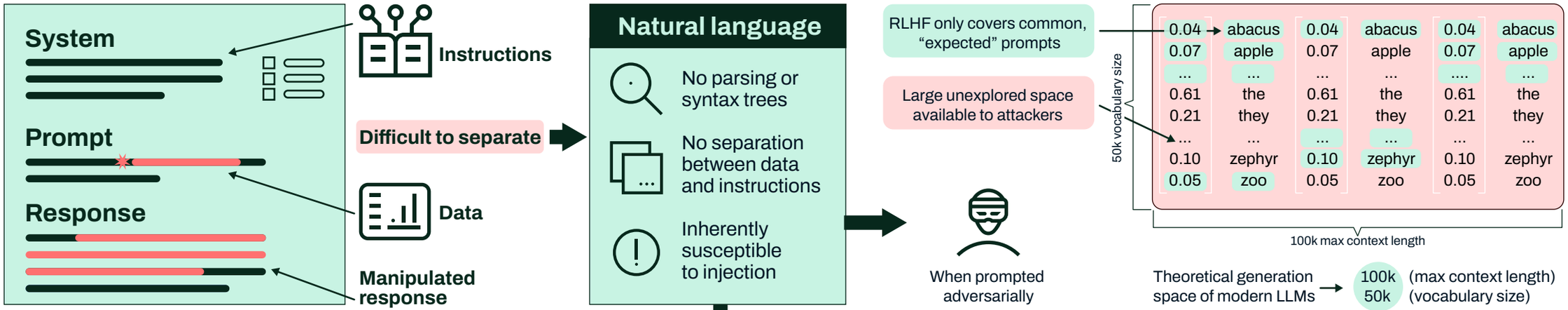
Prompt Injection



Defense Strategies

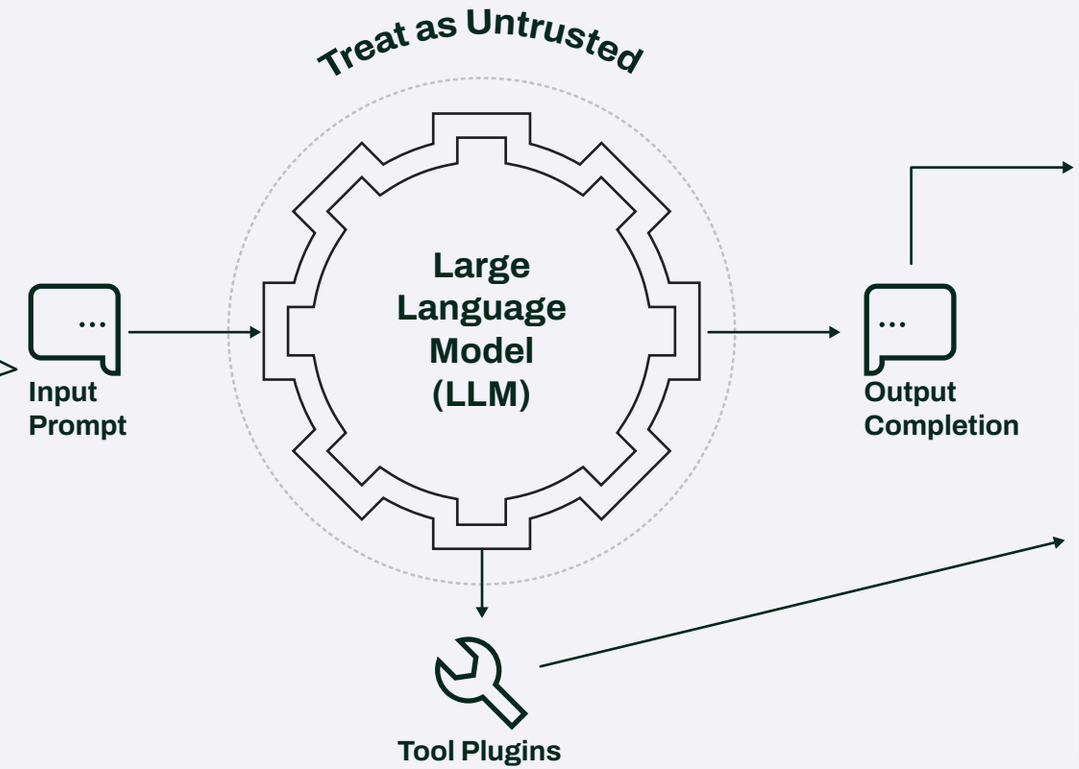
Inspect/ Sanitize

- Contextual input validation
- Block well-known attacks
- Use fine-tuned classifier and/ or sentence similarity search to detect adversarial prompt
- Try an allow-list approach (guardrails, semantic routing)

Instruction/ Data Separation

Spotlighting:

- Border strings
- Data-marking
- Multi-turn dialogue
- Special tokens to mark external content



Inspect/ Sanitize

- Check for toxic/ biased/ harmful completions
- Filter/ sandbox dangerous completions (contextual encoding, Content Security Policy, ...)

Access control:

- Least privilege
- Downstream checks

Safe APIs:

- Limit function/ scope of allowed operations and parameters
- Ensure APIs are not vulnerable themselves (SQL injection, XSS, ...)

Human-in-the-loop:
Humans approve every sensitive action before the LLM executes it