

Detecting AI Mistakes Before They Reach the Client

Fact-Checker & LLM Hallucination Detection for AI tools from McKinsey

McKinsey mentors: Suzana Jacob, Jan Hlousek
MIT faculty advisor: Prof. Donald Sull



Hindy Rossignol



Kevin Luo

01 Problem Statement

GenAI tools are used by companies to accelerate problem-solving and research. Without robust human oversight, AI mistakes can occur with **reputational and financial risks**. As seen in the media where several companies were flagged to have delivered reports containing AI hallucinations. McKinsey wants to prevent any such scenarios before they occur and pair human reviews with an automatic hallucination detection

EXAMPLE OF HALLUCINATION BY PUBLIC AI TOOLS

Source document: "Revenue grew 12% in FY24; profit margin was flat."
AI output: "Profit grew 12% in FY24."
Hallucination Type: Correct number assigned to distorted metrics

Taxonomy of Hallucinations

One Taxonomy set drives synthetic-data injection, judge prompting and evaluation.

Elements mix-up	Content distortion	Content add / loss	Wrong conclusion
Change of logical relationship	Change of quality degree / modality	Content added by the LLM	Wrong conclusion from the source data
Rows / columns / order mismatch	Correct number assigned to distorted metrics	Meaning change due to content loss	Wrong interpretation of an abbreviation
Parties / items / attribute mix-up			

02 Methodology & Deliverables

IMPLEMENTED & ADOPTED · LLM-as-Judge



EXPLORED & BENCHMARKED

Traditional ML classifiers (XGBoost, LogReg, MLP) 64 features → inexpensive but caught only half of the hallucinations	Hughes Hallucination Evaluation Model (HHEM) Fine-tuned cross-encoder → strong recall, but emits no error type	DSPy Automated Prompt Optimization MIPROv2 + GEPA → kept the baseline prompt; data is the bottleneck
--	--	--

DELIVERABLES — two products that integrate in existing workflows without friction

Reviewer web app

Upload a document, get back a UI to quickly review AI outputs

- Hallucinated spans highlighted in place
- Info provided: type, rationale, evidence, fix

**The quoted examples are from public reports that were detected to have AI hallucinations, the hallucinated content was produced by public or opensource models, not McKinsey proprietary tools.*

Agent Skill

Use our Detector inside Claude, Cursor, etc.

Check this report for hallucinations — verify every claim against the web. 2026_Q2_Financial_Report.pdf

I'll run the **detect-rag-hallucination** skill on all 47 pages.

detect-rag-hallucination

skill

✓

Ingested 47 pages · 17,683 words

✓

Decomposed 50 atomic claims

✓

Web-verified every statistic

⚠

3 contradicted · 43 supported

Open report →

- High-stakes findings get an adversarial recheck
- Ships a self-contained report.html

**The quoted examples are from public reports that were detected to have AI hallucinations, the hallucinated content was produced by public or opensource models, not McKinsey proprietary tools.*

03 Results

Detector	Binary detection F1	Error-type classification (macro-F1)
ML classifier (64 features + TF-IDF)	0.52	0.04
HHEM (Vectara cross-encoder)	0.91	0.00
LLM-as-Judge (adopted)	0.99	0.28

43-example expert-labelled held-out set of real hallucinations. All three detectors scored on the identical taxonomy and harness.

Configuration	Detection accuracy
Standalone web app	45%
Agent Skill (Claude / Cursor)	90%

Moving the judge into the agent runtime was the single largest accuracy gain of the project.

100% precision — measured on internal test set
81.3% F1 score — best of three detectors
0 false negatives — catches all mistakes
+6% On type--F1 score from per-claim decomposition

04 Business impact

5 min

saved per source review

2–3 hours

freed per consultant, weekly

38%

of McKinsey are monthly active users

55%

of consultants use Cursor or Claude

Integrated with the 3 largest platforms at McKinsey for applications in AI knowledge capture and research and communication tools

Deployed to McKinsey Intelligence for firmwide user download to uphold integrity of McKinsey's work across the industry