

## LLM Solutions for Lack of Knowledge

### Training / Fine-tuning with the Missing Knowledge

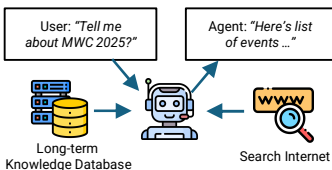


Domain-specific knowledge: "RIS stands for Reconfigurable Intelligent Surface ..."

Latest news and updates: "During the Mobile World Congress 2025 event, ..."

## Agentic Solutions for Lack of Knowledge

### Retrieval-augmented Generation



## LLM Solutions for Computation Capability

### Chain-of-thought or Tree-of-thought Reasoning

"What's wavelength of 300 GHz signal?"



Agent: "1mm"

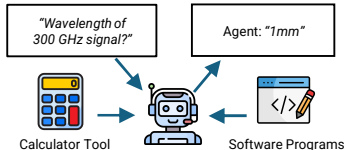
Step 1:  $\lambda = \frac{c}{f}$

Step 2:  $\lambda = \frac{3 \times 10^8}{3 \times 10^{11}}$

Step 3:  $\lambda = 10^{-3} \text{ m}$

## Agentic Solutions for Computation Capability

### Use Tools or Write Programs



## LLM Solutions for Value Alignment

### Training / Fine-tuning with Value Alignment Dataset

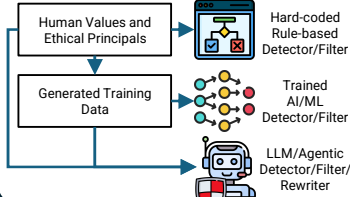


### Activation Engineering to Steer Generation Direction



## Agentic Solutions for Value Alignment

### Input/Output Detectors, Filters & Rewriters



## LLM Solutions for Jailbreak, Backdoor and Injection attacks

- Training / Fine-tuning with Adversarial Alignment Data
- Activation Engineering to Steer Generation Direction

## Agentic Solutions for Jailbreak, Backdoor and Injection attacks

### Input/Output Detectors, Filters & Rewriters

## Solutions for Malicious Tool/Program Runs



**Offline Certification:** formally verify the safety boundary or constraints of tools



**Online Guards:** introduce guardrail rules, Temporal Logic monitors, AI/ML detectors or agents pre/post processors



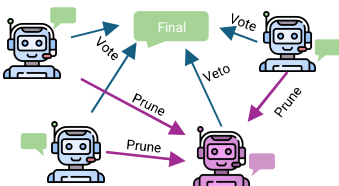
**Online Verification:** verify the tool calls in digital twins



**Online Isolation:** execute tools/programs in an isolated environment with fine-grained data/memory access

## Solutions for Agent Infection

**Multi-agent collaborative defence against harmful responses and malicious agents.**



## Downside of Guardrail Techniques



### Extra Attack Surface

e.g., exploiting overzealous safety mechanism to mark benign user requests as unsafe (false positive) for DoS attack



### Fast but Brittle

Rule-based detectors or filters are fast but can never exhaust all the possibilities



### Limited Transferability

For LLM-based guards, for instance, a guard model trained for LLaMa might not calibrate well for GPT-3.5



### Extra computation and latency

Stacking up guards, especially AI/ML/Agent models consume extra computation resources and incur additional latency.