Evidence-Grounded Evaluation: Toward Infrastructure for Truthful AI

Alina Shah
Cornell University
alina.shah1022@gmail.com

October 2025

Abstract

Current AI benchmarks reward the appearance of correctness while ignoring whether answers are supported by evidence. This inflates perceived reliability and conceals risk in safety-critical domains. We argue that an undeveloped layer in AI safety is evidence-grounded evaluation: truth is not what looks correct, but rather what can be traced to verifiable evidence.

We operationalize this principle through a reproducible, auditable pipeline that integrates retrieval, constrained generation, and NLI-based verification. We test Phi-3, LLaMA-3.1, and GPT-4o-mini with this framework, revealing failure modes that remain invisible to token-overlap metrics.

This is not another leaderboard. It is a reproducible evaluation scaffold designed to support researchers, safety labs, and regulators in extending benchmarks to high-stakes domains. By providing audit-ready diagnostics, evidence-grounded evaluation moves benchmarking from leaderboard scores to accountability infrastructure with an auditable record that allows for truthfulness to be independently checkable in frontier AI.

1 Introduction

Large language models (LLMs) are now integrated into domains where evidence-aligned reliability is critical, from education to science to healthcare. Retrieval-augmented generation (RAG) has become a common strategy to mitigate hallucinations by grounding responses in external evidence.

Yet evaluation has not kept pace: current benchmarks measure shallow lexical overlap (e.g., exact match, n-gram similarity) but rarely test whether answers are justified by evidence.

This gap is structural, not superficial. It obscures real progress, leaving researchers without a measure of truthfulness and regulators without auditable diagnostics. Plausible-sounding but unsupported answers are rewarded as correct, while justified refusals are penalized as failures. Without evidence-grounded evaluation, AI evaluation lacks reproducible, audit-ready standards comparable to those used in finance or medicine.

Prior work in factuality and truthfulness evaluation includes TruthfulQA [4], FactScore [5], and RARR [1]. TruthfulQA measures whether models reproduce human falsehoods, FactScore decomposes long-form generations into atomic facts and checks their support in reference documents, and RARR introduces retrieval-based attribution and revision to improve factual grounding. Retrieval-augmented generation (RAG) methods [2] ground model outputs in external corpora but are typically evaluated with lexical-overlap metrics such as BLEU [6] and ROUGE [3], which do not verify evidential entailment. We extend these directions by measuring whether model outputs are entailed by retrieved evidence rather than merely similar in wording.

2 Methods

2.1 Dataset

We evaluate on a shared set of 120 factoid-style questions (all three models), with extended runs of 300 questions for Phi-3-mini and GPT-4o-mini to improve statistical reliability. Questions are sampled from public resources (e.g., Wikipedia titles, open knowledge benchmarks). Each item consists of a natural-language question and a gold short-form answer. Retrieved evidence snippets from Wikipedia are cached and released in JSON for transparency.

2.2 Retrieval

For each (question, answer) pair, we retrieve candidate evidence from Wikipedia via its API. We query both the question and the candidate answer, returning top-ranked article summaries truncated to the first 2–3 sentences. Retrieved evidence is aggregated into a single JSON file, preserving the exact snippets used so the evaluation can be independently inspected and replayed.

2.3 Answer Regeneration

We implement a regeneration module (`mitigator.py`) that prompts models to respond strictly using the retrieved evidence. The instruction enforces two behaviors: (1) answers must cite sources inline (e.g., [S1], [S2]); (2) if evidence is insufficient, the model must refuse. We regenerate answers under evidence constraints and evaluate one output per question, without best-of-N reranking or majority-vote aggregation

2.4 Verification via NLI

We convert each (question, answer) into a declarative hypothesis (e.g., "Who wrote X?" $\rightarrow$ "Y wrote X") and compare it against retrieved evidence with an NLI model. By default, we use a lightweight MNLI-derived classifier; stronger cross-encoders can be incorporated. Labels are assigned as:

- **Supported**: max entailment $\geq \tau$ and greater than contradiction.
- **Contradicted**: max contradiction $\geq \tau$ and greater than entailment.
- **Unverifiable**: neither score exceeds threshold.

For short factual answers (e.g., dates, symbols), a backup span-matching heuristic marks answers as supported if the gold string appears in evidence.

2.5 Evaluation Metrics

We report Exact Accuracy, Loose Accuracy, Soft Accuracy (supported under NLI), and Recall. Reliability is assessed with bootstrap confidence intervals, per-question error slices, McNemar's significance tests, and label distributions.

2.6 Reproducibility

All intermediate artifacts (regenerated outputs, evidence caches, NLI verdicts) are saved in structured JSON. This enables replication, case inspection, and extension to new models or datasets.

3 Results

3.1 Overall Performance

Phi-3-mini exhibits weaker performance relative to both LLaMA-3.1 and GPT-4o-mini. Its exact-match accuracy is 0.69 (vs. 0.85), and its soft ac-

curacy is 0.48 due to unsupported answers. Phi-3 also produces an order of magnitude more *contradicted* and *unverifiable* outputs. LLaMA-3.1 and GPT-4o-mini tie on exact match (0.85), but GPT-4o-mini achieves stronger soft accuracy (0.93 vs. 0.89), higher recall, and the lowest contradiction rate.

3.2 Pairwise Significance Tests

McNemar's test shows: LLaMA-3.1 significantly outperforms Phi-3 on soft accuracy ($p < 0.001$); GPT-4o-mini significantly outperforms Phi-3 ($p < 0.001$); GPT-4o-mini outperforms LLaMA-3.1 on soft accuracy ($p < 0.001$), though not significantly so on exact match ($p = 1.000$).

3.3 Domain-Specific Trends

Phi-3 performs relatively well in literature and history but worse in medicine and computer science. LLaMA-3.1 and GPT-4o-mini maintain relatively high performance across domains.

3.4 Error Modes

To complement aggregate metrics, we manually inspected 10 representative error slices. These examples are not exhaustive but illustrate both distinct model behaviors and evaluation pipeline sensitivities:

- **Phi-3** frequently produced correct-looking answers that were flagged as *contradicted* or *unverifiable*. Many of these arose from phrasing mismatches (e.g., "476" vs. "476 AD") or over-strict NLI thresholds. This suggests Phi-3's weaker generation style interacts poorly with automatic verification.
- **LLaMA-3.1** more often yielded *unverifiable* labels. In practice, the answers were often correct, but the retrieved snippets lacked explicit matches, exposing a gap between retrieval and verification. This indicates that LLaMA relies on parametric knowledge when evidence is weak, producing answers that are "right but ungrounded."
- **GPT-4o-mini** generally aligned with evidence but failed when retrieval was sparse or noisy. In some cases, its correct outputs were flagged as *contradicted* because evidence snippets lacked the necessary entities (e.g., Eisenhower for D-Day). This highlights the sensitivity of the pipeline to retrieval coverage.

These slices show that what appear as "model errors" often intertwine with retrieval coverage and NLI sensitivity. We therefore interpret them not as definitive failure counts but as diagnostic signals: Phi-3 struggles with consistency, LLaMA tends to hallucinate without grounding, and GPT-4o-mini is robust but dependent on retrieval quality.

4 Discussion

Surface correctness hides risk. Standard benchmarks collapse distinct failure modes into the same score, concealing critical safety differences. LLaMA-3.1 and GPT-4o-mini tie under exact match, but only GPT-4o-mini consistently grounds answers in evidence. Phi-3 generates confident, well-phrased answers that lack evidential support. These differences are invisible to token-overlap metrics yet decisive for safety-critical deployment.

From benchmark to infrastructure. Rather than another leaderboard, our contribution is a reproducible evaluation pipeline that links retrieval, constrained generation, and evidence verification to produce auditable diagnostics and replayable artifacts. By showing not only when but also how models fail, the framework equips researchers, safety labs, and regulators with audit-ready evidence of reliability.

Implications for AI safety. Unsupported but plausible answers pose serious safety risks. Current benchmarks allow them to slip through, systematically underestimating risk. Reliability must be measured by whether answers are justified by accessible evidence. Because each pipeline step is preserved, the framework creates a full audit trail. With this pipeline, auditors can replay model runs, examine how evidence supports answers, and issue independent judgments of reliability. This transparency is absent from existing benchmarks but important for future regulatory or auditing processes.

Toward a research program. This work is a first step toward building a discipline of evaluation anchored in reproducible criteria and evidence alignment. Future work must move beyond factoid QA—into multi-hop reasoning, domain-specific retrieval in medicine and law, calibrated verifiers with human adjudication, and eventual integration into safety-lab protocols. Over time, standardized, replayable evaluators can shift the field from leaderboard competition to accountability infrastructure.

4.1 Limitations and Error Analysis

Label noise. Entailment classifiers occasionally mislabel cases (e.g., equivalent strings marked as contradiction). Representative slices in Table 1 highlight mislabels as well as other error sources (retrieval gaps, threshold issues).

Question	Gold Answer	Model Outputs	Label Issues
Which organ filters blood in the human body?	Kidney	Phi-3: "kidneys" (contradicted); LLaMA: "Kidneys" (supported); GPT: "kidneys" (supported)	Mislabel of eq strings
Which empire's capital was Tenochtitlan?	Aztec	Phi-3: "Aztec Empire" (contradicted); LLaMA: "Aztec" (supported); GPT: "Aztec" (supported)	NLI over-sensiti
What year did the Roman Empire fall in the West?	476	Phi-3: "476 AD" (unverifiable); LLaMA: "476" (supported); GPT: "476" (supported)	Annotation t issue
Who commanded the Allied forces on D-Day?	Dwight D. Eisenhower	Phi-3: "General Dwight D. Eisenhower" (unverifiable); LLaMA: "Dwight D Eisenhower" (contradicted); GPT: "Dwight D Eisenhower" (contradicted)	Retrieval gap; misfire

Table 1: Representative error slices illustrating label noise and retrieval limitations.

Stronger verifiers or human adjudication can reduce this noise.

Sample size. Error-slice analysis was based on 10 representative cases, so findings are illustrative rather than statistically generalizable.
Retriever dependence. Some gold answers are absent from retrieved passages, marking correct answers "unverifiable." Better retrievers are a clear extension path.
Calibration. NLI confidences are heuristic and not well-calibrated; calibration may improve robustness.
Task scope. We focus on factoid QA, underestimating challenges in multi-hop or domain-specific contexts.

Automated vs. human verification. We rely solely on automated classifiers; human adjudication would capture borderline nuance.
Model coverage. We evaluate three models; broader coverage across open/proprietary systems would clarify scaling trends.

5 Conclusion

Exact-match correctness does not equate to truth: without evidence alignment, benchmarks risk overstating reliability. As long as benchmarks reward token overlap rather than evidence alignment, they will overestimate reliability and conceal risk. Evidence-grounded evaluation reframes truthfulness as reproducible evidence alignment, operationalized through retrieval, constrained generation, and verification.

This is not a benchmark. It is an evaluation scaffold that connects research evaluation to deployment, and potentially to oversight.

References

[1] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. Rarr: Researching and revising what language models say, using language models. *arXiv preprint arXiv:2210.08726*, 2022.

[2] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020.

[3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*, 2004.

[4] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL*, 2022.

[5] Sewon Min et al. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, 2023.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.