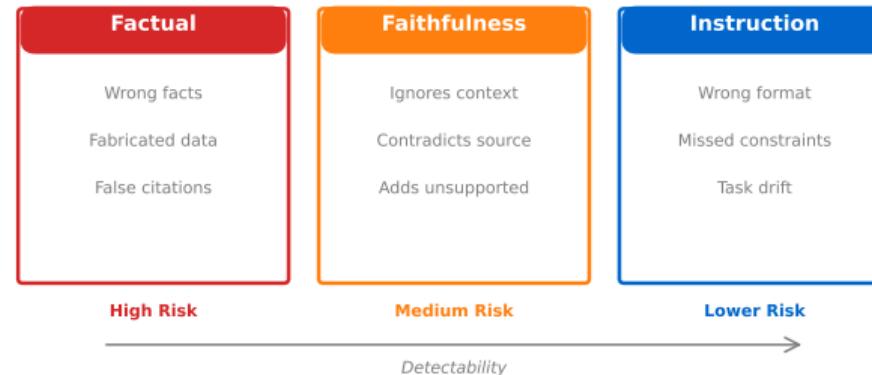


# Hallucination Prevention

Week 9: Verification, Grounding, and Factuality

PhD Course in Agentic Artificial Intelligence

## Types of LLM Hallucinations



hallucinations are highest risk; instruction drift is most common.

Factu

## Chain-of-Verification Pipeline



Example:

"Paris is capital of France"

[Paris, France, capital]

"What is France capital?"

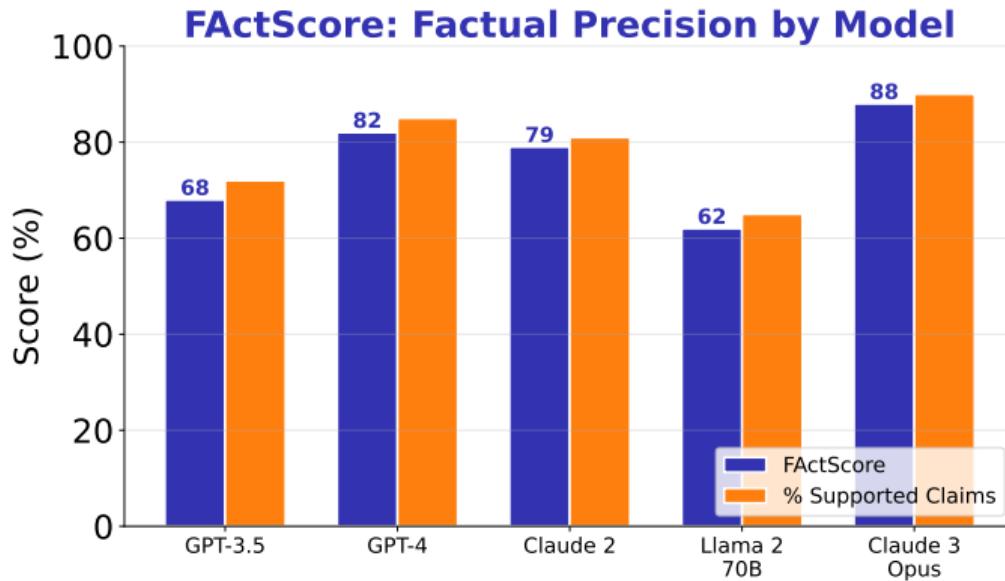
"Paris" (verified)

**Correct!**

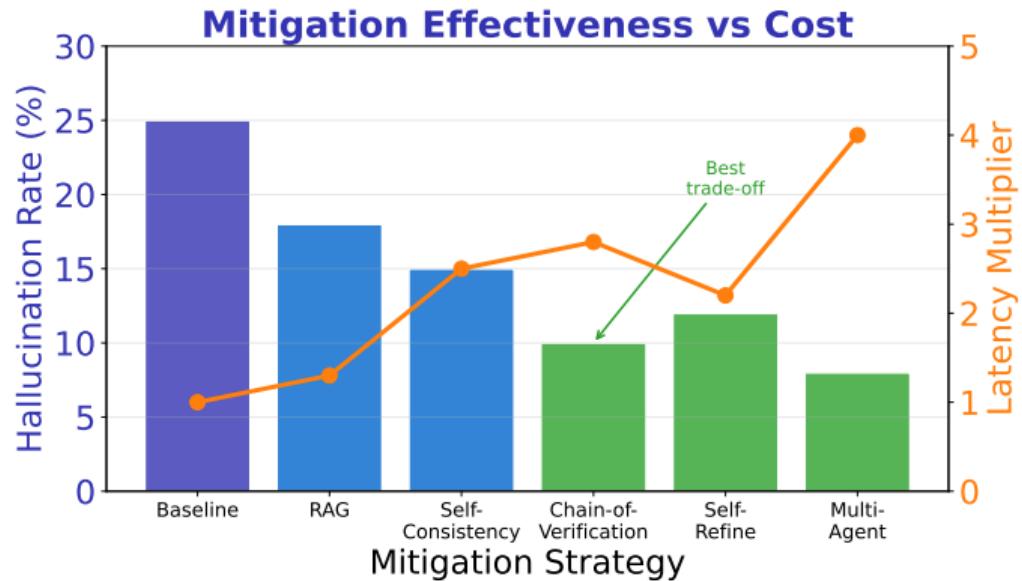
*Key: Independent verification prevents confirmation bias*

verification prevents confirmation bias.

Indep



measures atomic fact precision against knowledge sources.



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

## 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.

## Key Concepts

- **Types:** Factual, faithfulness, instruction hallucinations
- **Detection:** FActScore, claim decomposition, verification
- **Prevention:** Grounding, self-consistency, multi-agent review

## Next Week

- Agent Evaluation and Benchmarking

*i* Detection *i* Correction for production systems.

Preve