## **1. Prompt Injection**

• Attacker manipulates user input or context to override intended behavior.

• Variants: direct injection, indirect injection via documents or external tools.

## **2. Data Leakage / Training Data Extraction**

• LLM reveals sensitive information memorized from training data.

• GPT-2 and GPT-3 have shown susceptibility to this in research.

## **3. Insecure Tool Use (e.g., Plugins, Function Calling)**

• LLMs calling external tools or APIs can be misled to execute unsafe actions.

• High risk when LLMs interface with databases, shells, or payment systems.

## **4. Jailbreaking**

• Bypassing safety filters through adversarial prompts (e.g., DAN-style attacks).

• Continually evolving with community and threat actor creativity.

## **5. Indirect Prompt Injection via Third-Party Content**

• Injection through documents, HTML, emails, or browser input rendered by the LLM.

• Crucial risk in AI assistants and browser-integrated agents.

## **6. Overreliance / Automation Bias**

• Humans defer critical decisions to LLM outputs without verification.

• Dangerous in medical, legal, or operational decision-making.

## **7. Model Theft / Extraction Attacks**

• Query-based extraction of model parameters or weights via API abuse.

• Can lead to IP theft or rehosting of stolen models.

## **8. Model Inversion**

• Reconstructing training samples (e.g., text, images) from embeddings.

• Risk increases with access to internal representations or output probabilities.

## **9. Data Poisoning**

• Malicious training or fine-tuning data introduced to manipulate model behavior.

• Especially dangerous in open-source or continual-learning systems.

## **10. Membership Inference**

• Determining whether a specific data point was used in training.

• Violates privacy and can expose individuals in sensitive datasets.

## **11. Hallucination**

• Fabrication of false or misleading information with high confidence.

• Can damage trust, cause misinformation, or legal liability.

## **12. Supply Chain Risk (Model, Data, Code)**

• Trust issues in base model weights, fine-tuning datasets, and third-party libraries.

• Threat of backdoored models or datasets.

## **13. Inference-Time Backdoors**

• Model behaves normally until triggered by a specific token/pattern.

• Hard to detect through normal testing.

## **14. Cross-Domain Leakage**

• Multi-modal or multi-context systems (e.g., Chat + PDF + Web) may bleed information across contexts.

• Serious risk for confidential or compartmentalized data.

## **15. Resource Exhaustion / DoS**

• Prompt engineering designed to consume excessive computation (e.g., token flooding, recursive prompts).

• Can crash LLM services or spike costs in hosted APIs.

## **16. Alignment Drift**

• Model behavior drifts over time due to updates, context learning, or reinforcement.

• Can bypass previously tested safety constraints.

## **17. Unintended Code Execution**

• LLM outputs malicious code that gets executed automatically (e.g., via Copilot or eval in scripts).

• High risk in dev environments or low-trust pipelines.

## **18. Model Misuse**

• Using LLMs for generating phishing emails, malware, or social engineering scripts.

• Policy, detection, and usage guardrails are essential.

## **19. Overfitting Fine-Tunes**

• Overfit fine-tunes reveal private data or exhibit brittle behavior.

• Often occurs with small or sensitive datasets.

## **20. Legal/Compliance Violations**

• GDPR, HIPAA, and copyright violations due to unvetted data usage or storage.

• Failure to provide explainability, opt-outs, or lawful basis for processing.