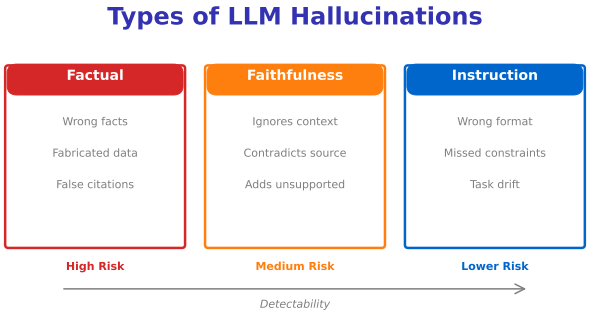
Hallucination Prevention

Week 9: Verification, Grounding, and Factuality

PhD Course in Agentic Artificial Intelligence

Hallucination Types



Factu

hallucinations are highest risk; instruction drift (straying from task) is most common.

Chain-of-Verification Pipeline

Example:

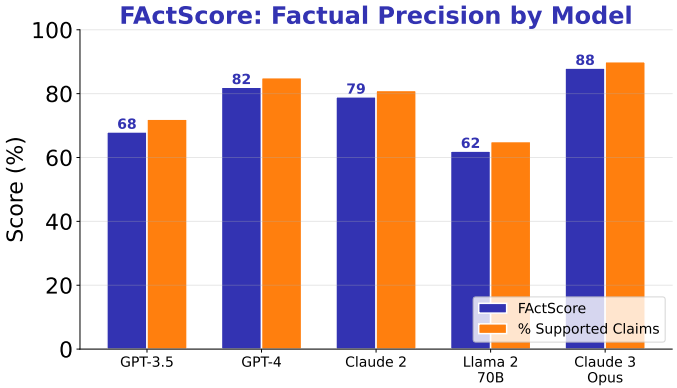
"Paris is capital of France" [Paris, France, capital] "What is France capital?" "Paris" (verified) **Correct!**

Key: Independent verification prevents confirmation bias

Indep

verification prevents confirmation bias (favoring initial beliefs).

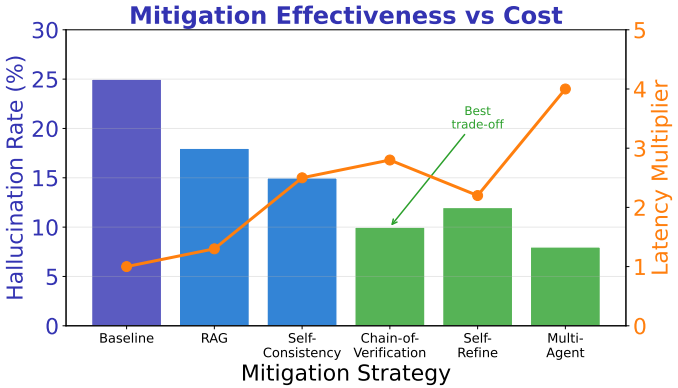
FActScore Evaluation



FActS

measures atomic fact precision against knowledge sources.

Mitigation Strategy Comparison



Chain-of-Verification offers best accuracy/latency trade-off.

Key Papers

This Week

- Ji et al. (2023). "Survey of Hallucination in NLG." arXiv:2202.03629
- Min et al. (2023). "FActScore: Fine-grained Atomic Evaluation." arXiv:2305.14251
- Dhuliawala et al. (2023). "Chain-of-Verification." arXiv:2309.11495

Start with the hallucination survey for taxonomy and scope.

Summary

Key Concepts

- **Types**: Factual, faithfulness, instruction hallucinations
- **Detection**: FActScore, claim decomposition (split into atomic facts), verification
- **Prevention**: Grounding (anchor to sources), self-consistency, multi-agent review

Next Week

- Agent Evaluation and Benchmarking

Preve

¿ Detection ¿ Correction for production systems.