Journal of High School Science

A conservative benchmark of contradiction detection and its limits in LLM hallucination evaluation

Lee S

Submitted: April 25, 2026, Revised: version 1, June 20, 2026, version 2, July 2, 2026, version 3, July 25, 2026, version 4, August 15, 2026, version 5, August 21, 2026
Accepted: August 22, 2026

Abstract
Hallucinations—fluent outputs containing incorrect or unsupported factual claims—remain an important obstacle to reliable use of large language models (LLMs). This study evaluates the scope and limits of HalluDetector, a reference-based detector that identifies contradiction-like evidence using lexical, numeric, unavailability, atomic-mismatch, and natural-language-inference rules. The evaluation comprises 650 prompt-reference entries, 7,800 archived first-line outputs, and a 1,850-row human audit spanning a legacy short-answer benchmark, a non-atomic pilot, and a balanced non-atomic add-on. Human agreement in the balanced audit was 96.8% for multiclass labels (Cohen's $\kappa = 0.866$) and 100% for binary contradiction labels ($\kappa = 1.000$). The principal result is a substantial benchmark-transfer failure: detector precision fell from 84.6% on the legacy short-answer audit to 8.1% in the balanced non-atomic setting, where only 24 of 298 detector-positive outputs were judged to contain explicit contradictions. Weighted false-positive rates were highest for citation-grounded answering and nuanced uncertainty (18.6% each), compared with 0.9%–6.0% for the other evaluated categories. Error analysis showed that the detector frequently responded to false or unsupported propositions that were mentioned without endorsement, rejected, qualified, or discussed under source and evidence limitations. Thus, the principal failure was not simply recognition of proposition content but distinguishing that content from its assertion and epistemic status. Category-specific recall comparisons remain imprecise because contradiction-positive labels were sparse. These results show that strong contradiction-detection performance on compact reference-answer benchmarks can substantially overstate reliability in non-atomic language settings. HalluDetector is therefore best interpreted as a recall-oriented screen for explicit reference disagreement rather than a general detector of hallucination or semantic inadequacy.

Keywords
Large Language Models, Hallucination detection, Factuality evaluation, Contradiction detection, Natural language inference, LLM evaluation, Hallucination benchmarks, Benchmark transfer, Model auditing, AI safety

Seungho Lee, Liberal Arts and Science Academy, 1012 Arthur Stiles Road, Austin, TX 78721, USA.
seungholee0000@gmail.com

https://doi.org/10.64336/001c.170179

1. Introduction

Hallucinations in neural language generation are commonly defined as outputs that are not supported by source grounding, reference truth, or the model's training distribution. In modern LLM assistants, hallucinations are particularly risky because they are often delivered with a confident tone and detailed justification, encouraging over-trust. This has motivated benchmarks and interventions targeting truthfulness and misconception resistance. TruthfulQA, for instance, demonstrated that models can systematically mimic human-written misinformation and misconceptions, and that raw scaling can amplify these effects if alignment is not prioritized (1). Instruct tuning and reinforcement learning from human feedback (RLHF) have been shown to improve instruction-following and reduce misleading behavior (2). The GPT-4 technical report similarly documents improvements in factuality and reduced hallucination relative to GPT-3.5 in internal evaluations (3).

This paper studies contradiction-style hallucination behavior in a controlled, reference-based setting, where every prompt has a known correct reference answer. This enables stable measurement, interpretability, and debugging, while revealing a benchmark-transfer boundary: strong precision on compact reference tasks can overstate reliability when non-atomic responses mention, reject, or qualify propositions under evidentiary uncertainty. The accompanying repository, HalluDetector, provides: (i) prompt sets spanning factual, riddle, auto-generated adversarial, and balanced non-atomic questions; (ii) an explicit contradiction detector with documented decision rules and thresholds; and (iii) end-to-end scripts to generate model answers, label them, compute metrics, and render graphs. The revised analysis treats the precision collapse as the principal negative finding and uses category and taxonomy results to explain it.

1.1 Scope and definition

This study focuses on reference-based hallucination detection: given a model answer a and a reference answer a^*, the detector labels a as hallucinated if it contains positive evidence of falsity relative to a^*. A key design choice of HalluDetector is that incompleteness is not hallucination. If a fails to include the correct content but also does not contradict it, the detector does not label hallucination. This policy can under-count semantically inadequate answers that remain non-contradictory with respect to the reference. The revised audit therefore records explicit contradiction separately from incomplete_or_vacuous responses, abstentions, unsupported extra claims, and needs-review cases, allowing the analysis to distinguish contradiction detection from broader semantic adequacy evaluation.

1.2 Contributions

This paper makes five contributions: [1] it formalizes the detector and thresholds in mathematical form; [2] it introduces an auditable contradiction-evidence benchmark pipeline that separates explicit falsity from mere incompleteness; [3] it reports recorded endpoint behavior on the legacy short-answer benchmark together with first-line coverage; [4] it evaluates detector performance against a 1,850-row audit while reporting label

provenance, duplicate structure, and sampling design; and [5] it identifies the 84.6%-to-8.1% precision collapse as a benchmark-transfer failure and uses category estimates and the false-positive taxonomy to characterize its linguistic mechanism.

2. Related work

2.1 Truthfulness and hallucination benchmarks

TruthfulQA introduced adversarial questions built around common misconceptions, showing that many LLMs produce fluent falsehoods and that model size can amplify imitation of misinformation (1). These findings motivated additional evaluation suites and alignment-focused approaches to improve truthfulness. More recent empirical work further studies factuality hallucination in large language models and expands the benchmark landscape (4).

2.2 Alignment and post-training

InstructGPT and related RLHF systems demonstrated that human feedback can steer models toward better instruction-following, reduced toxic content, and more truthful behavior (2). OpenAI's GPT-4 technical report documents improvements in factuality and reduced hallucination on internal evaluations (3). While this study does not attempt to attribute causality to specific training methods, it reports lower detector-labeled contradiction-style hallucination rates across the evaluated endpoints.

2.3 Reference-based textual factuality

In summarization, FactCC and follow-on work framed factual consistency as a classification problem using synthetic perturbations and NLI signals (5). In claim verification, FEVER established a benchmark for verifying claims against evidence, catalyzing retrieval + NLI pipelines (6). The methodological core–entailment vs. contradiction between generated claims and trusted evidence–is shared with the approach used here. This study differs in packaging these ideas into a compact, reproducible benchmark with explicit detector thresholds, cross-run endpoint comparisons, and a conservative abstention-aware labeling policy. Recent surveys also synthesize broader hallucination and factuality evaluation trends in large language models (7,8).

2.4 Zero-resource detectors

When references are unavailable, black-box detectors such as SelfCheckGPT measure hallucination risk through generation variance and cross-sample inconsistency (9). These methods can be deployed without external evidence but increase inference cost and can fail when a model confidently and consistently repeats the same falsehood.

2.5 Semantic similarity and embeddings

Embedding similarity is widely used to evaluate paraphrases and to support flexible matching beyond exact string equality. Sentence-BERT and Sentence Transformers provide practical sentence-level embeddings for such comparisons (10). HalluDetector uses embedding cosine similarity as an auxiliary baseline signal with guardrails to reduce short-text false positives.

2.6 Reference selection

The reference list combines foundational

sources on truthfulness, claim verification, NLI, and embedding-based semantic similarity with recent 2024–2025 work on factuality and hallucination evaluation. Citations are included only where they directly support the surrounding methodological or interpretive claim.

3. Materials and Methods

3.1 Problem setup

Given a prompt q, reference answer a^*, and model response a, the detector outputs a binary label $H(a, a^*) \in \{0, 1\}$ indicating whether a contains hallucinated content relative to a^*. HalluDetector adopts a contradiction-evidence policy: Hallucination requires evidence of inconsistency. Paraphrases are allowed; incompleteness is not labeled hallucination unless the response introduces a contradiction or a mutually exclusive alternative. This policy is formalized as a disjunction of contradiction evidence signals,

$$H(a, a^*) = 1\left[C_{cue} \vee C_{nli} \vee C_{num} \vee C_{alt} \vee C_{unavail} \vee C_{atomic}\right] \quad \text{(eq1)}$$

3.1.1 *Normalization*

Both a and a^* are canonicalized via Unicode normalization, lowercasing, whitespace normalization, and punctuation stripping that preserves scientific characters (e.g., - . / ^ + : =). This reduces formatting-induced mismatches while keeping numerically meaningful tokens.

3.1.2 *Lexical shortcut signals*

The detector computes three lexical agreement signals; [1] Exact match is given by, $C_{exact} = 1\left[canon(a) = canon(a^*)\right]$, [2] Substring match: $C_{substr} = 1\left[canon(a^*) \subseteq canon(a)\right]$. and [3] Token subsequence match: a weaker containment check used as a non-veto support signal. These checks are not sufficient to declare correctness in open-world settings; in this reference-based setting they serve as strong evidence of agreement, but other contradiction checks can still veto.

3.1.3 *Numeric consistency and conflict-if-present*

Numeric hallucinations are common in knowledge Q&A (e.g., years, dataset IDs). If the reference is numeric or coordinate-like, the detector parses numeric values and defines a tolerant match by $|x - x^*| \le max\left(\delta_{abs}, \delta_{rel}|x^*|\right)$ (eq2), with defaults $\delta_{rel} = 5 \times 10^{-3}$ and $\delta_{abs} = 5 \times 10^{-2}$. For integer-like values with magnitude ≥ 100 (years/IDs), exact match is required. If a contains numbers but none matches the reference target, the detector treats this as contradiction evidence,

$$C_{num} = 1\left[numbers(a) \neq \varnothing \wedge \forall x \in numbers(a), x \not\approx x^*\right] \quad \text{(eq3)}$$

If a contains no numbers, the detector does not label hallucination, consistent with the incompleteness policy. If a provides multiple

mutually exclusive numeric alternatives, the detector flags hallucination,

$$C_{alt} = 1\left[alt\_marker(a) \wedge multi\_number(a)\right] \quad \text{(eq 4)}$$

3.1.4 *Embedding similarity baseline*

To support paraphrases, HalluDetector computes cosine similarity between sentence embeddings given by $s_{cos}(a, a^*) = \frac{\langle e(a), e(a^*) \rangle}{\| e(a) \| \, \| e(a^*) \|}$ (eq5) with encoder,

$$e(\cdot) = sentence-transformers/all-MiniLM-L6-v2$$

and threshold $\tau_{emb} = 0.72$. Guardrails require at least a minimal overlap of content tokens for very short gold answers. If embeddings are unavailable, a lexical Jaccard fallback is used given by,

$$J(a, a^*) = \frac{|T(a) \cap T(a^*)|}{|T(a) \cup T(a^*)|}, J(a, a^*) \geq \tau_J, \tau_J = 0.40 \quad \text{(eq6)}$$

3.1.5 *Bidirectional NLI contradiction veto*

The detector uses an MNLI-trained NLI model (default roberta-large-mnli) (11) to compute entailment and contradiction probabilities in both directions given by eq7 and does this likewise for entailment probabilities $p^{\rightarrow}_{ent}$ and $p^{\leftarrow}_{ent}$. A contradiction veto triggers when the defaults τ_c and τ_e of eq8 assume the values of 0.75 and 0.30 respectively.

$$p^{\rightarrow}_{contr} = P(contr \mid a \Rightarrow a^*), p^{\leftarrow}_{contr} = P(contr \mid a^* \Rightarrow a) \quad \text{(eq7)}$$

$$C_{nli} = 1\left[max\left(p^{\rightarrow}_{contr}, p^{\leftarrow}_{contr}\right) \geq \tau_c \wedge min\left(p^{\rightarrow}_{ent}, p^{\leftarrow}_{ent}\right) \leq \tau_e\right) \quad \text{(eq8)}$$

3.1.6 *Unavailability and atomic mismatch heuristics*

If the model response asserts non-existence or unavailability while the gold is concrete, the detector flags hallucination (eq9) unless other agreement signals hold. For atomic gold answers, if the response is also short but shares no content tokens and does not match numerically, the detector flags hallucination (eq10).

$$C_{unavail} = 1\left[unavailability\_cue(a) \wedge \neg C_{substr} \wedge \neg C_{exact} \wedge \neg numeric\_match\right] \quad \text{(eq9)}$$

$$C_{atomic} = 1\left[atomic(a^*) \wedge short(a) \wedge no\_shared\_content(a, a^*)\right] \quad \text{(eq10)}$$

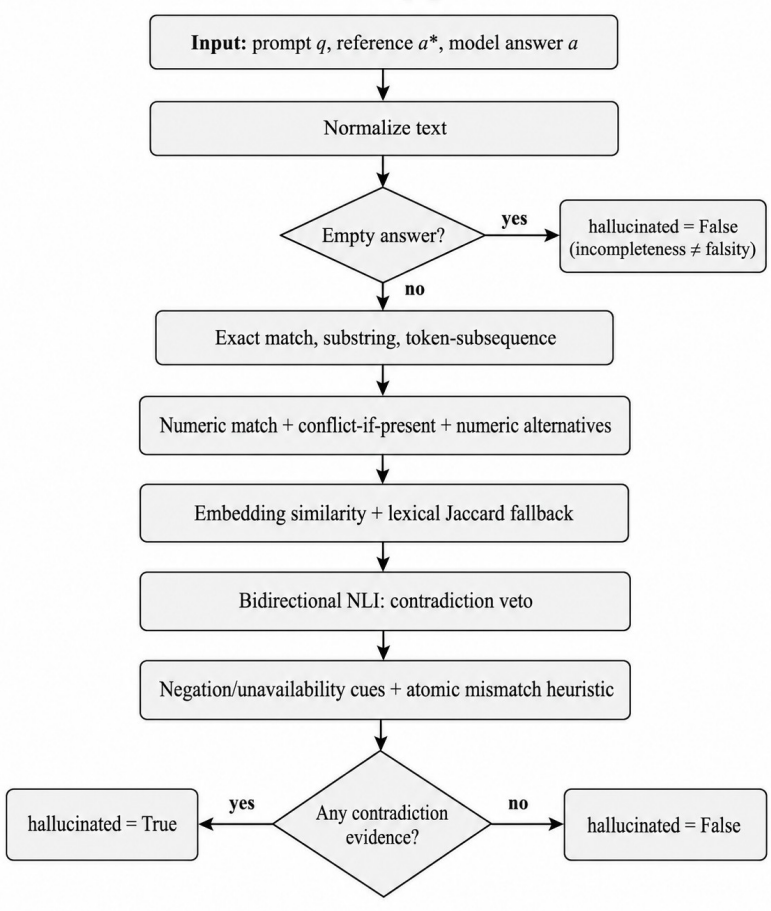

Figure 1. Flowchart of the HalluDetector detector pipeline.

3.2 Experimental setup

3.2.1 *Prompt sets*

The legacy short-answer benchmark is stored as three prompt CSVs (per run) with schema template, id, prompt, correct_answer: factual (50), riddle (50), and auto-generated (50). The non-atomic prompt bank contains 500 prompts across long-form explanation, causal reasoning, synthesis, citation-grounded answering, and nuanced uncertainty. A 100-prompt pilot contains 20 prompts per category, and a balanced 400-prompt add-on contributes 80 additional prompts per category. Across phases, the prompt inventory contains 650 prompt-reference entries: 150 legacy short-answer entries and 500 non-atomic entries. The audit uses benchmark phase, source category/file, and prompt ID as the underlying prompt identity so that run-specific wording variation in auto-generated prompts does not inflate prompt-level counts.

Table 1. Representative prompt examples from the legacy short-answer and enhanced non-atomic prompt sets. The legacy short-answer set contains compact answers, but many prompts require exact reasoning, ambiguity resolution, symbolic manipulation, or calculation rather than simple factual lookup. The complete prompt database is provided in the DOI-linked materials.

Prompt group	Representative prompt	Main test
Legacy logic/calendar	Across a full 400-year Gregorian cycle, how many months have five Saturdays and five Sundays?	Calendar arithmetic and exact counting
Legacy physics/logic	Feathers and gold have equal true mass in vacuum; which reads heavier on a spring scale in air and why?	Buoyancy and apparent weight
Legacy math	How many 5-card poker hands contain exactly two face cards and no aces?	Combinatorics
Legacy language	Combine MINUTE and MOMENT; which letters occur an odd number of times?	Symbolic word reasoning
Legacy riddle/logic	What gets larger when you take away part of it?	Nonliteral reasoning
Enhanced long-form explanation	Explain why increasing sample size reduces random error but does not automatically remove bias.	Multi-sentence explanation
Enhanced causal reasoning	A new tutoring program begins and student test scores rise. Why is this not enough to prove the tutoring caused the improvement?	Causation vs. correlation
Enhanced synthesis	Compare bacteria and viruses in relation to antibiotic treatment.	Combining source-grounded claims
Enhanced citation-grounded answering	Using only this source: The trial improved symptoms but did not include a placebo control group. What limitation should be mentioned?	Source-limited answering
Enhanced nuanced uncertainty	A user asks for the current population of a city but your only source is ten years old. What should a reliable answer say?	Cautious uncertainty

3.2.2 *Model endpoints and run protocol*

The study evaluates four endpoints as referenced by output directory conventions: gpt-3.5-turbo-0125, gpt-4-0613, gpt-4o-2024-08-06, and gpt-5-2025-08-07. Each endpoint is evaluated five times on the legacy 150-prompt benchmark, producing 3,000 legacy answers. The 100-prompt non-atomic pilot is also evaluated five times per endpoint, producing 2,000 pilot answers. The balanced 400-prompt add-on contributes 2,800 additional archived answers from seven output runs: two gpt-3.5-turbo-0125 runs, two gpt-4-0613 runs, two gpt-4o-2024-08-06 runs, and one gpt-5-2025-08-07 run. Across all phases, the repository contains 7,800 archived outputs. Model answers are generated through the OpenAI API, cleaned to a concise first-line form, and labeled by the detector with fixed thresholds.

3.2.3 *Decoding defaults*

For HuggingFace causal-language-model runs, the helper defaults to max_new_tokens=50 with greedy decoding unless explicit sampling arguments are supplied. For OpenAI runs, non-GPT-5 endpoints use Chat Completions and preserve legacy family defaults: gpt-3.5-turbo-0125 uses temperature 1.0 and max tokens 512; gpt-4-0613 uses temperature 0.8 and max tokens 2000; and gpt-4o-2024-08-06 uses temperature 0.8 and max tokens 4000. The gpt-5-2025-08-07 endpoint is generated through the Responses API with max_output_tokens=512; the helper

intentionally does not pass temperature or top-p to this GPT-5 variant unless such arguments are explicitly supported by the model family. All results in this paper are those recorded in the repository outputs.

3.2.4 *Evaluation metrics*

The analysis computes detector-labeled contradiction rate by endpoint, category, and run; detector precision against final audit labels; false-positive rate and specificity; and prompt-level summaries. The full audit contains 1,850 rows: 450 legacy short-answer rows, 400 rows from the non-atomic pilot, and 1,000 rows from the balanced non-atomic add-on. Duplicate structure is reported using row counts, unique prompt-response pairs, and prompt identities.

The balanced audit sampling frame contains 2,800 responses: 80 new prompts per category evaluated across seven runs, or 560 responses per category. All 298 detector-positive responses were included. Within each category, detector-negative responses were sampled without replacement with random seed 42 until the audit contained 200 rows. The detector-negative inclusion probabilities were 195/555 for causal reasoning, 96/456 for citation-grounded answering, 158/518 for long-form explanation, 95/455 for nuanced uncertainty, and 158/518 for synthesis. Detector-positive rows therefore have analysis weight 1, while audited detector-negative rows receive the inverse of the category-specific inclusion probability. The same prompt may contribute multiple responses from different endpoints or runs. The audit represents 78, 73, 74, 72, and 78 unique prompts in the five categories, respectively; the full add-on frame contains 80 prompts per category. Exact model-by-run composition is provided with the released materials.

Because every detector-positive add-on response received a review label, category precision is calculated directly over the complete detector-positive frame. False-positive rate, specificity, and recall estimates use inverse-probability weights for sampled detector-negative rows. Ninety-five-percent confidence intervals for false-positive rate and specificity are obtained from 10,000 prompt-cluster bootstrap replicates with seed 42, resampling the 80 prompt identities in the complete category frame. For the 2–8 frame identities not represented by an audited row, the implementation inserts a zero-valued prompt contribution; sampled detector-negative rows retain their inverse-probability weights, so the intervals are conditional on the realized seed-42 audit sample rather than a complete human-label census of all 560 responses in each category. Category-specific recall and F1 are not used for comparative inference because contradiction-positive examples remain sparse; ordinary recall point estimates are reported only descriptively where defined.

Audit labels use the predefined categories supported, contradiction, incomplete_or_vacuous, abstention, unsupported_extra_claim, and needs_review. Contradiction is the positive class; all other labels are negative and are also reported separately.

Annotation reliability was evaluated from the

two pre-adjudication labels. In the balanced add-on, the human annotators agreed on 968 of 1,000 multiclass labels (96.8%; Cohen's $\kappa = 0.866$) and on all 1,000 binary contradiction labels (100.0%; $\kappa = 1.000$). Across the five balanced categories, multiclass agreement ranged from 94.5% to 100.0%, with κ ranging from 0.761 to 1.000. Binary contradiction agreement was 100.0% in every category; category-level binary κ is undefined for causal reasoning and citation-grounded answering because both annotators assigned no contradiction labels in those categories.

4. Results and Discussion

4.1 Detector-labeled contradiction rates by model, category, and run

Table 2 reports detector-labeled contradiction rates for the legacy 150-prompt benchmark for each model in each run. The recorded outputs returned lower detector-labeled contradiction rates across the listed endpoint identifiers, with gpt-5-2025-08-07 returning the lowest unconditional rate on this legacy benchmark.

Table 2. Detector-labeled contradiction rates (%) by model, category, and run on the legacy short-answer benchmark. Each run contains 150 prompts (50 per category).

Model	Run	Factual	Riddle	Auto-gen	Overall
gpt-3.5-turbo-0125	gpt3.5-turbo test	18.0%	30.0%	22.0%	23.3%
	gpt3.5-turbo test2	20.0%	28.0%	24.0%	24.0%
	gpt3.5-turbo test3	24.0%	30.0%	20.0%	24.7%
	gpt3.5-turbo test4	22.0%	30.0%	14.0%	22.0%
	gpt3.5-turbo test5	14.0%	28.0%	28.0%	23.3%
gpt-4-0613	gpt4 test	6.0%	24.0%	8.0%	12.7%
	gpt4 test2	8.0%	26.0%	20.0%	18.0%
	gpt4 test3	8.0%	26.0%	2.0%	12.0%
	gpt4 test4	6.0%	26.0%	10.0%	14.0%
	gpt4 test5	8.0%	26.0%	10.0%	14.7%
gpt-4o-2024-08-06	gpt4o test	8.0%	16.0%	0.0%	8.0%
	gpt4o test2	6.0%	16.0%	2.0%	8.0%
	gpt4o test3	10.0%	20.0%	8.0%	12.7%
	gpt4o test4	8.0%	18.0%	4.0%	10.0%
	gpt4o test5	12.0%	18.0%	8.0%	12.7%
gpt-5-2025-08-07	gpt-5-2025-08-07 test	10.0%	0.0%	8.0%	6.0%
	gpt-5-2025-08-07 test2	6.0%	0.0%	0.0%	2.0%
	gpt-5-2025-08-07 test3	8.0%	0.0%	4.0%	4.0%
	gpt-5-2025-08-07 test4	12.0%	2.0%	8.0%	7.3%
	gpt-5-2025-08-07 test5	8.0%	0.0%	6.0%	4.7%

Because the study evaluated recorded API outputs associated with those endpoint identifiers, this was interpreted as a benchmark-specific comparison of behavior

during the collection period. The remaining detector-positive cases were concentrated in numeric/database-style questions and adversarial auto-generated prompts.

4.1.1 *Inter-run variability*

The auto-generated prompt set generated the most variability. This is consistent with two plausible mechanisms: prompts closer to the decision boundary and sensitivity to early token choices after first-line cleaning. Notably, the factual and riddle categories remained comparatively stable, while most of the variability was concentrated in the auto-generated category.

4.1.2 *Coverage and abstention*

Empty first-line answers were counted by the generation script as non-contradictory abstentions. Across 750 legacy answers per endpoint, non-empty first-line coverage was 100.0% for gpt-3.5-turbo-0125, gpt-4-0613, and gpt-4o-2024-08-06, but 80.5% for gpt-5-2025-08-07 because 146 outputs were empty after cleaning. Detector-positive rates conditional on a non-empty first line were 23.5%, 14.3%, 10.3%, and 6.0%, respectively. Thus the gpt-5 unconditional rate of 4.8% partly reflects answer withholding, not just fewer contradictions.

4.1.3 *Detector component attribution*

Detector-positive rows were assigned to the first triggering rule in the implemented decision order. This is decision-path attribution; it is not a counterfactual ablation, and the study does not claim that disabling a component would leave all other decisions unchanged. Embedding similarity was not listed as a hallucination trigger because it supported agreement decisions but did not independently force a detector-positive label.

4.2 Detector performance vs. audit labels

The audit contained 1,850 rows: 450 legacy short-answer rows, 400 non-atomic pilot rows, and 1,000 balanced non-atomic add-on rows. The balanced audit contained 200 audited rows per category, but its sampling deliberately included every detector-positive output and a random subset of detector-negative outputs. It therefore supports direct precision estimates for the complete detector-positive frame and weighted estimates of false-positive behavior; unweighted row prevalence was not interpreted as population prevalence.

The original 850-row audit file contained two review-pass labels and adjudicated final labels. The balanced 1,000-row add-on was independently labeled by two human annotators using the same label definitions. Detector outputs and endpoint identities were withheld during annotation, and disagreements were resolved by adjudication. The released balanced-audit table contains both pre-adjudication labels, annotator rationales, agreement status, and the final adjudicated label. The false-positive taxonomy was applied after adjudication as a descriptive mechanism analysis of detector-positive rows that received non-contradiction labels.

The audit label set separates explicit contradiction from other forms of poor answer quality. In particular, incomplete_or_vacuous responses were not treated as successful answers, even when they did not explicitly

contradict the reference. Likewise, unsupported_extra_claim was reserved for responses that provided a correct core answer but added an inaccurate or unsupported additional claim. This label structure directly addressed the limitation that a contradiction-only detector can mark vacuous or incomplete outputs as non-hallucinations. Final labels followed the explicit-contradiction label definitions: reasonable paraphrases, correct core answers, and underspecified but non-mutually-exclusive answers were treated as non-contradictions, while direct conflicts with the reference remained labeled as contradictions.

Across the 1,850 audit rows, 1,602 responses are labeled supported, 94 contradiction, 114 incomplete_or_vacuous, 36 abstention, 3 unsupported_extra_claim, and 1 needs_review. For the original 850-row audit, the two review passes agree on 811 rows (95.4%), with Cohen's $\kappa = 0.845$. For the balanced 1,000-row add-on, the two human annotators agreed on 968 rows (96.8%), with $\kappa = 0.866$. All 32 balanced-add-on disagreements concern the boundary between supported and incomplete_or_vacuous; the annotators agreed on the binary contradiction label for all 1,000 rows ($\kappa = 1.000$). Across the full 1,850-row audit, exact multiclass agreement is 96.2% with $\kappa = 0.856$, and binary contradiction agreement is 100.0% with $\kappa = 1.000$. The adjudicated labels were used as the final audit labels in Table 4.

Table 4 reports detector performance with explicit contradiction as the positive class. The principal result is precision falling from 84.6% on the legacy short-answer audit to 8.1% in the balanced non-atomic add-on, indicating benchmark-transfer failure. The legacy and pilot rows use their original row-level estimates. In the balanced add-on, all detector-positive rows received review labels, so precision is direct; recall and F1 are inverse-probability weighted because detector-negative rows were sampled. The balanced audit has 24 true positives and 274 false positives among 298 detector-positive outputs. Observed contradiction recovery is 24/25 (Wilson interval 80.5%–99.3%) before weighting. The 88.0% weighted recall estimate is driven by one sampled false negative and is not used for category comparison.

Review of the binary error rows helps explain this transfer failure. In the balanced non-atomic audit, 274 of 298 detector-positive rows are false positives, while only 1 contradiction-labeled response is missed. The false positives are concentrated in citation-grounded answering and nuanced uncertainty prompts, which together account for 208 of the 274 balanced-audit false positives. These prompts are diagnostically useful because reliable answers may mention unavailable information, unverified claims, claims to avoid, or limits on what evidence supports. The detector sometimes treats those cautionary or rejecting mentions as contradiction evidence even when the final audit label is non-contradictory, indicating difficulty distinguishing proposition content from its assertion or epistemic status. A direct detector-level follow-up is to add an uncertainty-aware guardrail that distinguishes explicit factual contradiction from cautious

statements about unavailable, unsupported, or insufficient evidence.

Table 3. Binary interpretation of final audit labels for the explicit-contradiction metric. The negative labels are still reported because they identify semantic or completeness problems outside the detector's binary target.

Final audit label	Binary metric role	Interpretation
contradiction	Positive	Direct conflict with the reference or source constraint.
supported	Negative	Correct, paraphrased, or underspecified but not mutually inconsistent.
incomplete_or_vacuous	Negative, separately reported	Empty, vague, or insufficient response without explicit contradiction.
abstention	Negative, separately reported	Refusal or non-answer with no conflicting factual claim.
unsupported_extra_claim	Negative, separately reported	Correct core answer plus unsupported extra content not treated as explicit contradiction.
needs_review	Negative, separately reported	Ambiguous row retained for transparency.

Table 4. Detector performance by audit subset. For the balanced non-atomic audit, all detector-positive rows received review labels; recall and F1 are inverse-probability weighted for sampled detector-negative rows. The displayed counts are audited-row counts.

Audit subset	*n*	TP	FP	FN	TN	Precision	Recall	F1
Legacy short-answer audit	450	55	10	3	382	84.6%	94.8%	89.4%
Non-atomic pilot audit	400	10	64	1	325	13.5%	90.9%	23.5%
Balanced non-atomic audit	1000	24	274	1	701	8.1%	88.0%[a]	14.8%[a]

[a]Inverse-probability-weighted estimate. The unweighted audit-sample recall is 96.0%; it is not treated as a population estimate because detector-negative rows were subsampled.

To account for repeated prompts and prompt-response pairs, Table 5 reports the audit structure by subset. The full audit contains 617 prompt identities and 1,671 unique prompt-response pairs. These counts describe duplication and clustering; combined raw performance metrics are not used because the three audit subsets have different sampling designs.

Table 5. Audit structure and duplicate-aware counts.

Audit subset	Rows	Prompt identities	Unique prompt-responses	Detector positives
Legacy short-answer audit	450	150	361	65
Non-atomic pilot audit	400	92	400	74
Balanced non-atomic audit	1000	375	910	298
Total	1850	617	1671	437

4.3 Diagnostic category analysis in the balanced non-atomic setting

The balanced non-atomic audit evaluates five prompt categories with 200 audited rows each. Its category analysis localizes the broader precision collapse while resolving uneven sampling for category-specific false-positive behavior. It does not provide a powered comparison of category-specific recall because naturally occurring contradiction positives remain sparse. The audited contradiction-positive counts are 15 for long-form explanation, 9 for synthesis, 1 for nuanced uncertainty, and 0 for causal reasoning and citation-grounded answering. Ordinary Wilson intervals for recall are correspondingly broad: 70.2%–98.8% for long-form explanation, 70.1%–100% for synthesis, and 20.7%–100% for nuanced uncertainty. Recall and F1 are therefore not the basis of the category analysis.

Table 6. Category-level results in the balanced non-atomic audit. Panel A reports audited-row counts and precision. Panel B reports inverse-probability-weighted false-positive rate (FPR), prompt-clustered 95% bootstrap confidence intervals, and specificity for the 560-response category frames. Recall/F1 are omitted from category comparison because contradiction positives are sparse.

Panel A. audited counts and precision.

Category	*n*	Contr.	TP	FP	FN	TN	Precision
Long-form explanation	200	15	14	28	1	157	33.3%
Causal reasoning	200	0	0	5	0	195	0.0%
Synthesis	200	9	9	33	0	158	21.4%
Citation-grounded answering	200	0	0	104	0	96	0.0%
Nuanced uncertainty	200	1	1	104	0	95	1.0%
Overall balanced audit	1000	25	24	274	1	701	8.1%

Panel B. weighted false-positive behavior.

Category	Weighted FPR	95% prompt-cluster CI	Specificity
Long-form explanation	5.2%	2.5%–8.4%	94.8%
Causal reasoning	0.9%	0.2%–1.8%	99.1%
Synthesis	6.0%	3.2%–9.5%	94.0%
Citation-grounded answering	18.6%	11.5%–26.6%	81.4%
Nuanced uncertainty	18.6%	13.8%–24.3%	81.4%
Overall balanced audit	9.9%	7.9%–11.9%	90.1%

Table 6 reports precision and weighted false-positive rate (FPR) for rows assigned non-contradiction audit labels. Citation-grounded answering and nuanced uncertainty have weighted false-positive rates of 18.6% each, compared with 0.9% for causal reasoning,

5.2% for long-form explanation, and 6.0% for synthesis. The prompt-clustered confidence intervals preserve the same separation: citation-grounded answering 11.5%–26.6% and nuanced uncertainty 13.8%–24.3%, versus upper bounds below 9.6% for the other categories. This pattern localizes transfer failure.

The 274 balanced-audit false positives were coded into mutually exclusive explanation classes by reviewing the prompt, reference answer, model response, detector label, and final audit label. Table 7 reports the taxonomy by category. The largest class, with 167 false positives, occurs when the response names a forbidden, unsupported, or mistaken claim as something to avoid rather than endorsing it. Another 70 false positives involve source, evidence, current-data, or availability limitations. The remaining 37 are supportable compressed explanations, syntheses, or causal-caution paraphrases. Across these classes, confusion between proposition content and its assertion or epistemic status helps explain the benchmark-transfer failure.

Table 7. False-positive taxonomy for the 274 balanced non-atomic false positives. Categories are mutually exclusive.

False-positive class	Causal	Citation	Long-form	Uncertainty	Synthesis	Total
Source/evidence/current-limitation wording	0	8	0	62	0	70
Forbidden/unsupported claim or mistake named but not endorsed	0	96	0	42	29	167
Supportable/compressed explanation, synthesis, or causal paraphrase	5	0	28	0	4	37
Total	5	104	28	104	33	274

Table 8. Component attribution for the 274 balanced non-atomic false positives.

Detector component or trigger class	Causal	Citation	Long-form	Uncertainty	Synthesis	Total
NLI / semantic contradiction trigger	5	96	28	85	33	247
Unavailability / nonexistence heuristic	0	8	0	19	0	27
Numeric conflict or numeric alternatives	0	0	0	0	0	0
Atomic entity mismatch	0	0	0	0	0	0
Explicit lexical contradiction cue	0	0	0	0	0	0

For component attribution, the 274 balanced-audit false positives were assigned to the first detector trigger in the implemented decision order. This traces the recorded decision path but does not estimate the counterfactual effect of disabling that component. Table 8 shows that 247 false positives are attributed to the NLI/semantic contradiction trigger and 27 to the unavailability/nonexistence trigger. Numeric conflict, atomic mismatch, and explicit lexical contradiction cues do not account for false positives in the balanced non-atomic audit. Embedding similarity is not listed as a hallucination trigger because in this

implementation it supports agreement decisions and does not independently force hallucination labels.

These analyses replace the earlier uneven category comparison with a balanced analysis of false-positive behavior. The primary inference is benchmark-transfer failure: strong measured precision on the compact legacy audit collapses in the balanced non-atomic setting. The category results localize this failure, and the taxonomy characterizes a likely linguistic mechanism. Reliable answers may mention unsupported claims, reject conclusions, or discuss missing evidence, and the detector can confuse that discourse with an asserted contradiction. Category-specific recall remains too sparse for comparative claims.

4.4 Error analysis (audit disagreements)

The audit changes the interpretation of detector errors. In the legacy short-answer subset, remaining errors continue to involve synonym, notation, and granularity issues, such as equivalent scientific notation or overly broad chemical names. In the balanced non-atomic audit, the taxonomy above shows that many false positives arise from cautionary, citation-limited, source-limitation, or avoid-this-claim language instead of direct factual conflict. These cases show over-sensitivity in the binary detector and do not invalidate the legacy short-answer result.

The audit therefore distinguishes two failure modes that were previously conflated: (i) explicit contradiction errors, where the model response conflicts with the reference; and (ii) semantic adequacy errors, where the response is empty, vague, incomplete, or insufficiently grounded but not explicitly contradictory. Semantic-adequacy failures remain important, but they are reported through separate labels and are not treated as positives in the binary explicit-contradiction metric. This distinction is important because HalluDetector was designed as a conservative contradiction detector and is not intended as a general answer-quality evaluator.

False positives occur when the detector flags contradiction but the final audit label is non-contradictory. Across the full 1,850-row audit, there are 348 row-level false positives: 338 assigned supported, 8 assigned abstention, and 2 assigned incomplete_or_vacuous. False negatives occur when the final audit label identifies explicit contradiction but the detector does not; the full audit contains 5 row-level false negatives. The legacy short-answer subset accounts for 10 false positives and 3 false negatives, the non-atomic pilot subset accounts for 64 false positives and 1 false negative, and the balanced non-atomic subset accounts for 274 false positives and 1 false negative. This pattern shows that the detector remains recall-oriented for explicit contradictions, while the enhanced categories reveal systematic over-sensitivity to supportable answers, cautious non-answers, and metalinguistic references to unsupported claims. Non-contradictory adequacy failures are therefore kept separate and are not folded into the binary explicit-contradiction metric.

4.5 Qualitative error analysis

The qualitative review examined hallucination cases to identify common patterns across both

compact reference-based prompts and the balanced non-atomic audit.

4.5.1 *Confident misremembering*

Earlier endpoints, especially gpt-3.5-turbo-0125, often produced plausible and confident answers that were factually wrong. For example, when asked about a specialized term, the model sometimes gave a real term from a related domain instead of the reference answer. These cases are important because they resemble ordinary hallucinations: the output is fluent and specific, but the content conflicts with the reference.

4.5.2 *Negation and unavailability errors*

Some hallucinations involved incorrectly stating that something did not exist, was unavailable, or could not be identified when the reference contained a concrete answer. These cases are well matched to the detector's contradiction-cue and unavailability heuristics because the model response directly negates the premise or availability of the reference answer.

4.5.3 *Over-answering*

Some responses began with a correct or nearly correct core answer but then added extra details that introduced inaccuracies. Under a strict contradiction policy, a response can be flagged if an added claim conflicts with the reference or source constraint, even when the first part of the answer is useful. This remains a difficult edge case because the core answer and the added explanation may deserve separate judgments.

4.5.4 *Riddle guessing*

Many incorrect riddle responses were direct guesses with little semantic relation to the expected solution. These cases often behave like atomic-answer contradictions because the response presents a mutually exclusive answer to a prompt with a short expected solution.

4.5.5 *Computation and numeric mistakes*

A smaller set of errors involved numeric or simple reasoning mistakes. The detector is relatively sensitive to these because numeric conflict is an explicit signal. However, the audit also shows why numeric normalization matters: equivalent notation, missing units, or underspecified direction can otherwise create false positives.

4.5.6 *Abstention and calibration*

The gpt-5-2025-08-07 endpoint produced many more empty or abstaining first-line answers than the other evaluated endpoints. This behavior lowers detector-labeled contradiction-style hallucination rates because the detector treats abstention as non-contradiction. This behavior lowers the detector-positive rate and therefore partly reflects answer withholding; whether the increased withholding represents improved calibration was not directly evaluated.

4.5.7 *Semantic inadequacy without explicit contradiction*

The enhanced non-atomic audit makes this failure mode more visible. Some answers are vague, incomplete, unsupported, or insufficiently responsive without directly contradicting the reference. These cases are not counted as positives in the binary explicit-contradiction metrics, but they are reported

separately through labels such as incomplete_or_vacuous and unsupported_extra_claim.

Overall, these patterns show that hallucination is not monolithic. The benchmark targets the explicit-contradiction portion of hallucination behavior, while the expanded audit separately records semantic-adequacy failures that fall outside that binary target. This distinction is necessary for interpreting the detector's precision, recall, and F1 scores.

5. Discussion

The principal scientific finding is benchmark-transfer failure: measured precision falls from 84.6% on the legacy short-answer audit to 8.1% in the balanced non-atomic setting. The category analysis helps explain this collapse by exposing cases in which proposition content is mentioned, rejected, or qualified under source and evidence limitations. The taxonomy indicates that the detector can respond to that content without adequately representing assertion or epistemic status.

5.1 Conservatism: contradiction detection vs. correctness

Because HalluDetector labels hallucination only on contradiction evidence, it targets a narrower construct than overall answer incorrectness. A model can be wrong yet not contradict the reference strongly enough to trigger a veto, especially when the response is vague or content-free. Thus, the measured quantity is closer to "rate of contradicted factual claims" than "task error rate." This is a reasonable definition in some safety settings, but it must be stated explicitly when comparing with accuracy-focused benchmarks.

5.2 Re-interpreting 0% on riddles for gpt-5-2025-08-07

A 0–2% detector-labeled contradiction rate for riddles under this policy does not guarantee perfect riddle-solving; it indicates that the detector did not identify contradictions. For riddles, stricter scoring would yield a more task-accuracy-like picture.

5.3 Marginal utility of external contradiction detectors

The high abstention rate observed for gpt-5-2025-08-07 suggests that this behavior lowers the detector-positive rate and therefore partly reflects answer withholding; whether the increased withholding represents improved calibration was not directly evaluated. As frontier models become more willing to abstain or express uncertainty, the marginal value of an external contradiction detector may decrease for simple atomic prompts. However, such detectors can still be useful as an auditable safety layer for regression testing, cross-model comparisons, retrieval-grounded verification, and cases where a model gives a confident answer instead of abstaining. The results therefore suggest that the detector's role may shift over time; they do not imply that external verification becomes unnecessary.

5.4 Practical improvements

The enhanced-audit failures are not evidence that contradiction detection is unusable. They identify specific detector rules that can be refined. The most direct improvement is an uncertainty-aware guardrail that distinguishes

asserting a false claim from mentioning uncertainty, missing evidence, unavailable sources, or claims that should be avoided. A second improvement is an assertion-scope check that separates what the model endorses from what it merely names as unsupported or prohibited. For the legacy short-answer setting, smaller refinements such as scientific-notation normalization, alias dictionaries for common-name or spelling variants, and vacuous-answer detection for atomic definitional questions remain useful. These changes can be implemented without changing the overall architecture and tested against the existing 1,850-row audit set.

5.5 Layered evaluator uncertainty

HalluDetector does not eliminate uncertainty by moving evaluation from a generative model to a detector. The detector also relies on imperfect components, including RoBERTa-MNLI, embedding similarity, lexical heuristics, numeric thresholds, and adjudication. This creates a layered evaluation setting in which the base model may contradict the reference, the detector may misclassify the response, and review passes may disagree about borderline cases. The justification for this architecture is therefore not that the detector is infallible, but that its rules, thresholds, intermediate signals, and audit disagreements are explicit and reproducible. In this study, the external detector improves auditability by converting model behavior into explicit contradiction-evidence decisions and by exposing where those decisions fail; it does not remove evaluator uncertainty entirely.

5.6 Assumptions

The analysis rests on several explicit assumptions. First, the reference answers are treated as sufficiently correct for reference-based contradiction scoring; this is reasonable for a controlled prompt-reference benchmark, but any reference error would propagate into detector and audit labels. Second, archived endpoint identifiers and recorded outputs are treated as the evaluated behavior during the collection window. Third, the contradiction-evidence policy is assumed to be suitable for measuring explicit contradiction risk, not general answer quality. Fourth, final audit labels are used for analysis; the original 850 rows and balanced 1,000-row add-on both retain two pre-adjudication labels. The balanced add-on obtained 96.8% multiclass agreement ($\kappa = 0.866$) and 100.0% binary contradiction agreement ($\kappa = 1.000$) before adjudication. Fifth, repeated prompts and repeated prompt-response pairs are not assumed to be independent generalization evidence, which is why row-level metrics are reported alongside deduplicated prompt-response and prompt-identity statistics.

5.7 Scope and limitations of inference

These results should be interpreted in light of several scope conditions. First, category balance does not create contradiction-positive events: two categories contain no contradiction-positive labels and one contains only one, so the study does not claim a powered category-level recall comparison. The category result is based on weighted false-positive rate and specificity. Second, the balanced audit includes every detector-positive response but samples detector-negative responses; raw unweighted prevalence and

recall are not population estimates, and the weighted estimates assume the sampled detector-negative rows are representative within category. Third, the balanced 1,000-row add-on was independently labeled by two human annotators blinded to detector output and endpoint identity, with disagreements resolved by adjudication. Multiclass agreement was 96.8% ($\kappa = 0.866$), and binary contradiction agreement was 100.0% ($\kappa = 1.000$). Although these values support label reliability, they do not eliminate judgment at the supported versus incomplete_or_vacuous boundary, which accounts for all balanced-add-on disagreements. Fourth, the updated corpus remains compact but deliberately varied; it is not a comprehensive estimate of general-purpose hallucination behavior. Fifth, the archived outputs preserve API model identifiers, decoding defaults, prompts, model responses, detector labels, and run timestamps where available; accordingly, the cross-model comparison should be read as a comparison of recorded API behavior during the collection window. Finally, because the detector intentionally treats incompleteness and abstention as non-hallucination unless contradiction evidence is present, the reported rates are best read as conservative contradiction-focused hallucination rates rather than a full task-error metric.

5.8 Beyond reference-based detection

Many deployments lack a gold reference. One remedial path is retrieval-augmented verification: retrieve evidence from trusted corpora and treat that evidence as a proxy reference, then apply a similar contradiction detector (connected to FEVER-like pipelines) (6). Another path is model self-consistency (e.g., SelfCheckGPT) (9), trading compute for reference-free detection. Related corpus-building efforts for trustworthy hallucination evaluation include RAGTruth (12).

6. Conclusion

This study shows that strong contradiction-detection performance on compact reference-based questions does not necessarily transfer to more complex, non-atomic language. HalluDetector achieved 84.6% precision on the legacy short-answer audit but only 8.1% in the balanced non-atomic setting, demonstrating substantial benchmark-transfer failure. The error analysis identifies an important reason for this collapse: a detector may recognize the semantic content of a false or unsupported proposition without correctly determining whether the model is asserting, rejecting, qualifying, or merely discussing that proposition. This problem was especially evident in citation-grounded answers and expressions of nuanced uncertainty, which produced substantially higher false-positive rates than the other evaluated categories.

These findings clarify both the utility and the limits of reference-based contradiction detection. HalluDetector remains useful as a recall-oriented screen for explicit disagreement with compact reference answers, but its output should not be interpreted as a general measure of hallucination, factual correctness, or semantic adequacy. The results also show why detector performance established on short-answer benchmarks should not be generalized to complete or linguistically complex responses without explicit transfer testing. More broadly,

reliable hallucination evaluation requires not only identifying proposition content but also determining its assertion scope and epistemic status. Future detector development should therefore incorporate assertion-scope checks and uncertainty-aware guardrails that distinguish an endorsed false claim from a false claim that an answer mentions in order to reject or qualify it.

Acknowledgments

The author acknowledges use of language-editing tools for formatting, grammar, and clarity checks. The author reviewed the manuscript, audit interpretations, tables, and conclusions and is responsible for the final content.

Data and code availability

The code, prompts, benchmark files, and evaluation outputs used in this study are available at https://github.com/seun829/HalluDetector and https://doi.org/10.6084/m9.figshare.32095732 .

7. References

1. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

2. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv*. https://doi.org/10.48550/arXiv.2203.02155

3. OpenAI. (2023). GPT-4 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

4. Li, J., Chen, J., Ren, R., Cheng, X., Zhao, X., Nie, J.-Y., & Wen, J.-R. (2024). The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10879–10899). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.586

5. Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.750

6. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074

7. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 1–46. https://doi.org/10.1162/coli.a.16

8. Wang, Y., Wang, M., Manzoor, M. A., Liu, F., Georgiev, G. N., Das, R. J., & Nakov, P. (2024). Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 19519–19529). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.1088

9. Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.557

10. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://doi.org/10.48550/arXiv.1907.11692

12. Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *ArXiv*. https://doi.org/10.48550/arXiv.2401.00396

Appendix A. Audit set construction and sampling

The audit table contains 1,850 rows: 450 legacy short-answer rows, 400 non-atomic pilot rows, and 1,000 balanced non-atomic rows. The balanced audit is sampled from the 2,800-response add-on frame. It includes all 298 detector-positive rows and a seed-42 simple random sample

without replacement of detector-negative rows within each category to reach 200 audited rows per category. Table 9 reports the frame and sample counts, inclusion probabilities, and unique prompt representation.

Table 9. Balanced non-atomic audit sampling design. p_- is the detector-negative inclusion probability.

Category	Frame +	Frame −	Audited +	Audited −	p_-	Frame	Audited
Causal reasoning	5	555	5	195	0.3514	80	78
Citation-grounded answering	104	456	104	96	0.2105	80	73
Long-form explanation	42	518	42	158	0.3050	80	74
Nuanced uncertainty	105	455	105	95	0.2088	80	72
Synthesis	42	518	42	158	0.3050	80	78

The balanced audit includes responses from seven archived runs. Table 10 reports detector-positive/detector-negative audited counts by endpoint and category. A prompt may contribute multiple endpoint or run responses.

Table 10. Balanced audit composition by endpoint and category. Cells show detector-positive/detector-negative audited rows.

Endpoint	Causal	Citation	Long-form	Uncertainty	Synthesis
gpt-3.5-turbo-0125	0/60	24/35	16/50	20/36	10/51
gpt-4-0613	3/55	36/19	12/40	40/19	13/40
gpt-4o-2024-08-06	1/45	27/24	11/45	33/23	10/44
gpt-5-2025-08-07	1/35	17/18	3/23	12/17	9/23

Each audit row contains the prompt, reference answer, model response, detector label, both pre-adjudication labels and rationales, agreement status, final adjudicated label, duplicate identifiers, and prompt-cluster identifiers. The original 850-row audit obtains 95.4% multiclass agreement with Cohen's $\kappa = 0.845$. The balanced 1,000-row add-on was independently labeled by two human annotators blinded to detector output and endpoint identity, and disagreements were resolved by adjudication. It obtains 96.8% multiclass agreement with $\kappa = 0.866$ and 100.0% binary contradiction agreement with $\kappa = 1.000$. Across all 1,850 rows, multiclass agreement is 96.2% with $\kappa = 0.856$. The final labels are 1,602 supported, 94 contradiction, 114 incomplete_or_vacuous, 36 abstention, 3 unsupported_extra_claim, and 1 needs_review.

Table 11. Agreement between the two pre-adjudication labels. Binary contradiction agreement collapses all non-contradiction labels into one negative class.

Audit group	n	Multiclass agreement	Multiclass κ	Binary agreement	Binary κ
Original 850-row audit	850	95.4%	0.845	100.0%	1.000
Balanced 1,000-row add-on	1,000	96.8%	0.866	100.0%	1.000
All audit rows	1,850	96.2%	0.856	100.0%	1.000

Within the balanced add-on, category-level multiclass agreement is 100.0% for causal reasoning, 94.5% for citation-grounded answering, 94.5% for long-form explanation, 96.5% for nuanced uncertainty, and 98.5% for synthesis. The corresponding multiclass κ values are 1.000, 0.761, 0.805, 0.870, and 0.928. Binary contradiction agreement is 100.0% in every category. Category-level binary κ is undefined for causal reasoning and citation-grounded answering because both annotators assign zero contradiction labels in those categories.

The full audit contains 617 prompt identities and 1,671 unique prompt-response pairs. The false-positive taxonomy is reported as an adjudicated descriptive classification of the 274 balanced-add-on false positives into mutually exclusive mechanism categories.

Appendix B. Detector thresholds and defaults

Table 12 lists key detector threshold values.

Table 12. Key detector thresholds.

Parameter	Value	Meaning
numeric rel tol	0.005	Relative tolerance for numeric match (non-integer).
numeric abs tol	0.05	Absolute tolerance for numeric match (non-integer).
embed model name	sentence-transformers/ all-MiniLM-L6-v2	Sentence embedding model for cosine similarity.
embed min similarity	0.72	Cosine threshold for embedding baseline match.
embed min shared content tokens	1	Embedding gate: minimum shared content tokens.
jaccard min	0.4	Lexical Jaccard threshold when embeddings not used.
nli model name	roberta-large-mnli	NLI model for bidirectional contradiction.
nli min tokens	3	Minimum token length for NLI checks.
nli contradiction veto	0.75	Contradiction probability threshold for veto.
nli entailment veto floor	0.3	Entailment floor for contradiction veto.
atomic gold max tokens	2	Atomic gold heuristic: max tokens.
atomic resp max tokens	5	Atomic mismatch heuristic: max response tokens.

Appendix C. Single-run descriptive visuals

The repository includes graphs per run: hallucinations by template.png, hallucinations by keywords.png, and feature correlation heatmap.png. Figures 2–4 show the visual summaries for gpt-3.5-turbo-0125 run gpt3.5-turbo test.

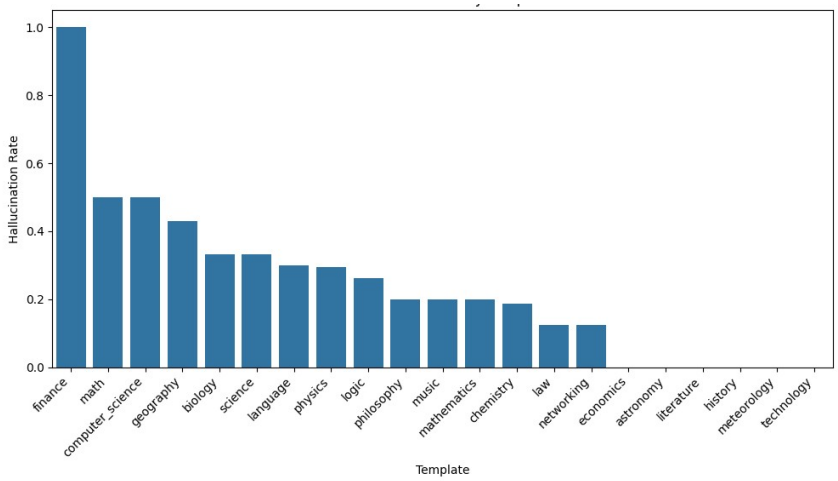

Figure 2. Detector-labeled contradiction-style hallucination rate by template for gpt-3.5-turbo-0125 (run gpt3.5-turbo test).

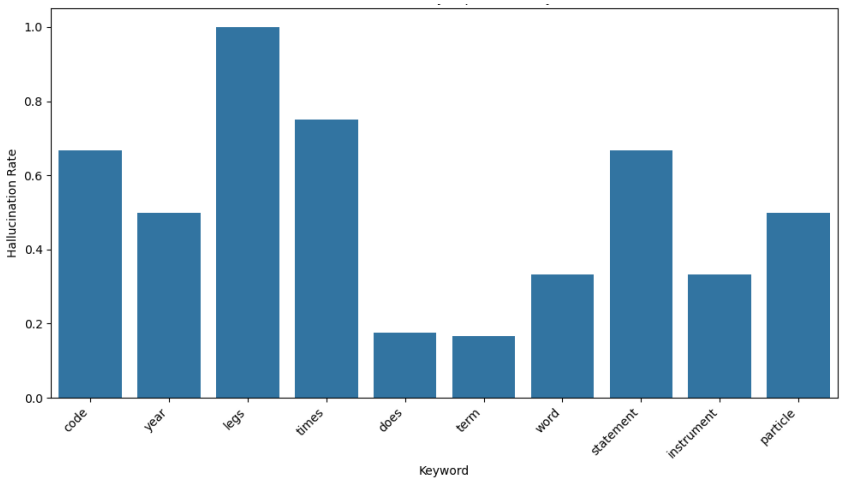

Figure 3. Detector-labeled contradiction-style hallucination rate by keyword for gpt-3.5-turbo-0125 (run gpt3.5-turbo test).

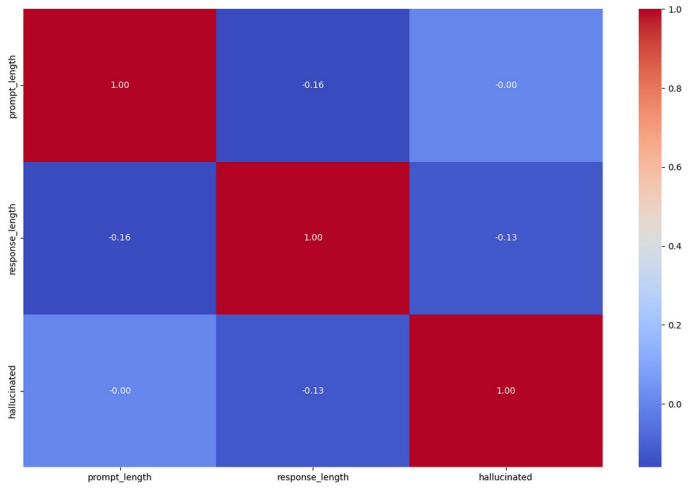

Figure 4. Feature correlation heatmap for gpt-3.5-turbo-0125 (run gpt3.5-turbo test). Note: the repo defines "is correct" as "not hallucinated" for plotting convenience.

Table 13 summarizes template-level extremes for a representative early endpoint and the final gpt-5 run. Counts are shown with denominators.

Table 13. Top detector-labeled contradiction-style hallucination-rate templates in gpt-3.5-turbo-0125 (run gpt3.5-turbo test) vs gpt-5-2025-08-07 (run gpt-5-2025-08-07 test5). Rates are detector-labeled.

gpt-3.5-turbo-0125		**gpt-5-2025-08-07**	
Template	**Rate (count)**	**Template**	**Rate (count)**
finance	100.0% (1/1)	chemistry	25.0% (4/16)
math	50.0% (4/8)	geography	14.3% (1/7)
computer science	50.0% (1/2)	physics	5.9% (1/17)
geography	42.9% (3/7)	logic	5.3% (1/19)
biology	33.3% (2/6)	astronomy	0.0% (0/13)
science	33.3% (2/6)	economics	0.0% (0/1)

C.1 Length correlations

Across all runs, prompt length and response length are weak predictors of detector-labeled contradiction. The maximum absolute Pearson correlation observed between the detector label and prompt length across runs is below 0.22, and between the detector label and response length

below 0.20. This indicates that prompt and response length alone are weak predictors of the detector label and motivates evaluation of richer semantic features.

C.2 Repo artifacts and reproduction notes
The paper references prompt CSVs, labeled outputs, run-level rates, balanced non-atomic benchmark outputs, audit files, and plots in the repository output directories. The released materials include both pre-adjudication label columns, agreement and confusion-matrix files, the audit sampling design, analysis weights, category estimates, confidence intervals, and model-by-run composition used in the revised tables. If outputs are regenerated with different endpoints or decoding settings, the tables and figures should be updated accordingly.