

Peer-Review

Lee, Seungho. 2026. "A Conservative Benchmark of Contradiction Detection and Its Limits in LLM Hallucination Evaluation." *Journal of High School Science* 10 (3): 582–607.

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

1. The detector marks incomplete responses as neutral (non-hallucinations). This means that its ontology treats meaningless placeholders, vacuous outputs, and semantic failure as acceptable unless contradiction is explicit. Hence, the detector is fundamentally measuring contradiction rather than semantic adequacy or hallucination more broadly. While the manuscript explicitly acknowledges this limitation, the repeated use of the term 'contradiction-style hallucination' still risks readers interpreting the detector as a more general hallucination detector than is actually operationalized. I therefore recommend simplifying the central framing toward 'contradiction detection' or otherwise further emphasizing throughout the manuscript that the benchmark measures explicit contradiction only.

2. The benchmark appears skewed toward atomic-answer QA. Many prompts are short factual retrievals, riddles or single-token answers. Examples: "How many white keys..." "What is the Oort cloud?" "Which 9 letter word...". This creates an important concern: The detector may work well precisely because the benchmark is unusually favorable to contradiction-style evaluation.

Real hallucination problems often involve:

long-form explanation,
causal reasoning,
synthesis,
citation generation,
and nuanced uncertainty.

The audit set strengthens the impression that this is fundamentally an atomic QA contradiction benchmark and not a broad hallucination benchmark. I will need to see a substantial # of prompts which are not favorable to contradiction-style hallucination added to your database. I anticipate that at least 50% of prompts fall in this category.

3. Human labeling criteria appear underdefined: The audit notes repeatedly say, "Agreed with detector.", but the CSV does not appear to include:

annotation guidelines,
inter-rater agreement,
annotator count,
or adjudication procedures.

That is a significant omission because hallucination is notoriously subjective. Please discuss this in the manuscript. Were there multiple annotators? Was Cohen's κ computed? Were disagreements adjudicated? Were annotators blinded to detector output?

The current audit structure strongly suggests a single-pass validation likely performed by the authors themselves. This substantially weakens claims of "human validated" detector performance.

4. The CSV file confirms the existence of identical prompts, identical responses, repeated multiple times. Due to this, the effective audit size sample is < 450 . The confidence is hence inflated since repeated identical samples can artificially stabilize precision/recall. The detector can memorize behavior. If the same failure pattern appears repeatedly, the metrics may overstate generalization. Please therefore report, statistics on unique-prompt auditing, per-prompt aggregation and clustered statistics.

5. The near-perfect riddle and auto-generated metrics compared with substantially lower factual-domain performance suggest that the benchmark categories may not stress the detector equally. This raises the possibility that the detector performs best on short atomic-answer contradiction tasks rather than broader factual semantic evaluation (see point 2).

6. If the detector itself uses probabilistic NLP models, then you have not eliminated contradiction risk — you have displaced it into another layer. The detector relies on RoBERTa-MNLI, embeddings, heuristics and thresholds. So now; GPT may contradict, the detector may misclassify,

and humans may disagree. This creates recursive evaluator uncertainty.

In particular, because the detector itself relies on probabilistic NLP components (e.g., NLI models and embedding similarity), the manuscript should discuss the epistemic limitations of layered AI-based verification and clarify why this architecture materially improves reliability rather than simply relocating uncertainty into a secondary evaluator. Discuss in the manuscript.

7. An important implication of the presented results is that the practical utility of the external contradiction detector may decrease as newer frontier models increasingly internalize contradiction control through abstention and uncertainty calibration. Indeed, the manuscript itself shows that GPT-5 exhibits substantially higher abstention rates alongside lower detector-labeled hallucination rates. This raises the possibility that the observed progression from GPT-3.5 to GPT-5 reflects improvements in the base models' intrinsic calibration mechanisms more than the effectiveness or necessity of the external detector architecture. The manuscript would benefit from discussing whether increasingly calibrated frontier models reduce the marginal utility of external contradiction-focused detectors over time.

Comment: The detector marks incomplete responses as neutral (non-hallucinations). This means that its ontology treats meaningless placeholders, vacuous outputs, and semantic failure as acceptable unless contradiction is explicit. Hence, the detector is fundamentally measuring contradiction rather than semantic adequacy or hallucination more broadly. While the manuscript explicitly acknowledges this limitation, the repeated use of the term 'contradiction-style hallucination' still risks readers interpreting the detector as a more general hallucination detector than is actually operationalized. I therefore recommend simplifying the central framing toward 'contradiction detection' or otherwise further emphasizing throughout the manuscript that the benchmark measures explicit contradiction only.

Response: The manuscript was revised to clarify that the detector measures explicit contradiction detection rather than general semantic adequacy or full hallucination detection. The Scope and Definition section now states that incompleteness is not treated as hallucination unless contradiction evidence is present. The revised audit separates contradiction from incomplete or vacuous, abstention, unsupported extra claim, and needs review. Precision, recall, and F1 are explicitly defined as binary metrics for the adjudicated explicit-contradiction class, and other audit labels are reported separately rather than being in a multi-class hallucination score. The title was revised to mention "explicit contradiction detection", and the abstract explicitly mentions that this is not a general hallucination or semantic adequacy detector.

Comment: The benchmark appears skewed toward atomic-answer QA. Many prompts are short factual retrievals, riddles or single-token answers. Examples: "How many white keys..." "What is the Oort cloud?" "Which 9 letter word...". This creates an important concern: The detector may work well precisely because the benchmark is unusually favorable to contradiction-style evaluation. Real hallucination problems often involve:

long-form explanation,
causal reasoning,
synthesis,
citation generation,
and nuanced uncertainty.

The audit set strengthens the impression that this is fundamentally an atomic QA contradiction benchmark and not a broad hallucination benchmark. I will need to see a substantial # of prompts

which are not favorable to contradiction-style hallucination added to your database. I anticipate that at least 50% of prompts fall in this category.

Human labeling criteria appear underdefined: The audit notes repeatedly say, “Agreed with detector.”, but the CSV does not appear to include:

annotation guidelines,

inter-rater agreement,

annotator count,

or adjudication procedures.

That is a significant omission because hallucination is notoriously subjective. Please discuss this in the manuscript. Were there multiple annotators? Was Cohen’s κ computed? Were disagreements adjudicated? Were annotators blinded to detector output?

The current audit structure strongly suggests a single-pass validation likely performed by the authors themselves. This substantially weakens claims of “human validated” detector performance.

Response: The benchmark and audit materials were revised to address both the prompt-distribution concern and the audit-protocol concern. First, the prompt database was expanded beyond the original 100 atomic/riddle prompts and 50 auto-generated prompts by adding a 100-prompt enhanced non-atomic benchmark. The added prompts are organized evenly across five template categories: long-form explanation, causal reasoning, synthesis, citation-grounded answering, and nuanced uncertainty. The updated corpus now contains 250 prompt-reference pairs and 5,000 archived model answers, including 2,000 answers from the enhanced non-atomic phase. The manuscript describes these enhanced prompt categories in the Experimental Setup section and treats the enhanced set as a stress test of the detector’s scope rather than as a straight replacement for the original/legacy atomic benchmark. The manuscript also now includes a compact representative-prompt table, with one example from each enhanced and original/legacy category, and the full prompt database remains available in the DOI-linked materials.

Second, the audit protocol was revised and described in the Detector performance vs. adjudicated audit and Audit Set Construction sections. The revised audit contains 850 adjudicated rows: 450 rows from the original/legacy atomic/riddle/auto-generated audit and 400 rows from the enhanced non-atomic audit sample. Each row includes the prompt, reference answer, model response, detector label, two independent annotator labels, annotator rationales, agreement status, adjudicated final label, duplicate identifiers, and prompt-cluster identifiers. Annotators were blinded to detector output and model identity, and disagreements were resolved by adjudication using the written label codebook. The manuscript now reports annotator agreement, Cohen’s kappa, the final adjudicated label distribution, duplicate-aware statistics, deduplicated prompt-response metrics, prompt-level aggregation, and prompt-cluster confidence intervals.

The revised materials substantially address the concern that the original/legacy benchmark favored short atomic contradiction tasks. The enhanced prompt set contributes 100 of 200 prompt-reference pairs excluding the 50 auto-generated adversarial prompts, and the enhanced audit sample contributes 400 of 850 audited rows. Because the enhanced prompts do not constitute a majority of the full prompt database, the manuscript does not claim to provide a comprehensive broad-hallucination benchmark. Instead, it reports the legacy and enhanced subsets separately and frames the enhanced subset as a substantial stress-test expansion showing how the detector behaves on less atomic, less contradiction-favorable prompts. The current split was used because the study primarily focuses on contradiction detection and retained the legacy 50 auto-generated adversarial prompts for continuity with the legacy benchmark and tables.

Comment: The CSV file confirms the existence of identical prompts, identical responses, repeated multiple times. Due to this, the effective audit size sample is < 450 . The confidence is hence

inflated since repeated identical samples can artificially stabilize precision/recall. The detector can memorize behavior. If the same failure pattern appears repeatedly, the metrics may overstate generalization.

Please therefore report statistics on unique-prompt auditing, per-prompt aggregation and clustered statistics.

Response: We revised the audit analysis to explicitly account for repeated prompts and repeated prompt-response pairs. The revised 850-row audit contains 242 underlying prompt identities represented in the sample audit and 761 unique prompt-response pairs. The manuscript now reports row-level metrics, deduplicated prompt-response metrics, and prompt-level aggregation. It also reports prompt-cluster bootstrap confidence intervals for the combined row-level explicit-contradiction metrics.

Comment: The near-perfect riddle and auto-generated metrics compared with substantially lower factual-domain performance suggest that the benchmark categories may not stress the detector equally. This raises the possibility that the detector performs best on short atomic-answer contradiction tasks rather than broader factual semantic evaluation (see point 2).

Response: We revised the audit analysis to distinguish the legacy atomic audit subset from the enhanced non-atomic stress-test subset. The manuscript now reports that the detector remains substantially stronger on the legacy atomic audit subset than on the enhanced non-atomic subset, where many detector-positive cases are adjudicated as supported or cautious non-contradictory answers rather than explicit contradictions. This reframes the enhanced audit as a scope stress test rather than a straight replacement for the legacy atomic audit.

Comment: If the detector itself uses probabilistic NLP models, then you have not eliminated contradiction risk — you have displaced it into another layer. The detector relies on RoBERTa-MNLI, embeddings, heuristics and thresholds. So now; GPT may contradict, the detector may misclassify, and humans may disagree. This creates recursive evaluator uncertainty. In particular, because the detector itself relies on probabilistic NLP components (e.g., NLI models and embedding similarity), the manuscript should discuss the epistemic limitations of layered AI-based verification and clarify why this architecture materially improves reliability rather than simply relocating uncertainty into a secondary evaluator. Discuss in the manuscript.

Response: The Discussion section was revised to add a concise paragraph titled “Layered evaluator uncertainty.” The manuscript now explicitly states that HalluDetector does not eliminate evaluator uncertainty and that the detector relies on imperfect components, including RoBERTa-MNLI, embedding similarity, lexical heuristics, numeric thresholds, and human adjudication. The revision clarifies that the architecture improves reliability not because the detector is infallible, but because its rules, thresholds, intermediate signals, and audit disagreements are explicit and reproducible. The manuscript now frames the detector as an auditable safety layer that exposes failure modes and converts model outputs into contradiction-evidence decisions, rather than as a mechanism that removes uncertainty entirely.

Comment: An important implication of the presented results is that the practical utility of the external contradiction detector may decrease as newer frontier models increasingly internalize contradiction control through abstention and uncertainty calibration. Indeed, the manuscript itself shows that GPT-5 exhibits substantially higher abstention rates alongside lower detector-labeled

hallucination rates. This raises the possibility that the observed progression from GPT-3.5 to GPT-5 reflects improvements in the base models' intrinsic calibration mechanisms more than the effectiveness or necessity of the external detector architecture. The manuscript would benefit from discussing whether increasingly calibrated frontier models reduce the marginal utility of external contradiction-focused detectors over time.

Response: The Discussion section was revised to add a paragraph titled “Marginal utility of external contradiction detectors.” The manuscript now explains that the high abstention rate observed for gpt-5-2025-08-07 suggests that part of the measured decline in detector-labeled hallucination reflects stronger intrinsic calibration, not only fewer underlying reasoning errors. The revision acknowledges that the marginal utility of an external contradiction detector may decrease for simple atomic prompts as frontier models become more willing to abstain or express uncertainty. It also explains why external detectors may still be useful for regression testing, cross-model comparison, retrieval-grounded verification, and cases where a model gives a confident answer rather than abstaining. The conclusion was also lightly revised to state that future work should evaluate when external contradiction detectors add value beyond a model's own abstention and uncertainty behavior.

Comment: Please verify that all links to the references point to the correct source.

Response: The reference links were checked against their DOI or official landing pages. The arXiv references resolve to the intended arXiv records, and the ACL/EMNLP/NAACL references resolve to the corresponding ACL Anthology records. During this check, one reference DOI was corrected: the 2025 Computational Linguistics survey by Zhang et al. now uses the DOI listed on its ACL Anthology record. No separate manuscript section was added for link verification, because this was a reference-list correction rather than a substantive methodological change.

Comment: The authors must disclose and acknowledge any assistance received in the preparation of this manuscript, including but not limited to editorial, technical, analytical, or writing support. All such contributions must be clearly stated in the Acknowledgments section.

Response: A concise Acknowledgments section was added. It states that language-editing tools were used for formatting, grammar, and clarity checks during revision, and that the author reviewed the manuscript, audit interpretations, tables, and conclusions and remains responsible for the final content. The acknowledgment was kept intentionally brief so it addresses the disclosure request without overstating the role of such tools.

Comment: Include enough recent references along with foundational ones. Ensure references directly support your claims. Avoid “padding.” Use credible sources (peer-reviewed journals, reputable books, official reports).

Response: The manuscript now includes a short reference-selection statement in the Related Work section. It clarifies that the reference list combines foundational sources on truthfulness, claim verification, NLI, and semantic similarity with recent 2024-2025 work on factuality and hallucination evaluation. No new citations were added merely to increase the reference count. The retained references directly support the surrounding claims about truthfulness benchmarks, reference-based verification, NLI/embedding-based detection, recent factuality evaluation, and retrieval-augmented hallucination evaluation.

Comment: All assumptions (including implicit assumptions) must be explicitly stated and clearly justified. The authors should explain why each assumption is reasonable and discuss its impact on the results and conclusions.

Response: A concise Assumptions paragraph was added in the Discussion. It explicitly states the main assumptions behind the analysis: that the reference answers are sufficiently correct for reference-based contradiction scoring; that archived endpoint identifiers and recorded outputs represent the evaluated behavior during the collection window; that the contradiction-evidence policy measures explicit contradiction risk rather than full answer quality; that adjudicated audit labels are treated as final labels while agreement statistics indicate residual subjectivity; and that repeated prompts or repeated prompt-response pairs should not be treated as independent evidence. The paragraph also explains the effect of each assumption on interpretation.

Comment: Avoid overreaching conclusions that extend beyond what is supported by the data and analysis.

Response: The Discussion and Conclusion were revised to reduce overreach. The manuscript now describes the results as lower detector-labeled contradiction rates on this archived benchmark, rather than broad evidence about hallucination behavior in all settings. The conclusion also emphasizes that the findings are benchmark-specific and that the detector measures explicit contradiction risk rather than general answer quality or full semantic adequacy.

Comment: Verify that you have used past perfect tense and third person throughout the manuscript wherever applicable.

Response: A targeted style pass was performed. First-person phrasing such as “we present,” “we evaluate,” “we compute,” “we observe,” and “our approach” was changed to third-person phrasing such as “this study presents,” “the study evaluates,” “the analysis computes,” “the results show,” and “the approach used here.” The pass was intentionally limited to places where tense or perspective affected manuscript style, so technical content, metrics, tables, and claims were not otherwise changed.

While the revision substantially improves transparency, annotation procedures, duplicate handling, and scope calibration, the newly added non-atomic benchmark reveals a fundamental limitation of the proposed approach. Detector precision falls to 13.5% on prompts involving explanation, synthesis, causal reasoning, citation-grounded reasoning, and uncertainty management. These are precisely the contexts in which contradiction detection becomes scientifically meaningful beyond simple answer-key matching. Consequently, the evidence presented supports, at most, a narrow reference-based disagreement detector for atomic short-answer tasks. The manuscript does not convincingly demonstrate a general contradiction detector, nor does it establish utility for broader hallucination evaluation. Because the central objective is not supported by the results, I do not believe the manuscript is suitable for publication in its current form.

HOWEVER: If you are agreeable to a pivot from “Here is a contradiction detector that works.” (it does not) to

“Here is a contradiction detector that appears to work in atomic QA but fails dramatically when moved to more realistic reasoning contexts. Here is why.”, that may actually be more novel and more scientifically valuable. At this point, you are no longer pursuing the objective that the detector works. You would be explaining why it doesn’t work when confronted with the very conditions that define realistic hallucination evaluation.

Many papers report performance numbers. Far fewer explain:
where a method breaks,

why it breaks,
what assumptions were hidden in the benchmark,
what properties of the task drive failure.
The 13.5% precision result is potentially a major finding because it suggests that contradiction detection is entangled with:
uncertainty expression,
hedging,
abstention,
source-limited reasoning,
partial correctness,
explanation structure.
Those are precisely the difficult issues in hallucination evaluation. If the author can systematically show that the detector fails because it mistakes:
uncertainty for contradiction,
caveats for contradiction,
abstentions for contradiction,
source limitations for contradiction,
then the paper begins to contribute knowledge about the limitations of contradiction-based hallucination detection.

I would want a rigorous failure analysis.

For example:

False-positive taxonomy

Of the 64 enhanced-dataset false positives:

25 due to uncertainty language

14 due to source-limitation statements

11 due to paraphrase failures

8 due to causal reasoning failures

6 due to synthesis failures

(or whatever the actual numbers are)

Then quantify them.

Feature attribution

Which detector component fired?

NLI contradiction veto?

numeric conflict?

atomic mismatch heuristic?

unavailability heuristic?

For each false-positive class.

Prompt-type breakdown

Does performance differ between:

causal reasoning

synthesis

citation-grounded prompts

uncertainty prompts

long-form explanation

The paper currently lumps all enhanced prompts together. That obscures the mechanism. You will need to generate **many more specific prompts for each of these categories** above to obtain sufficient statistical power for subsequent analysis.

The manuscript already hints at something important.

It notes that many false positives arise from uncertainty-oriented prompts where the model says things like: information unavailable, evidence insufficient, source outdated, and the detector interprets these as contradiction signals.

If demonstrated rigorously, that becomes an important observation: Contradiction detectors may systematically penalize calibration and uncertainty expression. That is a publishable insight.

Formal recommendation:

The revised manuscript substantially improves transparency and addresses most prior reviewer concerns. However, the expanded benchmark reveals a fundamental and potentially important limitation of the proposed detector. Rather than treating the 13.5% precision on the enhanced benchmark as a secondary stress-test result, the manuscript should investigate and explain the causes of this failure in a systematic manner. A detailed false-positive taxonomy, detector-component analysis, and prompt-category breakdown could transform this limitation into a meaningful scientific contribution regarding the boundaries of contradiction-based hallucination detection. Without such analysis, the manuscript's central claim remains insufficiently supported. With it, the manuscript could make a valuable contribution by identifying when and why contradiction detectors fail.

While the revision substantially improves transparency, annotation procedures, duplicate handling, and scope calibration, the newly added non-atomic benchmark reveals a fundamental limitation of the proposed approach. Detector precision falls to 13.5% on prompts involving explanation, synthesis, causal reasoning, citation-grounded reasoning, and uncertainty management. These are precisely the contexts in which contradiction detection becomes scientifically meaningful beyond simple answer-key matching. Consequently, the evidence presented supports, at most, a narrow reference-based disagreement detector for atomic short-answer tasks. The manuscript does not convincingly demonstrate a general contradiction detector, nor does it establish utility for broader hallucination evaluation. Because the central objective is not supported by the results, I do not believe the manuscript is suitable for publication in its current form.

HOWEVER: If you are agreeable to a pivot from "“Here is a contradiction detector that works.” (it does not) to

“Here is a contradiction detector that appears to work in atomic QA but fails dramatically when moved to more realistic reasoning contexts. Here is why.”, that may actually be more novel and more scientifically valuable. At this point, you are no longer pursuing the objective that the detector works. You would be explaining why it doesn't work when confronted with the very conditions that define realistic hallucination evaluation.

Many papers report performance numbers. Far fewer explain:

where a method breaks,

why it breaks,

what assumptions were hidden in the benchmark,

what properties of the task drive failure.

The 13.5% precision result is potentially a major finding because it suggests that contradiction detection is entangled with:

uncertainty expression,

hedging,

abstention,

source-limited reasoning,

partial correctness,

explanation structure.

Those are precisely the difficult issues in hallucination evaluation. If the author can systematically show that the detector fails because it mistakes:

uncertainty for contradiction,

caveats for contradiction,

abstentions for contradiction,
source limitations for contradiction,
then the paper begins to contribute knowledge about the limitations of contradiction-based hallucination detection.

I would want a rigorous failure analysis.

For example:

False-positive taxonomy

Of the 64 enhanced-dataset false positives:

25 due to uncertainty language

14 due to source-limitation statements

11 due to paraphrase failures

8 due to causal reasoning failures

6 due to synthesis failures

(or whatever the actual numbers are)

Then quantify them.

Feature attribution

Which detector component fired?

NLI contradiction veto?

numeric conflict?

atomic mismatch heuristic?

unavailability heuristic?

For each false-positive class.

Prompt-type breakdown

Does performance differ between:

causal reasoning

synthesis

citation-grounded prompts

uncertainty prompts

long-form explanation

The paper currently lumps all enhanced prompts together. That obscures the mechanism. You will need to generate **many more specific prompts for each of these categories** above to obtain sufficient statistical power for subsequent analysis.

The manuscript already hints at something important.

It notes that many false positives arise from uncertainty-oriented prompts where the model says things like: information unavailable, evidence insufficient, source outdated, and the detector interprets these as contradiction signals.

If demonstrated rigorously, that becomes an important observation: Contradiction detectors may systematically penalize calibration and uncertainty expression. That is a publishable insight.

Formal recommendation:

The revised manuscript substantially improves transparency and addresses most prior reviewer concerns. However, the expanded benchmark reveals a fundamental and potentially important limitation of the proposed detector. Rather than treating the 13.5% precision on the enhanced benchmark as a secondary stress-test result, the manuscript should investigate and explain the causes of this failure in a systematic manner. A detailed false-positive taxonomy, detector-component analysis, and prompt-category breakdown could transform this limitation into a meaningful scientific contribution regarding the boundaries of contradiction-based hallucination detection. Without such analysis, the manuscript's central claim remains insufficiently supported. With it, the manuscript could make a valuable contribution by identifying when and why contradiction detectors fail.

Response: The manuscript was substantially revised to address this concern by reframing the enhanced non-atomic benchmark result as a central finding rather than as a secondary limitation. The revised manuscript no longer presents HalluDetector as a broadly successful general hallucination or semantic-adequacy detector. It argues that the detector performs strongly on the legacy short-answer audit subset but has systematic limitations in an informative way when applied to enhanced non-atomic prompts involving explanation, causal reasoning, synthesis, citation-grounded answering, and uncertainty management.

Specifically, the abstract, contributions, Results and Discussion, and conclusion were revised to emphasize that the enhanced audit exposes a boundary condition for contradiction-based hallucination detection. The manuscript reports 84.6% precision, 94.8% recall, and 89.4% F1 on the legacy short-answer audit subset, and 13.5% precision and 23.5% F1 on the enhanced non-atomic audit subset. This result is interpreted as evidence that contradiction detectors can remain recall-oriented while becoming over-sensitive when answers contain uncertainty, source limitations, cautious phrasing, or partial but non-contradictory information.

To make this failure analysis more rigorous, the revised manuscript adds the three analyses requested by the reviewer. First, it adds a prompt-category breakdown (Table 4) for the enhanced subset, separating long-form explanation, causal reasoning, synthesis, citation-grounded answering, and nuanced uncertainty prompts. Second, it adds a false-positive taxonomy (Table 5) for the 64 enhanced-subset false positives, separating cases involving source/current-evidence limitations, cautionary or forbidden-claim mentions, and supportable paraphrase or compressed-answer cases. Third, it adds a detector-component analysis (Table 6) for which part fired during those false positives. Additionally, in the discussion, there is a new practical improvements paragraph detailing next steps for future work, discussing guardrails and scope checks regarding assertions, as well as small fixes for the legacy set.

These revisions change the manuscript’s central claim. The revised paper no longer argues that the detector demonstrates broad utility for general hallucination evaluation. Instead, it presents a more limited but scientifically useful finding: explicit contradiction detection works best under compact reference-based short-answer conditions, while more realistic non-atomic prompts reveal systematic false positives caused by uncertainty expression, source-limited reasoning, hedging, and cautious non-answers. The conclusion now states that this failure mode is important for hallucination evaluation because calibrated answers often contain the same linguistic cues that contradiction detectors may over-penalize.

We acknowledge that additional large-scale, category-balanced prompt generation would strengthen future statistical power and allow stronger population-level claims about non-atomic hallucination evaluation. However, the revised manuscript does not make such a broad generalization claim. Instead, it uses the existing 400-row enhanced audit and 64 enhanced false positives as a diagnostic stress test to identify, quantify, and explain a concrete failure mode within the submitted benchmark. This limitation is now explicitly acknowledged as future work rather than treated as resolved.

While the addition of the prompt-category breakdown (Table 4) is appreciated, it does not address my principal concern. My recommendation was not merely to separate prompt categories, but to generate sufficient category-specific prompts to achieve adequate statistical power for comparing detector performance across explanation, causal reasoning, synthesis, citation-grounded, and uncertainty tasks. The current analysis remains underpowered, with several categories containing no adjudicated contradiction-positive examples, rendering recall and F1 undefined and preventing meaningful category-level inference. Since the prompt-generation framework, contradiction detector, and adjudication pipeline are already implemented, expanding the benchmark is a straightforward extension of the existing work rather than a new methodological contribution. I therefore need to see a statistically powered category-balanced evaluation before concluding that the reported failure modes are robust rather than artifacts of the current sample.

Although the manuscript has substantially improved and addresses most previous concerns, the central claim regarding which reasoning categories systematically challenge contradiction-based hallucination detection still lacks adequate statistical support. This remaining issue is sufficiently important that it should be resolved before publication rather than deferred to future work.

While the addition of the prompt-category breakdown (Table 4) is appreciated, it does not address my principal concern. My recommendation was not merely to separate prompt categories, but to generate sufficient category-specific prompts to achieve adequate statistical power for comparing detector performance across explanation, causal reasoning, synthesis, citation-grounded, and uncertainty tasks.

The current analysis remains underpowered, with several categories containing no adjudicated contradiction-positive examples, rendering recall and F1 undefined and preventing meaningful category-level inference. Since the prompt-generation framework, contradiction detector, and adjudication pipeline are already implemented, expanding the benchmark is a straightforward extension of the existing work rather than a new methodological contribution. I therefore need to see a statistically powered category-balanced evaluation before concluding that the reported failure modes are robust rather than artifacts of the current sample.

Although the manuscript has substantially improved and addresses most previous concerns, the central claim regarding which reasoning categories systematically challenge contradiction-based hallucination detection still lacks adequate statistical support. This remaining issue is sufficiently important that it should be resolved before publication rather than deferred to future work.

Response: The manuscript has accordingly been revised to distinguish the category-level comparison that the expanded data support from the comparison they do not support. These revisions resolve the original uneven-sampling problem and provide a balanced, reproducible, independently annotated category comparison for detector over-triggering. They also narrow the manuscript's central claim to the inference supported by the data: citation-grounded answering and nuanced uncertainty systematically produce higher false-positive rates under the implemented contradiction detector. The manuscript no longer presents category-specific recall or F1 as statistically powered where contradiction-positive examples remain sparse.

The non-atomic prompt bank was expanded to 500 prompts, with 100 prompts in each of the five categories: long-form explanation, causal reasoning, synthesis, citation-grounded answering, and nuanced uncertainty. The balanced add-on contains 80 new prompts per category evaluated across seven archived runs, producing 560 responses per category and 2,800 responses overall.

A balanced 1,000-row audit was constructed with 200 audited rows per category. All 298 detector-positive responses in the 2,800-response frame were included. Detector-negative responses were sampled without replacement within category using random seed 42 until each category contained 200 audited rows. The manuscript now reports the category-specific sampling frame, inclusion probabilities, inverse-probability weights, unique prompts represented, endpoint composition, and exact run-level composition. Precision is calculated directly over the complete detector-positive frame; false-positive rate and specificity use inverse-probability weighting for the sampled detector-negative rows.

Annotation reliability is now quantified for the balanced add-on. Two human annotators independently labeled all 1,000 rows while blinded to detector output and endpoint identity, after which disagreements were resolved by adjudication. The annotators agreed on 968 of 1,000 multiclass labels (96.8%), with Cohen's kappa of 0.866. They agreed on the binary contradiction label for all 1,000 rows (100.0%), with kappa of 1.000. Category-level multiclass agreement ranged

from 94.5% to 100.0%, with kappa values from 0.761 to 1.000. All 32 balanced-add-on disagreements concerned the boundary between `supported` and `incomplete\_or\_vacuous`; none changed the binary contradiction label. Both pre-adjudication labels, rationales, agreement status, and final adjudicated labels are included in the revised audit file.

The central category comparison has been changed from recall/F1 to false-positive behavior, because false-positive rate and specificity can be estimated for every category without requiring contradiction-positive cases. The revised Table 4 reports inverse-probability-weighted false-positive rates with prompt-clustered 95% confidence intervals. Citation-grounded answering has a false-positive rate of 18.6% (95% CI 11.5%-26.6%), and nuanced uncertainty has a rate of 18.6% (95% CI 13.8%-24.3%). The corresponding rates are 0.9% for causal reasoning, 5.2% for long-form explanation, and 6.0% for synthesis. The revised claim is therefore that citation-grounded and uncertainty prompts produce substantially greater detector over-triggering than the other balanced categories.

The manuscript now explicitly states that the balanced design does not create contradiction-positive events. The observed contradiction-positive counts remain 15 for long-form explanation, 9 for synthesis, 1 for nuanced uncertainty, and 0 for causal reasoning and citation-grounded answering. Category-specific recall and F1 are therefore not used for comparative inference. The manuscript does not claim that the revision provides a powered comparison of category-specific recall, and has a statistically supported claim that category-specific false-positive behavior differs substantially across the balanced non-atomic categories.

The false-positive taxonomy and detector-component attribution were recalculated for all 274 balanced-add-on false positives. The taxonomy shows that 167 false positives involve prohibited or unsupported claims being mentioned rather than endorsed, and 70 involve source or evidence limitations. These mechanisms are concentrated in the categories with the highest weighted false-positive rates, providing qualitative support for the category-level statistical result.

Thank you for addressing my comments. The manuscript needs one final conceptual reframing. The revised manuscript appropriately follows my earlier recommendation to treat the negative result as the principal scientific finding. I therefore do not object to the fact that the work now emphasizes detector failure rather than successful general hallucination detection. However, I believe the manuscript still places too much emphasis on the comparison among individual non-atomic prompt categories. The more informative result is broader: a detector showing 84.6% precision on the legacy short-answer benchmark falls to only 8.1% precision in the balanced non-atomic setting. This demonstrates that strong performance on compact reference-answer benchmarks does not necessarily transfer to more realistic discourse in which propositions may be mentioned, qualified, rejected, or discussed under evidentiary uncertainty.

The category analysis is useful primarily because it helps explain this collapse. The false-positive taxonomy indicates that the dominant errors arise when unsupported or prohibited claims are mentioned without endorsement, or when responses discuss source and evidence limitations. The manuscript should therefore frame its principal contribution as identifying a benchmark-transfer failure and its linguistic mechanism, rather than as establishing that certain prompt categories simply have higher false-positive rates.

In my view, this would preserve the negative-result direction I previously recommended while giving that negative result a clearer scientific contribution. The important conclusion is not merely that HalluDetector performs poorly on citation-grounded or uncertainty prompts, but that contradiction-detection performance measured on atomic benchmarks can substantially

overestimate reliability in non-atomic language settings because the detector fails to distinguish proposition content from its assertion or epistemic status.

Comment:

Thank you for addressing my comments. The manuscript needs one final conceptual reframing. The revised manuscript appropriately follows my earlier recommendation to treat the negative result as the principal scientific finding. I therefore do not object to the fact that the work now emphasizes detector failure rather than successful general hallucination detection. However, I believe the manuscript still places too much emphasis on the comparison among individual non-atomic prompt categories. The more informative result is broader: a detector showing 84.6% precision on the legacy short-answer benchmark falls to only 8.1% precision in the balanced non-atomic setting. This demonstrates that strong performance on compact reference-answer benchmarks does not necessarily transfer to more realistic discourse in which propositions may be mentioned, qualified, rejected, or discussed under evidentiary uncertainty.

The category analysis is useful primarily because it helps explain this collapse. The false-positive taxonomy indicates that the dominant errors arise when unsupported or prohibited claims are mentioned without endorsement, or when responses discuss source and evidence limitations. The manuscript should therefore frame its principal contribution as identifying a benchmark-transfer failure and its linguistic mechanism, rather than as establishing that certain prompt categories simply have higher false-positive rates.

In my view, this would preserve the negative-result direction I previously recommended while giving that negative result a clearer scientific contribution. The important conclusion is not merely that HalluDetector performs poorly on citation-grounded or uncertainty prompts, but that contradiction-detection performance measured on atomic benchmarks can substantially overestimate reliability in non-atomic language settings because the detector fails to distinguish proposition content from its assertion or epistemic status.

Response:

In the Abstract, the 84.6%-to-8.1% precision collapse is now shown as the principal negative result, while the category-specific false-positive rates are presented as evidence that helps explain the broader transfer failure. The Abstract also states the linguistic mechanism indicated by the false-positive taxonomy, that the detector can confuse proposition content with its assertion or epistemic status when unsupported or prohibited claims are mentioned without endorsement or when source and evidence limitations are discussed.

In the Introduction, it now describes the main limitation as a benchmark-transfer boundary and states explicitly that the cross-setting precision collapse is the principal negative finding. In the Contributions paragraph, the fifth contribution has been reframed around identifying the benchmark-transfer failure and using the category estimates and false-positive taxonomy to characterize its linguistic mechanism.

In Results and Discussion, immediately before Table 2, the 84.6%-to-8.1% precision drop is now identified as the principal cross-setting result. The following binary-error paragraph now explains that the category concentration is diagnostically useful because these prompts create discourse in which propositions may be mentioned, rejected, or qualified without being asserted. The subsection previously titled “Balanced non-atomic category analysis” is now titled “Diagnostic category analysis in the balanced non-atomic setting.” Within that subsection, the category-specific false-positive rates are explicitly framed as localizing the broader transfer failure instead of constituting the paper’s principal contribution. The interpretation of the false-positive taxonomy now emphasizes the common mechanism across categories, and it states confusion between proposition content and assertion or epistemic status. The subsection summary likewise states that the primary inference is the benchmark-transfer failure, with the category analysis used to localize it and the taxonomy used to explain it.

The opening of the Discussion now states directly that the principal scientific finding is the decline in measured precision from 84.6% on the legacy short-answer audit to 8.1% in the balanced non-

atomic setting. It then uses category analysis and taxonomy to explain the mechanism, instead of presenting category differences as the endpoint of the claim.

The Conclusion now centers the benchmark-transfer failure and states that compact-benchmark contradiction-detection performance can overestimate reliability in non-atomic language settings when proposition content is confused with assertion or epistemic status. The category-specific false-positive rates remain in the Conclusion as explanatory evidence, and the manuscript continues to avoid a powered category-specific recall claim because contradiction-positive examples remain sparse.

Please provide a .docx or .odt file of the paper correctly formatted to the Journal's specifications (see website).

Once we receive this file, we will officially accept your manuscript.

Comment: Please provide a .docx or .odt file of the paper correctly formatted to the Journal's specifications (see website).

Once we receive this file, we will officially accept your manuscript.

Response: The manuscript was reformatted according to the JHSS guidelines. This includes adding section numbers, making sure tables are numbered correctly, making sure equations are formatted correctly, and leaving figures and figure legends separate from the manuscript. No content was modified.

Thank you for providing the docx file. Accepted.