Peer-Review Dialogue Examples

Author

We use Adam (lr=1e-3).

Reviewer/LLM

Who's Adam?

"Who's Adam?"

⚠ Hallucination Risk

Reviewer/LLM

Performance increases significantly as model size grows, **opposite to the actual trend observed in the paper.**

Correct Trend

⚠ Trend Reversed

~27k papers

~1M claims

>7M evidence objects

Ground-Truth Paper: Multimodal Evidence

Evidence Retrieval
Paper as Truth Source

E1

...propose Adam, a method for stochastic optimization...

E2

Model	LR	Batch Size	Accuracy
Adam	1e-3	128	92.5%

E3

Figure 2: Training Loss vs. Epochs

E4

$$\theta_{t+1} = \theta_t - \alpha \frac{\cdot m_t}{\sqrt{v_t} + \varepsilon}$$

Evidence Objects (EObjs)

E1 E2 E3 E4

ADAM-Bench Evaluation

ClaimCard

CLAIM: "Adam optimizer with lr=1e-3 achieves 92.5% accuracy."

Evidence: {E2}

Minimal Evidence Sets

Option 1: E2 ✓ OR Option 2: E1 E2 ✓

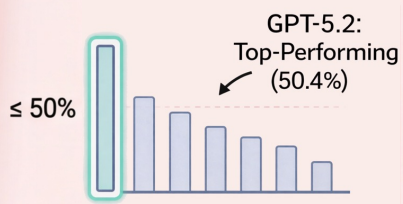
Task A – Hallucination Detection

- ✓ SUPPORTED
- ✗ CONTRADICTED
- ? NOT_FOUND
- UNDECIDABLE

Task B – Minimal Evidence Set Localization

E1 E2 E3 E4

Highlighted Minimal EObj Sets



Benchmark Performance

LLM LMM

LMM > LLM at similar scale

34 LLMs + 10 LMMs