18:56:19 UTC (201 KB)

<https://arxiv.org/abs/2402.17764>

LLMs are getting more specialised

arXiv > cs > arXiv:2310.01728

Search...

Help | Adv

Computer Science > Machine Learning

[Submitted on 3 Oct 2023 (v1), last revised 29 Jan 2024 (this version, v2)]

Time-LLM: Time Series Forecasting by Reprogramming Large Language Models

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, Qingsong Wen

Time series forecasting holds significant importance in many real-world dynamic systems and has been extensively studied. Unlike natural language process (NLP) and computer vision (CV), where a single large model can tackle multiple tasks, models for time series forecasting are often specialized, necessitating distinct designs for different tasks and applications. While pre-trained foundation models have made impressive strides in NLP and CV, their development in time series domains has been constrained by data sparsity. Recent studies have revealed that large language models (LLMs) possess robust pattern recognition and reasoning abilities over complex sequences of tokens. However, the challenge remains in effectively aligning the modalities of time series data and natural language to leverage these capabilities. In this work, we present Time-LLM, a reprogramming framework to repurpose LLMs for general time series forecasting with the backbone language models kept intact. We begin by reprogramming the input time series with text prototypes before feeding it into the frozen LLM to align the two modalities. To augment the LLM's ability to reason with time series data, we propose Prompt-as-Prefix (PaP), which enriches the input context and directs the transformation of reprogrammed input patches. The transformed time series patches from the LLM are finally projected to obtain the forecasts. Our comprehensive evaluations demonstrate that Time-LLM is a powerful time series learner that outperforms state-of-the-art, specialized forecasting models. Moreover, Time-LLM excels in both few-shot and zero-shot learning scenarios.

Comments: Accepted by the 12th International Conference on Learning Representations (ICLR 2024)

<https://arxiv.org/abs/2310.01728>

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI)

Cite as: [arXiv:2310.01728](https://arxiv.org/abs/2310.01728) [cs.LG]

(or [arXiv:2310.01728v2](https://arxiv.org/abs/2310.01728v2) [cs.LG] for this version)

<https://doi.org/10.48550/arXiv.2310.01728> 

Hallucination is Inevitable

arXiv > cs > arXiv:2401.11817

Search...

Help | Advan

Computer Science > Computation and Language

[Submitted on 22 Jan 2024]

Hallucination is Inevitable: An Innate Limitation of Large Language Models

Ziwei Xu, Sanjay Jain, Mohan Kankanhalli

Hallucination has been widely recognized to be a significant drawback for large language models (LLMs). There have been many works that attempt to reduce the extent of hallucination. These efforts have mostly been empirical so far, which cannot answer the fundamental question whether it can be completely eliminated. In this paper, we formalize the problem and show that it is impossible to eliminate hallucination in LLMs. Specifically, we define a formal world where hallucination is defined as inconsistencies between a computable LLM and a computable ground truth function. By employing results from learning theory, we show that LLMs cannot learn all of the computable functions and will therefore always hallucinate. Since the formal world is a part of the real world which is much more complicated, hallucinations are also inevitable for real world LLMs. Furthermore, for real world LLMs constrained by provable time complexity, we describe the hallucination-prone tasks and empirically validate our claims. Finally, using the formal world framework, we discuss the possible mechanisms and efficacies of existing hallucination mitigators as well as the practical implications on the safe deployment of LLMs.

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)

Cite as: [arXiv:2401.11817](https://arxiv.org/abs/2401.11817) [cs.CL]

(or [arXiv:2401.11817v1](https://arxiv.org/abs/2401.11817v1) [cs.CL] for this version)

<https://doi.org/10.48550/arXiv.2401.11817> 

<https://arxiv.org/abs/2401.11817>

Submission history

From: Ziwei Xu [\[view email\]](#)

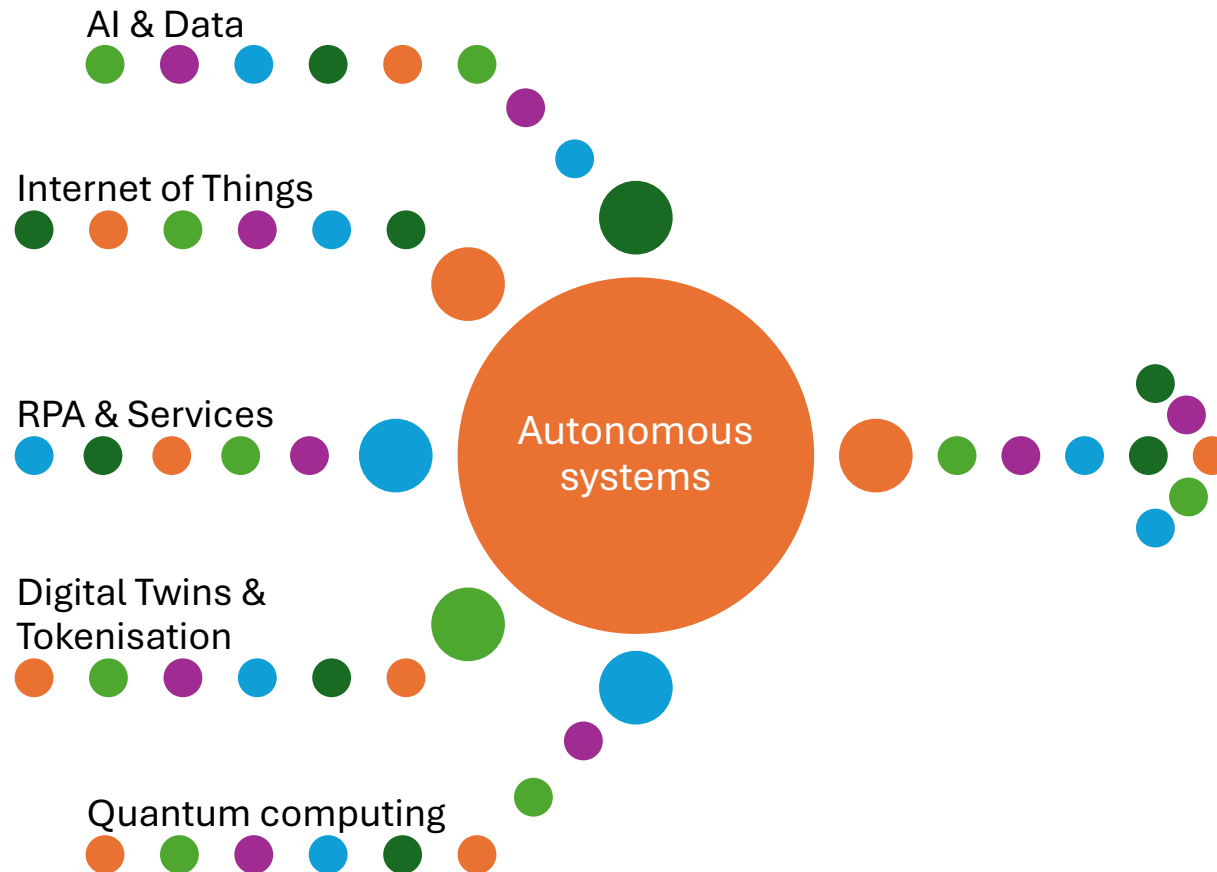
[v1] Mon, 22 Jan 2024 10:26:14 UTC (291 KB)

25/9/2024

A close-up, slightly blurred photograph of a colorful board game. In the center foreground is a red die with circular holes. To its left is a blue pawn, and to its right is a red pawn. The game board features a green field with yellow sheep, a brown brick path, and circular tiles with numbers like 7, 8, and 9. The background is out of focus, showing more of the game board and other pieces.

Current state of play

Where we are heading



Agentic AI is the new big thing

“Agentic AI goes beyond traditional AI by incorporating a "chaining" capability. This means it can take a sequence of actions in response to a single request, breaking down complex tasks into smaller, manageable steps.”

<https://www.forbes.com/sites/bernardmarr/2024/09/06/agentic-ai-the-next-big-breakthrough-thats-transforming-business-and-technology/>

Agentic AI is coming our way

Artificial Intelligence

Salesforce Unveils Agentforce—What AI Was Meant to Be

September 12, 2024 • 2 min read

 MEDIA LIBRARY



<https://www.salesforce.com/news/press-releases/2024/09/12/agentforce-announcement/>



25/9/2024

Agentic AI

- We are facing a democratisation of AI development with the emergence of Agentic AI
- Citizen developers will be developing these
- It is Robotic Process Automation (RPA) on steroids

Retrieval Augmented Generation (RAG) Language Models

“RAG, or Retrieval Augmented Generation, is a technique that combines the capabilities of a pre-trained large language model with an external data source. This approach combines the generative power of LLMs like GPT-3 or GPT-4 with the precision of specialised data search mechanisms, resulting in a system that can offer nuanced responses.”

<https://www.datacamp.com/blog/what-is-retrieval-augmented-generation-rag>

RAG Model

<https://www.rungalileo.io/blog/mastering-rag-how-to-architect-an-enterprise-rag-system>

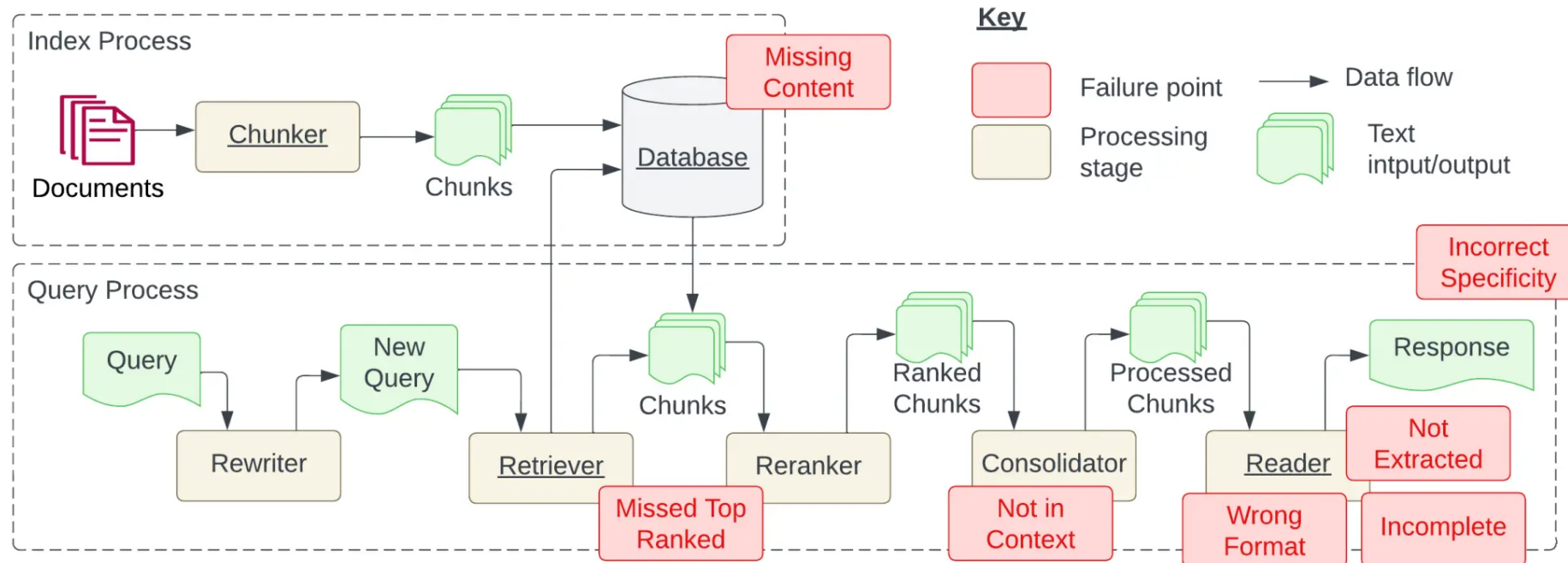


Figure 1: Indexing and Query processes required for creating a Retrieval Augmented Generation (RAG) system. The indexing process is typically done at development time and queries at runtime. Failure points identified in this study are shown in red boxes. All required stages are underlined. Figure expanded from [19].



Risks?

Data Security and Privacy Risks

AI systems process vast amounts of sensitive data
= **prime targets for data breaches**

Unauthorised access can lead to significant privacy violations and financial penalties for organisations

Robust **encryption** methods, secure communication protocols, & compliance with data protection regulations are essential

Rapidly evolving nature of AI complicates traditional **privacy frameworks**

Shift towards **ethical data governance** & stewardship that emphasises responsible handling of information once collected

Bias and Discrimination



AI models can perpetuate biases present in their training data, leading to discriminatory outcomes



Addressing this issue requires diverse training datasets, fairness-aware algorithms, and regular audits to ensure equitable decision-making processes



Organisations must also establish ethical guidelines to oversee AI applications and mitigate potential biases

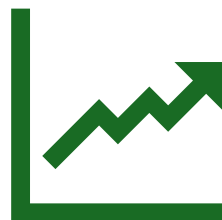
Reliability, reproducibility and Trust Issues

- AI systems can make **errors** that undermine trust in their outputs
- AI systems may not produce **reproducible** results
- Black box AI decision-making processes make it challenging for stakeholders to understand or predict system behaviour, leading to hesitancy in relying on these technologies for critical security decisions.
- Building trust through **transparency, explainability**, and robust performance metrics is vital for fostering confidence in AI applications

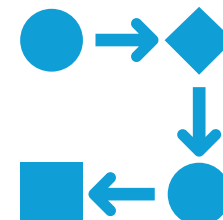
Regulatory Challenges



As AI technologies evolve, so do the regulatory landscapes governing their use



Increased complexity of global compliance requirements for AI operations



Ongoing re-evaluation of existing policies and practices.

Technical Integration Challenges

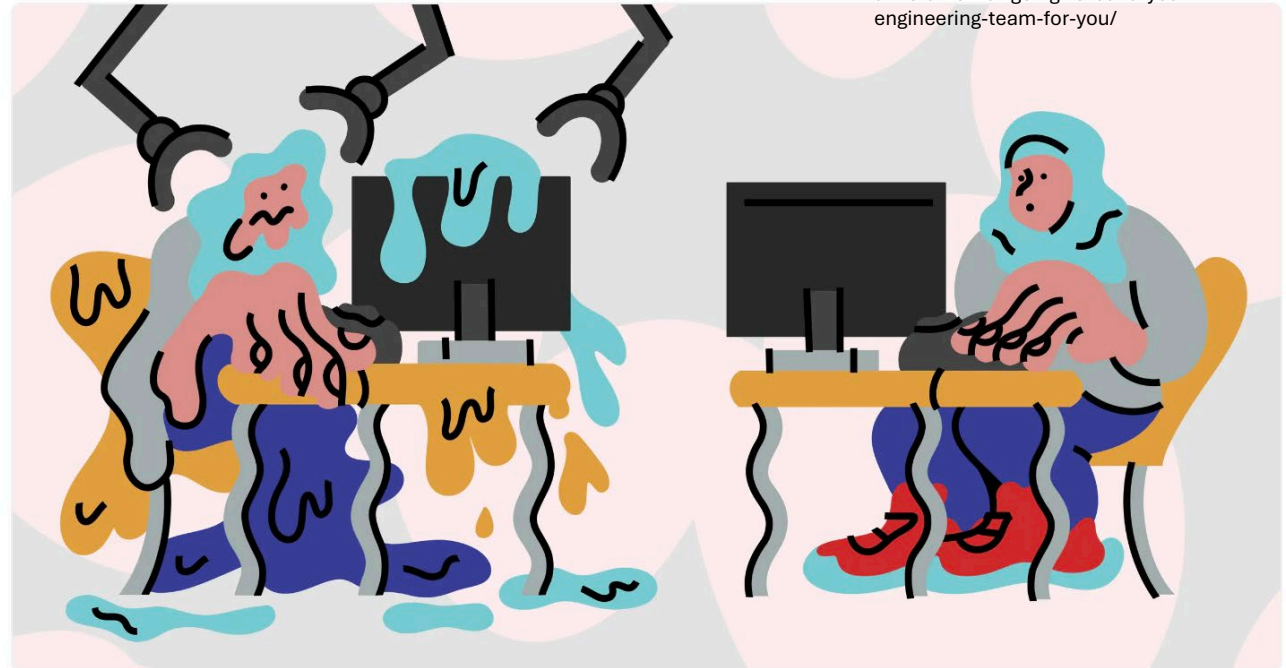
- Integrating AI with legacy systems poses significant hurdles due to compatibility issues and the need for substantial technical expertise
- Need to retrofit infrastructure and adapt data formats to accommodate new AI technologies

**People and
skills**

Generative AI Is Not Going To Build Your Engineering Team For You

It's easy to generate code, but not so easy to generate good code.

<https://stackoverflow.blog/2024/06/10/generative-ai-is-not-going-to-build-your-engineering-team-for-you/>

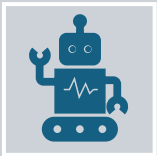


 Credit: Alexandra Francis

Networking for AI



AI/ML applications require low-latency, lossless networks to perform optimally.



High latency and packet drops can significantly increase the time it takes to complete AI/ML jobs, and in some cases, can even cause them to fail

**The big issues are
how to minimise
congestion and
reduce latency**



Cyber risks?

Increased Attack Surface



Integration of AI technologies broadens the attack surface, exposing them to new vulnerabilities.



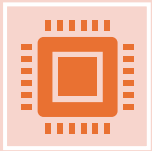
Observability: Need comprehensive visibility over AI infrastructure to identify and remediate potential risks effectively.



Organisations must **adapt** existing cybersecurity controls to address the unique challenges posed by AI systems while ensuring that all components are securely managed.

Adversarial Attacks

<https://rareconnections.io/adversarial-machine-learning-attacks/>



AI systems are vulnerable to adversarial attacks where malicious actors manipulate input data to deceive models into making incorrect predictions or decisions.



Techniques such as adversarial training—exposing models to both normal and adversarial examples—can help enhance resilience against such attacks.



Implementing input validation and anomaly detection mechanisms further protects AI systems from manipulation attempts.

Adversarial Training

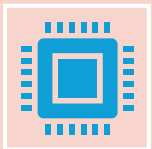
<https://www.linkedin.com/pulse/defending-ai-systems-combating-adversarial-attacks-rama-krishna>



Adversarial training involves incorporating adversarial examples into the training dataset.



By exposing AI models to these crafted inputs during training, they learn to recognise and respond to similar manipulations in real-world scenarios.



This method has shown promise in improving model robustness against various types of attacks.



Adversarial Attacks

Defensive Distillation



Defensive distillation is a technique that involves training a model with softened output probabilities, which helps reduce its sensitivity to small perturbations in input data.



This method can make models more resilient by effectively compressing knowledge from a complex model into a simpler one, thereby enhancing resistance to adversarial examples.

Input Transformation Techniques



Applying transformations to input data before it reaches the model can act as a first line of defense.



Techniques such as adding noise, filtering, or other modifications can obscure adversarial perturbations while preserving essential features of the data.

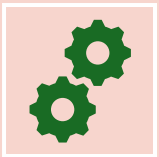


This helps to mitigate the impact of attacks by distorting malicious inputs.

Robust Model Architecture



Developing and utilising robust model architectures that are inherently less sensitive to adversarial manipulations is crucial.



Research into new architectures and training methodologies can lead to models that are better equipped to handle adversarial inputs without significant performance degradation.

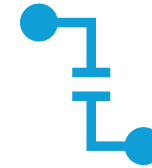
Evasion Attacks



Evasion attacks involve altering input data to mislead an AI model during inference



Modifications are often imperceptible to humans but can significantly affect model outputs

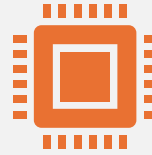


Evasion attacks can be further categorised into:

Targeted Attacks: The attacker aims for a specific incorrect output (e.g., misclassifying a stop sign as a yield sign).

Non-Targeted Attacks: The goal is simply to cause any misclassification, without a specific target output in mind

Poisoning Attacks



Poisoning attacks occur during the training phase, where attackers inject malicious data into the training dataset

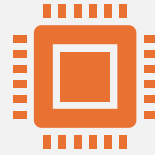


This tainted data skews the model's learning process, leading to incorrect associations and predictions once the model is deployed



Identifying the effects of poisoning can be challenging, as they may resemble other issues like overfitting

White-Box Attacks



In white-box attacks, the attacker has full access to the AI model's architecture, parameters, and training data.



This knowledge allows for precise manipulations tailored to exploit specific vulnerabilities within the model.

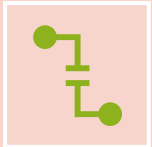


For example, an attacker could subtly modify an image of their face to be recognised as someone else by a facial recognition system[1][5].

Black-Box Attacks (inference)



Black-box attacks occur when attackers have no direct access to the model's internal workings



They rely solely on observing input-output behaviour to infer vulnerabilities

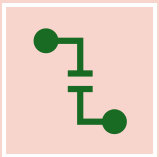


Despite limited information, these attacks can still be effective by systematically probing the model through various inputs

Adversarial Examples



Adversarial examples are inputs that have been altered to deceive AI systems into making errors.



These examples are typically generated using optimisation techniques that exploit the gradients of the model's loss function to create perturbations that lead to misclassification while remaining visually similar to original inputs.

Gradient-Based Attacks



These attacks utilise the gradients of a model's loss function with respect to its inputs to generate adversarial examples efficiently.



Techniques like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) fall under this category, allowing attackers to create adversarial inputs quickly by exploiting how models learn from data.

Takeaways

**Robust Security
Protocols**

**Ethical
Guidelines and
Oversight**

**Bias and
Fairness
Monitoring**

**Transparency
and
Explainability**

**Collaboration
and Information
Sharing**

**User Education
and Awareness**



25/9/2024

Thank you

Kate Carruthers

katec@unsw.edu.au