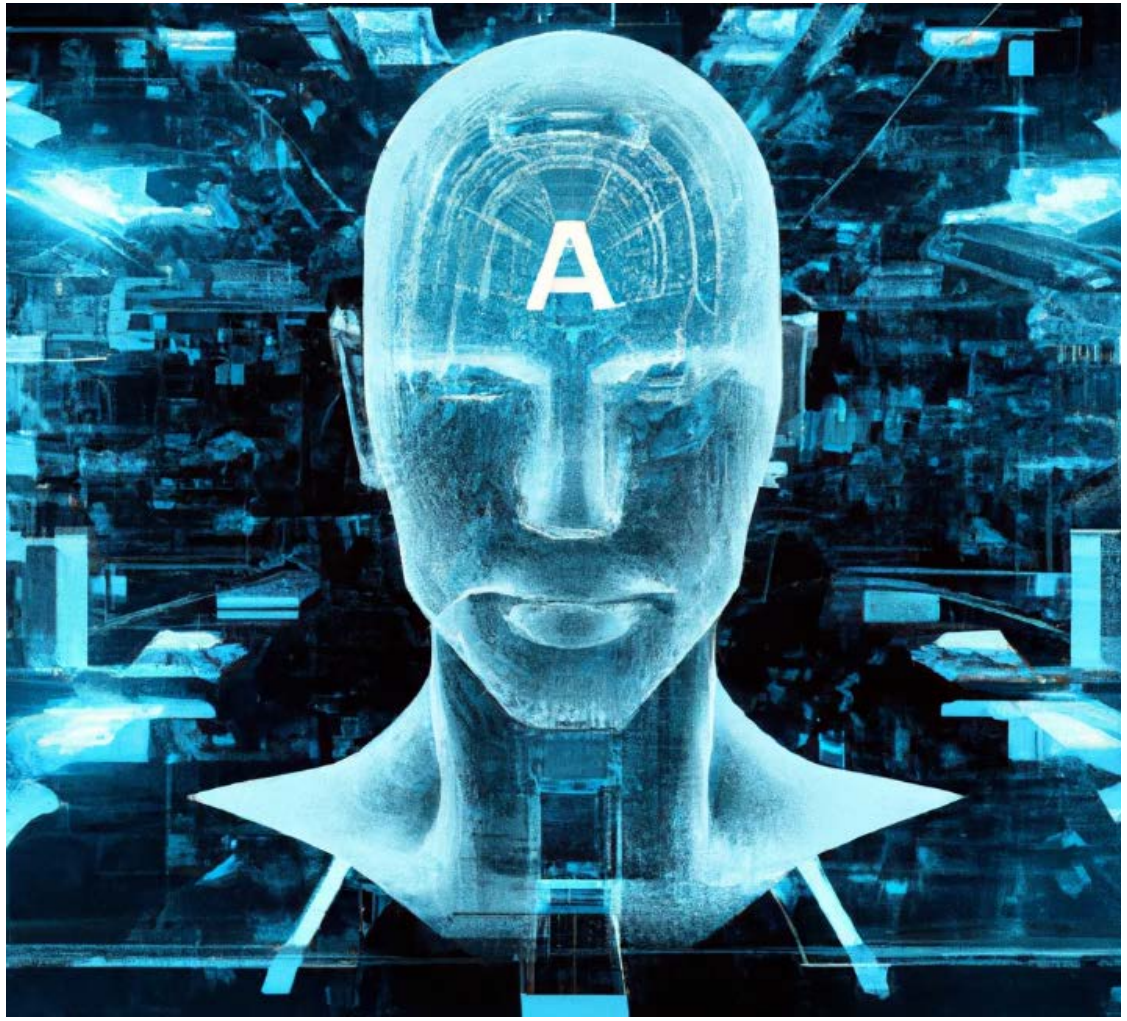
# Convergence of Cybersecurity and Large Language Models

## Exploring Threat Modeling in the World of LLMs

**Mike Arbrouet, MS/CISSP/CISM**
Information Security Strategist|Cyber Security|
Risk & Technology Management|Security Strat...

# Contents



# Generative AI

HOW I EXPLAIN IT TO MY KIDS

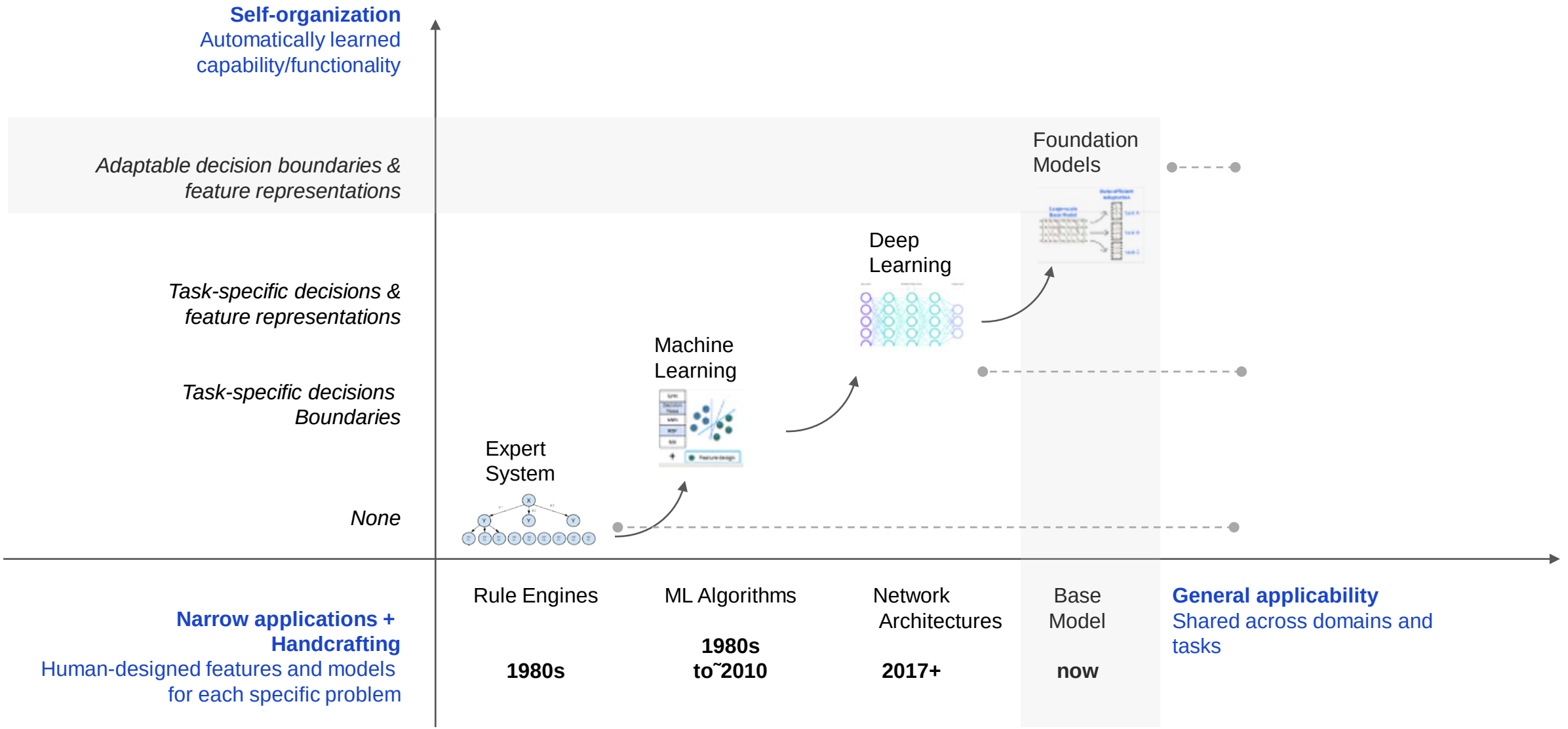

1. **Generative AI**: It's like a creative computer program.
2. **Two Players**: Imagine a Creator (makes stuff) and a Judge (tells real from fake).
3. **Game of Improvement**: Creator makes, Judge evaluates, Creator learns and improves.
4. **Outcome**: Creator gets really good at making things almost as good as humans

# LLMs

## HOW I EXPLAIN IT TO MY KIDS


- **What are LLMs?**: LLMs are super-smart computer programs that understand and generate human-like text.
- **Massive Knowledge**: They're trained on tons of books, articles, and websites to know a lot about different topics.
- **Understanding Language**: LLMs can understand and answer questions, write essays, translate languages, and more.
- **Chatting with LLMs**: You can have conversations with them, and they give helpful responses.
- **Applications**: Used in virtual assistants, content generation, translation, and much more.

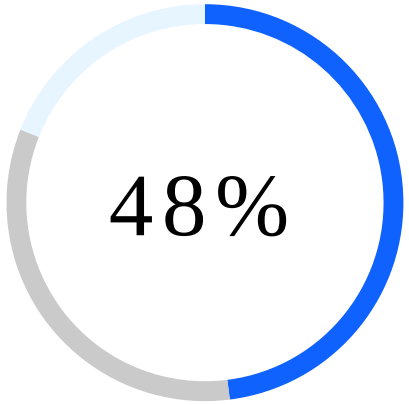
# Generations of AI



# Generative AI adoption considerations, inhibitors and fears

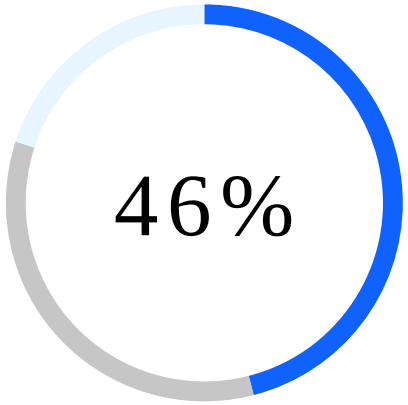
80% of business leaders see at least one of these ethical issues as a major concern

Explainability



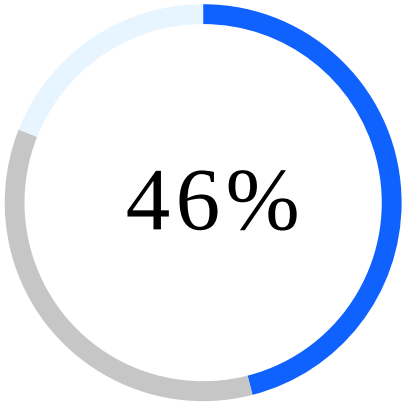

Believe decisions made by generative AI are not sufficiently explainable.

Ethics



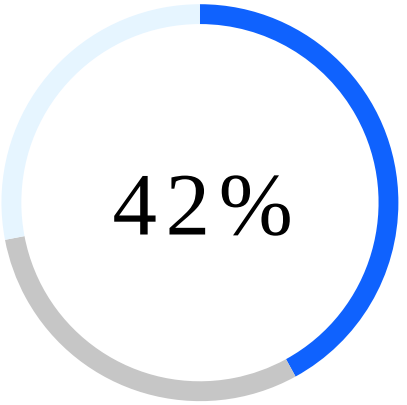

Concerned about the safety and ethical aspects of generative AI.

Bias




Believe that generative AI will propagate established **biases**.

Trust




Believe generative AI cannot be **trusted**.

Agree Neutral Disagree

Enterprises need more than an AI solution - they need a comprehensive and sound strategy for generative AI.

# Impact of generative AI

The speed, scope, and scale of generative AI impact is unprecedented

## Massive early adoption

**80%**

of enterprises are working with or planning to leverage foundation models and adopt generative AI

## Broad-reaching and deep impact

Generative AI could raise global GDP by

**7%**

within 10 years

## Critical focus of AI activity and investment

Generative AI expected to represent

**30%**

of overall market by 2025

Sources: Statista; Reuters; Goldman Sachs; IBM Institute for Business Value; Gartner. Scale Zeitgeist: AI Readiness Report, a survey of more than 1,600 executives and ML practitioners

# Risks associated with Generative AI

## Security

These models are susceptible to data and security risks including prompt injection attacks.

## Bias

The training data has an impact on the results the model produces. Foundation Models are trained on large portions of data crawled from the internet.

Consequently, the biases that inherently exist in internet data are picked up by the trained models and can show up in the results.

## Opacity

Foundation Models are also not fully auditable or transparent because of the “self-supervised” nature of the algorithm’s training.

## Hallucination

LLMs can produce “hallucinations," results that satisfy a prompt syntactically but are factually incorrect.

## IP

There are unanswered questions concerning the legal implications and who may own the rights to content generated by models that are trained on potentially copywritten material.

# Most common generative AI tasks implemented today

## Summarization

Transform text with domain-specific content into personalized overviews that capture key points.

Conversation summaries, insurance coverage, meeting transcripts, contract information

## Classification

Read and classify written input with as few as zero examples.

Sorting of customer complaints, threat and vulnerability classification, sentiment analysis, customer segmentation

## Generation

Generate text content for a specific purpose.

Marketing campaigns, job descriptions, blog posts and articles, email drafting support

## Extraction

Analyze and extract essential information from unstructured text.

Medical diagnosis support, user research findings

## Question-answering

Create a question-answering feature grounded on specific content.

Build a product specific Q&A resource for customer service agents.

# Navigating the Maze of AI and ML Threats


| |
|---|
| **Prompt Injection** |
| **Data Poisoning** |
| **Data Leakage** |
| Model Evasion/Abuse |
| Safe Behaviour |

# RISK– Vulnerability – IMPACT: 7th GRADER EXPLANATION

Alright, imagine you're playing with a toy castle

## RISK

The chance that your toy castle might break

## THREAT

Someone, maybe your little brother, who might want to push or knock down your castle.

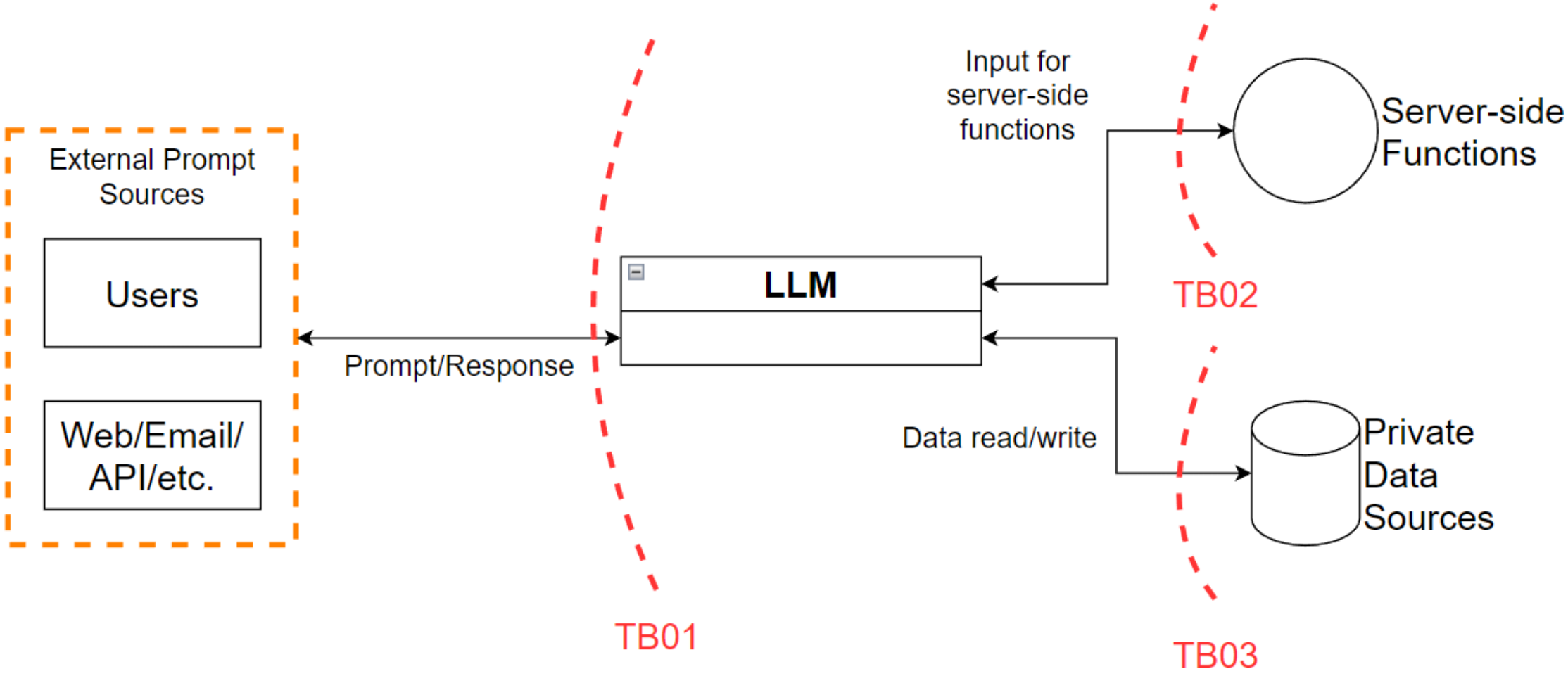
## VULNERBAILITY

The weak part of your castle, like a loose brick.

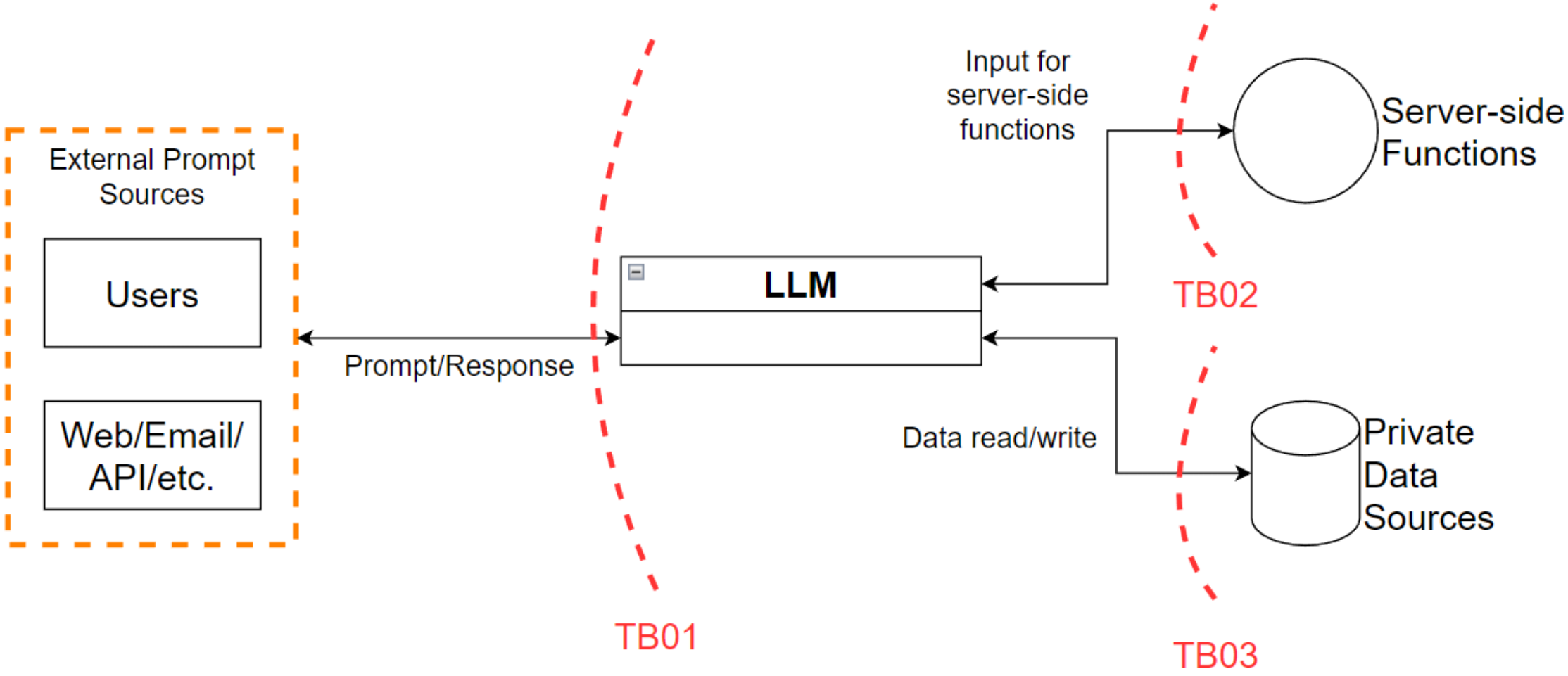
## IMPACT

If the castle breaks, you can't play with it until it's fixed

# DATA FLOW DIAGRAM



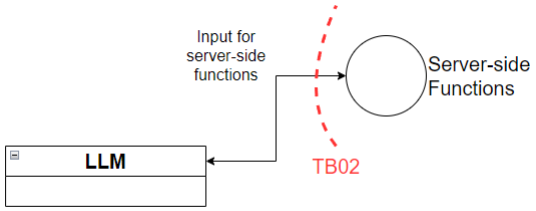
# DATA FLOW DIAGRAM



# STRIDE

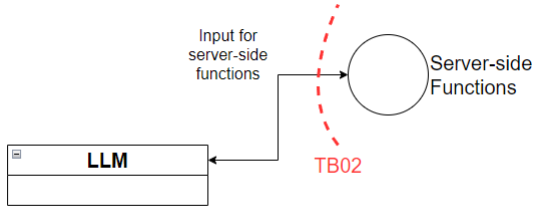
| | |
|---|---|
| **Spoofing** | Pretending to be someone or something you're not.<br>For example, sending emails from someone else's address. |
| **Tampering** | Changing data or code without permission.<br>Like secretly altering a message sent between two people. |
| **Repudiation** | Denying having done something, even if you did it.<br>Like sending a message and later claiming you never sent it. |
| **Information Disclosure** | Getting access to data that's supposed to be kept secret.<br>Like someone reading your private messages without permission. |
| **Dos** | Preventing legitimate users from accessing a service.<br>Like overwhelming a website with traffic so it crashes. |
| **Escalation of Privilege** | Gaining higher privileges than you should have, often to do malicious activities. Like a student getting admin rights on a school computer. |

# Trust boundary TB01 - STRIDE table



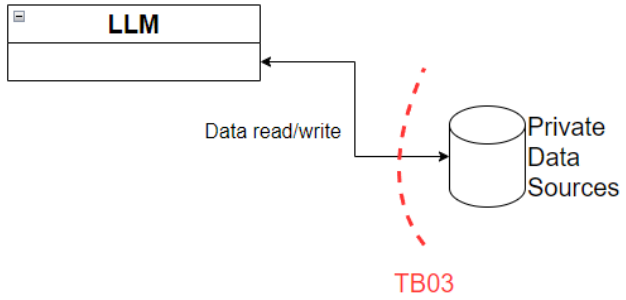

| Vuln ID | Description | Examples |
|---|---|---|
| VULN01 | Modify System prompt (prompt injection) | Users can modify the system-level prompt restrictions to "jailbreak" the LLM and overwrite previous controls in place |
| VULN02 | Modify LLM parameters (temperature, length, model, etc.) | Users can modify API parameters as input to the LLM such as temperature, number of tokens returned, and model being used. |
| VULN03 | Input sensitive information to a third-party site (user behavior) | Users may knowingly or unknowingly submit private information such as HIPAA details or trade secrets into LLMs. |
| VULN04 | LLMs are unable to filter sensitive information (open research area) | LLMs are not able to hide sensitive information. Anything presented to an LLM can be retrieved by a user. This is an open area of research. |

# Trust boundary TB02 - STRIDE table




| Vuln ID | Description | Examples |
|---|---|---|
| VULN05 | Output controlled by prompt input (unfiltered) | LLM output can be controlled by users and external entities. Unfiltered acceptance of LLM output could lead to unintended code execution. |
| VULN06 | Server-side output can be fed directly back into LLM (requires filter) | Unrestricted input to server-side functions can result in sensitive information disclosure or server-side request forgery (SSRF). Server-side controls would mitigate this impact |

# Trust boundary TB03 - STRIDE table




| Vuln ID | Description | Examples |
|---|---|---|
| VULN05 | Output controlled by prompt input (unfiltered) | LLM output can be controlled by users and external entities. Unfiltered acceptance of LLM output could lead to unintended code execution. |
| VULN07 | Access to sensitive information | LLMs have no concept of authorization or confidentiality. Unrestricted access to private data stores would allow users to retrieve sensitive information. |

# Recommendations

| |
|---|
| **VULN01** - Modify System prompt (prompt injection): Mitigate effects by ensuring the LLM doesn't train on confidential data and always treat its output as untrusted. |
| **VULN02** - Modify LLM parameters: Limit API exposure to external prompts and filter all external inputs. |
| **VULN03** - Input sensitive information to a third-party site: Educate users about the risks every time they connect to the LLM. |
| **VULN04** - LLMs are unable to filter sensitive information: Don't train LLMs on sensitive data and apply external data source controls. |
| **VULN05** - Output controlled by prompt input: Always treat LLM output as untrusted and restrict its use for other functions. |
| **VULN06** - Server-side output can be fed back into LLM: Filter server output and sanitize sensitive information before retraining or returning to users. |
| **VULN07** - Access to sensitive information: Treat LLM access like any user and enforce standard data access controls. |