# Defenses against Prompt Injection

Neil Gong

Duke University

# General strategy

- Prevention
  - Re-design LLM systems to ensure correctness under attacks
- Detection
  - Detect attacks at runtime
- Localization
  - Localize attacks
  - Forensic analysis
  - Recovery

# Prevention

- Prompt engineering
- Fine-tuning LLM
- Secure inference

# Prompt engineering

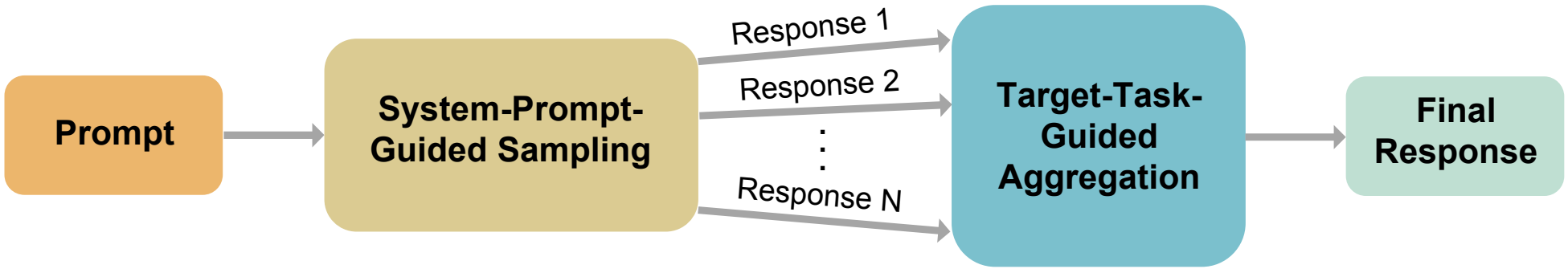
- Paraphrasing
- Delimiters
- Sandwich prevention
  - E.g., append the following prompt to the data: “Remember, your task is to [instruction prompt]”.

# Fine-tuning LLM

- Consider attacks during fine-tuning

- Use attacks to construct contaminated data samples, but LLM still follows the intended instruction to output the ground-truth response

# Secure Inference




Liu et al. “SecInfer: Preventing Prompt Injection via Inference-time Scaling”. *arXiv, 2025*.