AI Red-Teaming

Unmasking the Vulnerabilities in LLMs

Presenters

Managing Director, Cybersecurity and Privacy
Tech, Media & Entertainment Cybersecurity Lead

E Caesar.Sedek@us.gt.com

Intro to AI and LLMs

- **What is AI**
  - Simulation of human intelligence by machines
  - Capabilities include learning, reasoning, and language understanding
- **Large Language Models (LLMs)**
  - AI models that process and generate human-like text
  - Key features: context understanding, coherent text generation
- **Importance of LLMs in Business**
  - Transform business operations with advanced analytics and automation
  - Improve decision making, efficiency and customer interactions
- **Applications Across Industries**
  - **Finance:** Automated reporting, risks assessment, conversational finance
  - **Healthcare:** Improved patient interactions, streamlined medical documentation, image analysis
  - **Customer Service:** Automated responses, enhanced customer engagement

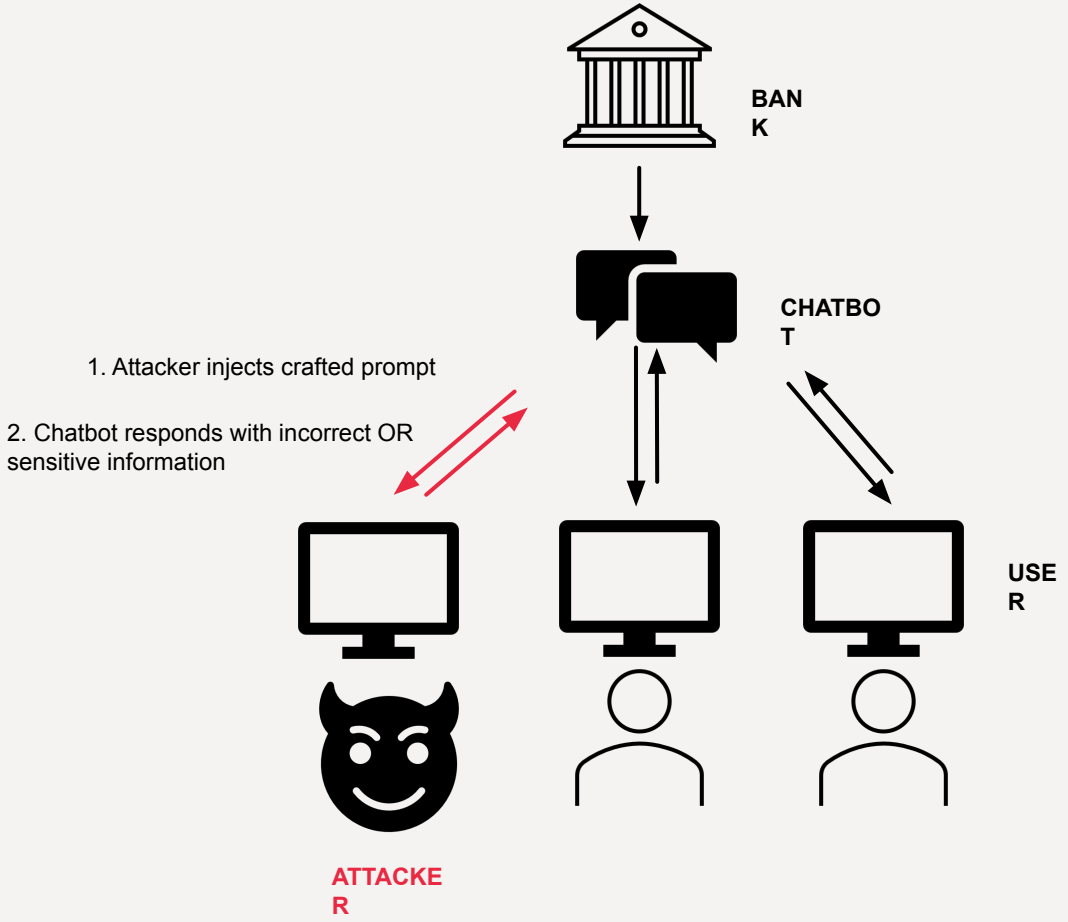
Clearly not the best at generating abstract infographics. Prompt: “Generate an Infographic illustrating different LLMs (e.g., GPT, BERT) along with their primary applications across industries such as finance, healthcare and customer service.

Result

Understanding the Security Risks & Threats

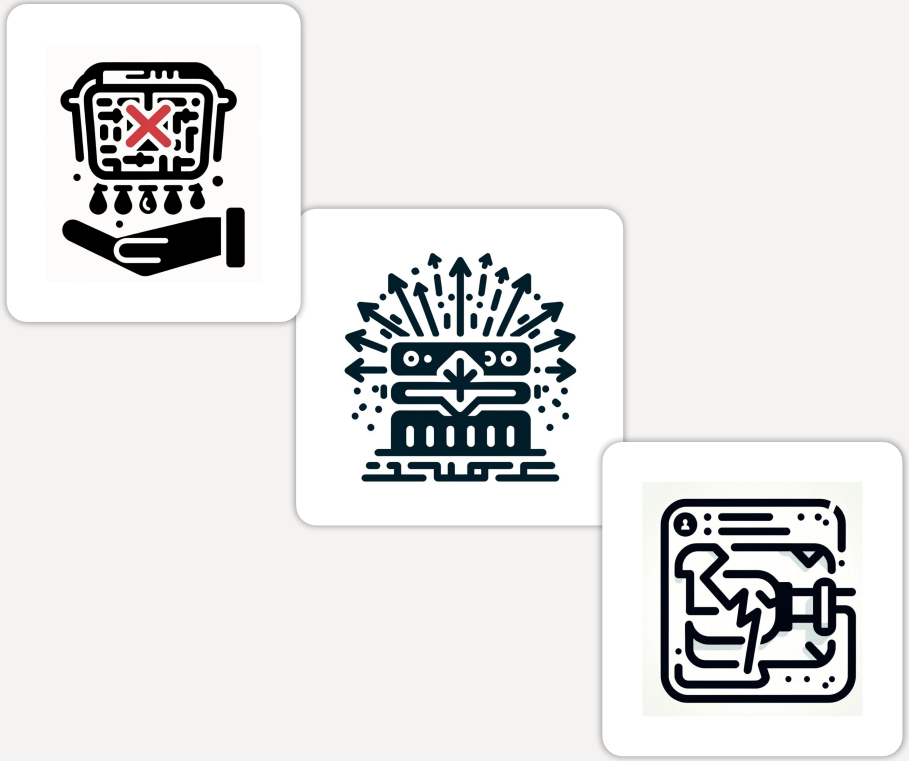
Because Large Language Models (LLMs) are secure by default, right?

- **Prompt Injection Attacks:**
  - Malicious prompts used to manipulate LLM outputs leading to unauthorized actions or data exposure
  - Undermine model's reliability
  - Incorrect or harmful output
- **Prompt Leaking**
  - LLM reveals its own prompts or internal processing logic
  - Differs from prompt injection – doesn't alter model's behavior – but extracts information about the model itself
- **Data Leakage**
  - Unintentional exposure of sensitive information through outputs of LLMs due to flaws in model's design or training data
- **Personally Identifiable Information (PII) in LLMs**
  - PII in LLMs prompts poses privacy risks
  - Exposure of personal identities
- **Compliance issues with data protection laws**
  - Use of LLMs to process PII may violate GDPR, CCPA, other privacy laws/regulations



Understanding the Security Risks & Threats

Because Large Language Models (LLMs) are secure by default, right?



- **Insecure Output Handling**
  - Unsafe or harmful content due to lack of output handling / filtering
- **Model Denial of Service (DoS)**
  - Overwhelming LLM with a flood of requests or inputs rendering it unable to function
- **Insecure Plugin Design**
  - Plug-ins or extensions can introduce vulnerabilities
  - Insecure interfaces/APIs
- **Excessive agency and overreliance**
  - Excessive agency gives LLM more autonomy or functions than necessary
  - Security risk due to blind trust in outputs and neglect of anomaly detection
- **Model theft**
  - Unauthorized access, copying or use of proprietary LLM models

Once again – let's give it up to the creative genius of DALL-E!

Mitigation Strategies

Best practices for security Large Language Models (LLMs) against prompt injection attacks.

- **Input Validation and Sanitization**
  - Implement Strict Validation rules for inputs based on format, type and length
  - Automatically sanitize inputs to remove or encode harmful characters or patterns
- **Allowlists**
  - Use allowlists (permitted inputs) over blocklists (forbidden inputs)
- **Role-Based Access Control / Zero Trust**
  - Limit permissions across the stack to only those strictly necessary
  - Minimize potential damage of a successful attack
- **Secure Prompt Design**
  - Design prompts that limit user's ability to influence execution path
  - Use structured data as input where possible vs. free text
  - Template use with variables
- **Regular Expression (RE) Check**
  - Use RE to identify and block potentially malicious patterns in inputs
- **Logging and Monitoring**
  - Log/monitor unusual patterns of use
  - Detailed audit trails

Prompt Leak Prevention

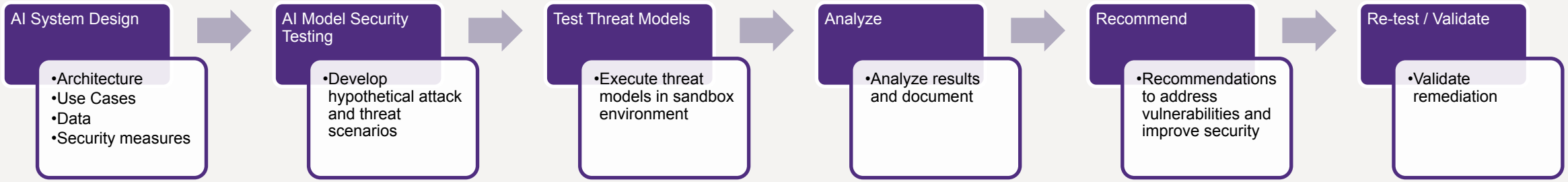
Strategies for addressing prompt leaking and maintaining the integrity and confidentiality of LLMs

- **Data Anonymization, Data Redaction and Pseudonymization**
  - Anonymize data sent to prompts
  - Automated Redaction
  - Replace sensitive data with non-identifiable placeholders that maintain reference integrity
- **Input Validation and Sanitization**
  - Remove or encode characters and patterns in input data
  - Ensure data sent to LLM adheres to expected formats and ranges
- **Encryption**
  - Use strong encryption to prevent man-in-the-middle (MITM) attacks
  - Encrypt sensitive data used as part of model's training data
- **Secure Data Handling Practices**
  - Encrypt data at rest and in transit
  - Use strict access controls to limit both read and write access to data
- **Secure Authentication Mechanisms**
  - Implement strong authentication (e.g., Multi-factor authentication)
- **API Gateways**
  - Use application programing interface (API) gateways with rate limiting and monitoring
- **Education & Awareness**
  - Inform users about types of data that system can process
  - Guidelines on what information shouldn't be submitted into the system
- **Regular Audits and Assessments**
  - Conduct regular security audits and privacy assessments
  - Penetration testing against internal systems
  - AI Red-Teaming
- **Privacy by Design**
  - Adopt a privacy-by-design approach

AI Red-Teaming

Assess your AI solutions before they are challenged by real-world adversaries

- **What is AI Red-Teaming**
  - Simulating attacks on AI systems to identify and address vulnerabilities before they can be exploited maliciously
- **Why Red-Teaming**
  - Maintain integrity and trustworthiness of AI system in critical applications
  - Proactive approach
- **AI Red-Teaming Strategies**
  - Attacking the Model: Techniques include input manipulation, exploiting model biases, finding loopholes in model logic
  - Attacking the Developer: Social engineering aimed at exploiting human factors and system configurations



Our Cybersecurity & Privacy solution overview

- Cybersecurity program risk and maturity assessment; Cybersecurity program implementation
- Design and implement governance, risk, and compliance (GRC) technology solutions

Cyber Defense Solutions

- Vulnerability assessment, penetration testing, and red teaming; Cyber incident tabletop exercises
- Cyber defense technology implementation; managed Cyber analytics (MCA) services

Privacy & Data Protection

- Personal data inventory, privacy program readiness assessment and implementation (GDPR, CCPA)
- Data protection assessments; privacy solution implementation (data discovery, classification, retention, leakage protection)

Identity and Access Management

- Identity and access management strategy; privileged & role-based access implementation
- Identity and access management technology implementation and application onboarding

Third Party Risk Management (TPRM)

- Program Design and Strategy
- TPRM Program Execution/Assessments and Technology Automation

Grant Thornton

Questions?

Grant Thornton

Thank You!