

## Peer-Review

Luo, Haoning. 2026. "An Investigation into LLMs' Abilities in Discerning between True and False Information." *Journal of High School Science* 10 (3): 484–504. <https://doi.org/10.64336/001c.169559>.

A good attempt. I propose an addition to your work which will elevate its already novel nature. Define 'Semantic truth' as a statement which is factually correct. Define 'epistemic truth' as whether believing the statement is justified given the available evidence.

Propose a two axis (four quadrant) model of LLM truth based on epistemic and semantic truths. Semantic truth (ST) on the X-axis and epistemic truth (ET) on the Y-axis. Hence, quadrant I is both ST and ET. Quadrant II is ET but not ST. Quadrant III is neither ET nor ST and Quadrant IV is ST but not ET.

X-axis: Semantic Truth (Grounded Correctness): Degree to which a statement corresponds to objective reality. Determined by: verifiable external evidence, authoritative sources, ground truth labels, Binary or continuous (e.g., % factual accuracy)

Y-axis: Epistemic Truth (Justified Believability): Degree to which an agent (human or LLM) is justified in believing the statement given available information. Determined by: linguistic coherence, internal consistency, familiarity with known patterns, prior probability in training data and naturally graded and probabilistic

Quadrant I: High Semantic / High Epistemic (Ideal Truth): Factually correct, Plausible and well-justified.

Examples: Well-established scientific facts, Neutral, accurate news reporting, Textbook knowledge.

LLM behavior: High confidence, Correct explanations, Appropriate reliability scores

Interpretation: LLM performance is best here

Quadrant II: Low Semantic / High Epistemic (Plausible Falsehoods): Factually incorrect, Highly believable

Examples: Fluent misinformation, Hallucinated but stylistically accurate articles, False claims aligned with common narratives

LLM behavior: High reliability scores, Confident explanations, Overconfidence failures

Interpretation: This is the most dangerous quadrant.

Quadrant III: Low Semantic / Low Epistemic (Obvious Falsehoods): Factually incorrect, Implausible

Examples: Self-contradictory claims, Nonsense text, Easily falsifiable statements

LLM behavior: Low confidence, Rejection or uncertainty, Correct skepticism

Quadrant IV: High Semantic / Low Epistemic (Unfamiliar Truths): Factually correct, Weakly justified given priors

Examples: Obscure historical facts, Counterintuitive scientific results, New discoveries

LLM behavior: Low or inconsistent confidence, Hedged language, Underestimation of reliability

Map your results into this two-axes model

False articles scoring >3.0 → Quadrant II

Biased but factually accurate texts scoring intermediately → Transition between Quadrants I and IV

Older/historical sources rated more reliably → Semantic truth and epistemic familiarity aligned

Score-explanation conflict → Epistemic reasoning flags issues, but scalar confidence remains high

This reframes "LLM failure" as axis misalignment, not random error.

A Single Scalar Score Is Fundamentally Insufficient because it collapses these two orthogonal dimensions.

Implications for Evaluation and Design

Evaluation

Fact-checking benchmarks should separately measure:

semantic accuracy (external verification) and epistemic confidence (model-internal belief).

Miscalibration is often axis confusion, not ignorance

Model Design

Future systems could:output (semantic\_estimate, epistemic\_confidence) pairs, flag Quadrant II outputs automatically, route Quadrant IV claims for external verification

Here is a one paragraph summary:

We propose a two-axis framework separating semantic truth (correspondence with external reality) from epistemic truth (justified belief given available information). Large language models naturally operate in epistemic space, estimating plausibility rather than factual correctness. Many apparent fact-checking failures arise when high epistemic confidence coincides with low semantic truth, producing fluent but incorrect outputs. Collapsing these dimensions into a single scalar reliability score obscures this distinction and leads to systematic overconfidence. Separating these axes provides a clearer interpretation of LLM behavior and suggests more robust evaluation and deployment strategies.

Apply it to your database.

**Manuscript title:** An Investigation into LLMs' Abilities in Discerning Between True and False Information

I thank the editor and reviewer for the constructive comments. I revised the manuscript to incorporate the proposed two-axis framework, strengthened Phase 2 of the experimental method, expanded the statistical analyses, and clarified how the updated results should be interpreted. The main revisions are summarized below.

| Comment             | Reviewer comment listed verbatim                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Response and location in revised manuscript                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| of Semantic Truth   | Define ‘Semantic truth’ as a statement which is factually correct. Define ‘epistemic truth’ as whether believing the statement is justified given the available evidence.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Addressed. I revised the theoretical framework and definitions to distinguish Semantic Truth Score from Epistemic Truth Score. Semantic Truth Score is now defined as the degree to which a text's propositional content corresponds to external facts or states of affairs. Epistemic Truth Score is defined as whether the text appears credible given available information, including contextual and evidential presentation, stylistic consistency, plausibility, and training-data familiarity.<br><br>Location in manuscript: Introduction; Section 2.3 Variables; Section 2.4 LLM Settings and Prompting Procedures; Section 4.3 Epistemic Credibility and Tarski's Semantic Conception of Truth.                                                                                    |
| Four-quadrant model | Propose a two axis (four quadrant) model of LLM truth based on epistemic and semantic truths. Semantic truth (ST) on the X-axis and epistemic truth (ET) on the Y-axis. Hence, quadrant I is both ST and ET. Quadrant II is ET but not ST. Quadrant III is neither ET nor ST and Quadrant IV is ST but not ET.<br><br>X-axis: Semantic Truth (Grounded Correctness):Degree to which a statement corresponds to objective reality. Determined by:verifiable external evidence, authoritative sources, ground truth labels, Binary or continuous (e.g., % factual accuracy)<br>Y-axis: Epistemic Truth (Justified Believability): Degree to which an agent (human or LLM) is justified in believing the statement given available information. Determined by:linguistic coherence, internal consistency, familiarity | Addressed. I revised Phase 2 so that truth evaluation is explicitly treated as a two-dimensional framework rather than a single reliability score. The revised manuscript describes ST as grounded factual correspondence and ET as justified believability based on available information. I added interpretation of results through semantic-epistemic alignment and mismatch, which operationalizes the reviewer's proposed two-axis model in the empirical analysis.<br><br>Location in manuscript: Introduction contribution statement; Section 2.3 Variables; Section 2.4 Prompting Procedures; Section 3 Results, especially Tables 3-8; Section 4 Training Data Coverage and Semantic Truth Evaluation; Section 4.3 Epistemic Credibility and Tarski's Semantic Conception of Truth. |

|                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|----------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                    | with known patterns, prior probability in training data and naturally graded and probabilistic                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <b>Comment on the model base</b>                   | <p>Map your results into this two-axes model</p> <p>False articles scoring &gt;3.0 → Quadrant II</p> <p>Biased but factually accurate texts scoring intermediately → Transition between Quadrants I and IV</p> <p>Older/historical sources rated more reliably → Semantic truth and epistemic familiarity aligned</p> <p>Score–explanation conflict → Epistemic reasoning flags issues, but scalar confidence remains high</p> <p>This reframes “LLM failure” as axis misalignment, not random error.</p>                                                                                                                                                                                                                                                                 | <p>Addressed. I expanded the Results and Discussion to integrate the empirical patterns into the two-axis interpretation. The updated analysis shows that fake texts often retain substantial epistemic credibility despite low semantic truth, especially in the newer CNN condition. Older CNN and historical texts showed stronger alignment between semantic truth and epistemic credibility, consistent with the point about familiarity and training-data coverage. I retained and interpreted the score–explanation conflict as evidence that a single scalar reliability score can mask the tension between semantic and epistemic judgments.</p> <p>Location in manuscript: Section 3 Results, Tables 1–3; Section 4.1 Data Analysis; Section 4.2; Section 4.3; Conclusion.</p> |
| <b>Comment on the single scalar score argument</b> | A Single Scalar Score Is Fundamentally Insufficient because it collapses these two orthogonal dimensions.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | <p>Addressed. I revised the manuscript's central interpretation to state that the Phase 1 single reliability score is less useful because it collapses factual correspondence and justified believability into one number. Location in manuscript: Abstract; Introduction; Section 3 Results, Table 8; Section 4.3; Conclusion.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| <b>Comments for the final design</b>               | <p>Implications for Evaluation and Design</p> <p>Evaluation</p> <p>Fact-checking benchmarks should separately measure: semantic accuracy (external verification) and epistemic confidence (model-internal belief). Miscalibration is often axis confusion, not ignorance</p> <p>Model Design</p> <p>Future systems could: output (semantic_estimate, epistemic_confidence) pairs, flag Quadrant II outputs automatically, route Quadrant IV claims for external verification</p>                                                                                                                                                                                                                                                                                          | <p>Addressed. I added implications for LLM evaluation and deployment. The revised manuscript argues that fact-checking systems should separately evaluate factual correspondence and epistemic credibility. It also clarifies that newer or unfamiliar but potentially true claims should be routed to external verification, while plausible but semantically weak outputs should be treated as high-risk cases.</p> <p>Location in manuscript: Section 4.3; Section 4.4 Implications and Future Works; Conclusion.</p>                                                                                                                                                                                                                                                                 |
| <b>Comment on the one paragraph summary</b>        | <p>Here is a one paragraph summary:</p> <p>I propose a two-axis framework separating semantic truth (correspondence with external reality) from epistemic truth (justified belief given available information). Large language models naturally operate in epistemic space, estimating plausibility rather than factual correctness. Many apparent fact-checking failures arise when high epistemic confidence coincides with low semantic truth, producing fluent but incorrect outputs. Collapsing these dimensions into a single scalar reliability score obscures this distinction and leads to systematic overconfidence. Separating these axes provides a clearer interpretation of LLM behavior and suggests more robust evaluation and deployment strategies.</p> | <p>Addressed. I used this summary as the conceptual foundation for the revisions, while adapting the language to match the manuscript's empirical evidence. The revised manuscript now emphasizes that the two-axis framework provides a more interpretable evaluation than a single reliability score, and that LLM behavior depends on both textual plausibility and training-data coverage.</p> <p>Location in manuscript: Abstract; Introduction conclusion paragraph; Section 4.3; Conclusion.</p>                                                                                                                                                                                                                                                                                  |
| <b>Comment on applying to the database</b>         | Apply it to your database.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | <p>Addressed. I updated Phase 2 and reanalyzed the data using Semantic Truth Scores and Epistemic Truth Scores across authentic and fabricated texts. I added descriptive statistics, authentic-fake comparisons, paired statistical tests, and a semantic-epistemic gap analysis, source-level correlations.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |

and Phase 1 vs. Phase 2 comparisons. These analyses directly apply the two-axis framework to the data and show where the framework succeeds and where it has persistent limitations, especially for newer CNN models.

Location in manuscript: Section 2.5 Statistical Analysis; Section 3 Results, Tables 3-8; Section 4.2; Section 4.4

Thank you for addressing my comments. I found a very interesting and possibly very discriminative facet of your LLM assigned scores. Instead of taking the difference of ST to ET; you would take the ratio of ET to ST.

For example in Table 3, if you take the ET to ST ratio, you can see that semantic false cluster at high ratios whereas semantic true cluster at low ratios. I asked chatgpt to construct 10 points per row using mean and standard deviation for that particular row and plot those points on the quadrant. Except now, the horizontal axis is ST and the vertical axis is ET/ST ratio.

Please see the attached chatgpt file.

I think that if you use this ratio and quadrant axes, you can separate semantically true and semantically false statements using your LLM scores very well.

Please incorporate this into the manuscript. Also provide a diagram of the quadrants. With this information, your manuscript is poised to make a significant difference in the field.

I look forward to your revised manuscript.

### Response to Reviewer's Comments

I thank the reviewer for the constructive second-round suggestion. In this revision, I adapted the ET/ST ratio into the manuscript, and strengthened the discussion of this ratio-based framework as a practical approach for LLM truth verification.

| Reviewer comment listed verbatim                                                                                                                                                                                                                                    | Response and location in revised manuscript                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| "Instead of taking the difference of ST to ET; you would take the ratio of ET to ST. For example in Table 3, if you take the ET to ST ratio, you can see that semantic false cluster at high ratios whereas semantic true cluster at low ratios."                   | Addressed. I incorporated the ET/ST ratio into the manuscript as the Epistemic Inflation Ratio. The revised manuscript now explains that the ratio measures whether epistemic credibility is proportionate to semantic grounding. This strengthens the analysis by showing cases in which a text appears credible despite weak factual correspondence.<br><br>Location in manuscript: Section 2.5 Statistical Analysis; Section 3 Results, Table 7; Section 4.4 Epistemic Inflation Ratio and a Practical Verification Framework; Conclusion. |
| "I asked chatgpt to construct 10 points per row using mean and standard deviation for that particular row and plot those points on the quadrant. Except now, the horizontal axis is ST and the vertical axis is ET/ST ratio. Please see the attached chatgpt file." | Partially addressed. I added a new figure that plots Semantic Truth Score (ST) on the horizontal axis and ET/ST on the vertical axis. The figure uses shape-based markers to distinguish authentic and fabricated texts and is presented as a diagnostic visualization. However, I am unable to fully recreate the other plot provided by the reviewer, which is based on simulated data. After checking, I find that my experimental results cannot form a pattern as significant and identifiable as given by the                           |

| Reviewer comment listed verbatim                                                                                                                                     | Response and location in revised manuscript                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                                                                      | <p>reviewer's simulation (probably due to the fact that the response variables in the experiment are discrete). In addition, because the scores are all integers from 1 to 5, it would not be very informative and elegant to display the ET/ST values of all individual articles in a scatter plot.</p> <p>Location in manuscript: Section 3 Results, Figure 1; Section 4.4.</p>                                                                                                                                                                                                                                                                             |
| <p>"I think that if you use this ratio and quadrant axes, you can separate semantically true and semantically false statements using your LLM scores very well."</p> | <p>Partially addressed. The revised manuscript explains that authentic older CNN and historical texts cluster near semantic-epistemic alignment, while many fabricated texts show high ET/ST ratios.</p> <p>However, after double-checking my actual experimental data, I cannot reach the same conclusion that the reviewer obtained from the simulated data. Instead, I interpreted the real-world significance of the ET/ST ratio – it can be a measure for the risk of text being harmful by spreading misinformation.</p> <p>Location in manuscript: Section 3 Results, Table 7 and Figure 1; Section 4.4; Section 4.5 Limitations and Future Works.</p> |
| <p>"Please incorporate this into the manuscript. Also, provide a diagram of the quadrants."</p>                                                                      | <p>Same as above</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |

Thank you for incorporating the ET/ST framework. The manuscript is significantly improved.

1. The analysis remains at the group-average level. The manuscript would be strengthened substantially by presenting article-level ST versus ET/ST distributions rather than only mean-based diagnostic plots. You have an availability of approximately 548 labeled observations. ROC/AUC, sensitivity, specificity, and threshold analyses would provide a direct evaluation of the proposed metric and substantially strengthen the manuscript.

2. While the introduction of the ET/ST ratio is a valuable improvement, I encourage the author to reconsider the quadrant framework. The most important contribution of the revised manuscript is not the ratio itself, but the recognition that semantic truth and epistemic credibility are distinct dimensions of evaluation.

The current ST versus ET/ST plot emphasizes a derived metric and partially obscures the underlying two-dimensional structure. A direct ST versus ET representation with quadrant boundaries would better illustrate the conceptual framework introduced in Phase 2. Such a visualization reframes truth evaluation from a one-dimensional true/false classification problem into a two-dimensional calibration problem between factual grounding and perceived credibility. This perspective also naturally distinguishes grounded information, plausible falsehoods, poorly supported truths, and obvious falsehoods. The ET/ST ratio can then be interpreted as a secondary measure of semantic-epistemic misalignment within this broader framework.

I therefore encourage the author to consider presenting the ST-ET quadrant diagram alongside the ET/ST analysis, as it more clearly communicates the central conceptual contribution of the manuscript.

## Response to Reviewer's Comments

I sincerely thank the reviewer for the careful and constructive comments. The revision follows the reviewer's central recommendation by shifting the manuscript away from a ratio-centered interpretation and toward direct article-level evaluation of Semantic Truth (ST) and Epistemic Truth (ET). The revision adds article-level ROC and threshold analyses, introduces a direct ST–ET quadrant visualization, and clarifies that ST and ET are analytically distinct but empirically overlapping dimensions rather than fully independent measures.

Finally, I also wish my work could be published soon, as college applications will be starting this autumn.

| Reviewer comment listed verbatim                                                                                                                                                                                                                                                                                                                                                                                                                                              | Response and location in revised manuscript                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| “The analysis remains at the group-average level. The manuscript would be strengthened substantially by presenting article-level ST versus ET/ST distributions rather than only mean-based diagnostic plots. You have an availability of approximately 548 labeled observations. ROC/AUC, sensitivity, specificity, and threshold analyses would provide a direct evaluation of the proposed metric and substantially strengthen the manuscript.”                             | Addressed with article-level analysis. All 548 Phase 2 observations (274 authentic and 274 fabricated texts) are now retained and analyzed individually rather than represented only by group means. The revision reports ROC/AUC, sensitivity, specificity, accuracy, Youden's J, and exploratory classification thresholds for ST and ET. Location in manuscript: Section 2.5 Statistical Analysis; Section 3 Results, Table 7 and Figure 1; Section 4.5 Limitations and Future Works.                                                                           |
| “While the introduction of the ET/ST ratio is a valuable improvement, I encourage the author to reconsider the quadrant framework. The most important contribution of the revised manuscript is not the ratio itself, but the recognition that semantic truth and epistemic credibility are distinct dimensions of evaluation.”                                                                                                                                               | Addressed. The manuscript has been re-centered on the direct distinction between ST and ET. It now states explicitly that the two scores are analytically distinct but not assumed to be statistically or computationally independent. ST represents the model's prompted estimate of factual correspondence, whereas ET represents perceived credibility based on coherence, plausibility, evidential presentation, and related cues. Location in manuscript: Abstract; Introduction contribution paragraph; Section 2.3 Variables; Sections 4.2–4.4; Conclusion. |
| “The current ST versus ET/ST plot emphasizes a derived metric and partially obscures the underlying two-dimensional structure. A direct ST versus ET representation with quadrant boundaries would better illustrate the conceptual framework introduced in Phase 2. Such a visualization reframes truth evaluation from a one-dimensional true/false classification problem into a two-dimensional calibration problem between factual grounding and perceived credibility.” | Addressed. The group-mean ST-versus-ratio plot has been replaced with a direct article-level ST–ET representation. The new diagram plots every Phase 2 observation and uses descriptive midpoint boundaries to display the four regions of the framework. Location in manuscript: Section 4.4 Article-Level ST–ET Calibration Framework, Figure 2.                                                                                                                                                                                                                 |
| “This perspective also naturally distinguishes grounded information, plausible falsehoods, poorly supported truths, and obvious falsehoods. The ET/ST ratio can then be interpreted as a secondary measure of semantic-epistemic misalignment within this broader framework. I therefore encourage the author to consider presenting the ST-ET                                                                                                                                | Addressed with a methodological refinement. The revised manuscript adopts the direct ST–ET framework. After examining the article-level data, I concluded that the ST/ET does not provide an independent dimension. Location in manuscript: Section 3 Results, Table 7 and Figure 1; Section 4.4 and Figure 2; Conclusion.                                                                                                                                                                                                                                         |

|                                                                                                                                        |  |
|----------------------------------------------------------------------------------------------------------------------------------------|--|
| quadrant diagram alongside the ET/ST analysis, as it more clearly communicates the central conceptual contribution of the manuscript.” |  |
|----------------------------------------------------------------------------------------------------------------------------------------|--|

My last remaining comment has to do with validation of the ratio using the same method as that used when evaluating the difference.

Although the ET/ST ratio substantially strengthens the manuscript conceptually, its utility as a practical verification framework is not quantitatively demonstrated. The manuscript proposes ET/ST as a diagnostic measure for identifying potentially misleading content, yet no classification-performance analysis is presented. If ET/ST is intended to function as a verification or screening metric, the authors should evaluate its discriminatory ability using appropriate diagnostic performance measures such as sensitivity, specificity, precision, F1-score, and receiver operating characteristic (ROC) analysis with the corresponding area under the curve (AUC). Without such analyses, the manuscript demonstrates that ET/ST differs statistically between groups but does not establish how effectively it classifies authentic versus fabricated texts or how it compares with the original reliability score.

### Response to Reviewer’s Comments

I sincerely thank the reviewer for the careful and constructive comments. The comments about ET/ST appear to refer to an earlier draft, yet the newest manuscript had already omitted the ratio and centered the analysis on direct article-level ST and ET evaluation. Nevertheless, the revision addresses the underlying concern by strengthening the theoretical basis of the ST and ET framework, retaining article-level ROC and threshold analyses.

| Reviewer comment                                                                                                                                                                                                                                                                                                                                                                                                                                  | Response                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| “The analysis remains at the group-average level. The manuscript would be strengthened substantially by presenting article-level ST versus ET/ST distributions rather than only mean-based diagnostic plots. You have an availability of approximately 548 labeled observations. ROC/AUC, sensitivity, specificity, and threshold analyses would provide a direct evaluation of the proposed metric and substantially strengthen the manuscript.” | Addressed with article-level analysis. All 548 Phase 2 observations (274 authentic and 274 fabricated texts) are analyzed individually. The manuscript reports statistical results for measuring and comparing ST and ET. These analyses were already present in the newest draft and are retained with clearer interpretation.<br>Location in manuscript: Section 2.5 Statistical Analysis; Section 3 Results, Table 7 and Figure 1; Section 4.5 Limitations and Future Works. |
| “While the introduction of the ET/ST ratio is a valuable improvement, I encourage the author to reconsider the quadrant framework. The most important contribution of the revised manuscript is not the ratio itself, but the recognition that semantic truth and epistemic credibility are distinct dimensions of evaluation.”                                                                                                                   | Already addressed. The reference to ET/ST appears to concern an earlier draft: the newest manuscript had already omitted the ratio and re-centered the analysis on ST and ET. The Discussion clarifies that ST and ET are conceptually distinct but empirically overlapping.<br>Location in manuscript: Introduction; Section 2.3 Variables; Sections 4.3–4.4; Conclusion.                                                                                                      |
| “The current ST versus ET/ST plot emphasizes a derived metric and partially obscures the underlying two-dimensional structure. A direct ST versus ET representation with quadrant boundaries would better illustrate the conceptual framework introduced in Phase 2. Such a visualization reframes truth evaluation from a one-dimensional true/false                                                                                             | Already addressed. The earlier group-mean ST-versus-ratio plot has been replaced by a direct article-level ST–ET representation. Every Phase 2 observation is plotted, and descriptive midpoint boundaries show the four regions of the framework. The Discussion maps the observed source-level patterns onto this space rather than treating the diagram as a separate classifier.<br>Location in manuscript: Section 4.4 ST–ET Calibration                                   |

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| classification problem into a two-dimensional calibration problem between factual grounding and perceived credibility.”                                                                                                                                                                                                                                                                                                                                                               | Framework, Figure 2.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| “This perspective also naturally distinguishes grounded information, plausible falsehoods, poorly supported truths, and obvious falsehoods. The ET/ST ratio can then be interpreted as a secondary measure of semantic-epistemic misalignment within this broader framework. I therefore encourage the author to consider presenting the ST-ET quadrant diagram alongside the ET/ST analysis, as it more clearly communicates the central conceptual contribution of the manuscript.” | Addressed with a clarification. The direct ST-ET framework is retained as the central contribution. The ratio was not restored because the newest manuscript no longer proposes ET/ST as an independent verification measure; the underlying ST and ET dimensions are instead evaluated directly at the article level.<br>This follows the reviewer’s central recommendation while avoiding the reintroduction of another scalar metric.<br>Location in manuscript: Introduction contribution paragraph; Section 3 Results, Table 7 and Figure 1; Section 4.4 and Figure 2; Conclusion. |

Thank you for addressing my comments. I have one minor concern that could perhaps be addressed in the ‘future work’ subsection of the manuscript.

The revised manuscript has satisfactorily addressed my principal concerns regarding article-level analysis and the direct representation of Semantic Truth (ST) and Epistemic Truth (ET). I agree that the ST-ET framework should remain the primary representation. However, I believe there remains value in the previously proposed ET/ST ratio, not as an alternative truth classifier, but as a potential quantitative measure of **truth inflation**—that is, *the degree to which epistemic credibility becomes elevated relative to semantic truth*. The current ET-ST analysis captures the absolute discrepancy between these dimensions, whereas ET/ST could potentially capture their relative discrepancy. I suggest that the authors mention this possibility under Future Work and propose testing whether ET/ST, or another normalized measure of semantic-epistemic discordance, provides information beyond ST, ET, and ET-ST in independent datasets. This would preserve an interesting conceptual contribution of the earlier analysis without requiring additional analysis in the present manuscript.

I sincerely thank the reviewer for the positive assessment and constructive final suggestion. Since I am currently preparing university applications, I would be sincerely grateful if an editorial decision could be communicated at the earliest convenient time.

Reviewer commentResponse

have one minor concern that could perhaps be addressed in the ‘future work’ subsection of the manuscript. Addressed. In response to the reviewer’s suggestion, Section 4.5 now identifies ET/ST

and other normalized measures of semantic-epistemic discordance as a future candidate measures. The added text explains that independent datasets should test whether such a measure provides information beyond ST, ET, and the absolute ET-ST difference.

The revised manuscript has satisfactorily addressed my principal concerns regarding article-level analysis and the direct representation of Semantic Truth (ST) and Epistemic Truth (ET). I agree that the ST-ET framework should remain the primary representation. However, I believe there remains value in the previously proposed ET/ST ratio, not as an alternative truth classifier, but as a potential quantitative measure of truth inflation—that is, the degree to which epistemic credibility becomes elevated relative to semantic truth. The current ET-ST analysis captures the absolute discrepancy between these

dimensions, whereas ET/ST could potentially capture their relative discrepancy. I suggest that the authors mention this possibility under Future Work and propose testing whether ET/ST, or another normalized measure of semantic–epistemic discordance, provides information beyond ST, ET, and ET–ST in independent datasets. This would preserve an interesting conceptual contribution of the earlier analysis without requiring additional analysis in the present manuscript.

Location in manuscript: Section 4.5, Limitations and Future Works.

---

Thank you for addressing my comments. Accepted.