Journal of High School Science

An investigation into LLMs' abilities in discerning between true and false information

Luo H

Submitted: January 1, 2026, Revised: version 1, May 19, 2026, version 2, June 1, 2026, version 3, July 12, 2026, version 4, July 29, 2026, version 5, August 18, 2026
Accepted: August 19, 2026

Abstract
Large language models (LLMs) may judge whether information appears credible without necessarily establishing whether it corresponds to external facts. This study investigates this distinction by operationally separating Semantic Truth (ST), defined as the correspondence of textual claims with external states of affairs, from Epistemic Truth (ET), defined as the credibility or justification conveyed by a text through coherence, plausibility, evidential presentation, and consistency. The dataset comprised 274 source texts, including newer BBC and CNN articles, older CNN articles, and historical articles, from which controlled variants differing in factual accuracy and presentation were generated. In Phase 1, a single LLM reliability score remained relatively high even as factual accuracy decreased, with completely fabricated texts still receiving mean scores above 3 on a 1–5 scale. In some cases, the model also assigned high numerical reliability despite identifying substantial factual problems in its written justification. Phase 2 separately evaluated ST and ET across 548 authentic and fabricated observations. ST provided stronger discrimination between authentic and fabricated texts than ET, achieving an overall AUC of 0.801, sensitivity of 0.766, and specificity of 0.810. However, semantic discrimination varied markedly with information familiarity, ranging from near-chance performance for newer CNN articles to nearly perfect discrimination for older and historical material. These findings demonstrate that targeted semantic prompting improves factual discrimination but does not fully separate semantic correspondence from information familiarity, plausibility, and other non-factual textual cues. More broadly, the ST–ET framework exposes a potentially important form of truth inflation: epistemic credibility may remain high as semantic correspondence is progressively degraded through increasing fabrication. This provides a basis for future studies to determine how far epistemic credibility can be sustained or inflated as factual grounding deteriorates, thereby defining and quantifying an LLM's tolerance for increasingly plausible fabrication.

Keywords
Large Language Models, Misinformation detection, Fact-checking, Truthfulness evaluation, Semantic truth, Epistemic credibility, Fake news detection, Hallucination detection, Uncertainty calibration, Artificial intelligence

Haoning Luo, Shanghai Starriver Bilingual School, 2588 Jindu Rd., Shanghai, P. R. China 201108.
jerryluoofficial@icloud.com

https://doi.org/10.64336/001c.169559

1. Introduction

Online false rumors—sometimes called misinformation or disinformation—spread rapidly across social media and online platforms. If left unchecked, such content can greatly influence public opinion, affect decision-making, and sometimes cause social or political harm (1). As the volume of online information grows, it becomes increasingly challenging for humans alone to detect and address false content quickly and accurately (2).

At the same time, Large Language Models (LLMs) have become widely used tools for natural language understanding, content generation, and information retrieval. These models can analyze large amounts of text data and generate human-like responses. From this perspective, using LLMs to detect false information seems like a plausible solution (2). However, while LLMs show impressive capabilities, they also have notable limitations, occasionally producing inaccurate, misleading, or fabricated information, also known as hallucinations. Evaluating their performance, especially in critical contexts such as rumor detection, remains a pressing research concern (3). Though powerful, an LLM itself has limited mechanistic interpretability (a black box); researchers are not fully clear about why a model works as intended. As a result, much existing research relies on empirical methods that identify patterns in LLM behavior (4).

LLMs have shown promising performance in fact-checking within specialized areas such as medicine. For example, ChatGPT showed high performance when identifying myths about Alzheimer's disease, and clinicians were satisfied with its explanations (5). A recent study comparing seven LLMs with different numbers of parameters concluded that models with more parameters generally perform better in fake-news detection. The study also noted that fabricated citations generally decrease model accuracy and that LLMs are more prone to misclassification when evaluating content generated by themselves (6). Another study by Saju et al. (7) investigated the performance of different LLMs in judging 61,514 claims across multiple languages. The researchers found substantial differences in accuracy across languages and observed that LLMs are often misled by factual-sounding claims.

The interaction between LLMs and users also presents challenges. DeVerna et al. (8) found that AI-generated fact-checks did not necessarily improve users' ability to discern truth from falsehood and could even reduce trust in accurate information when the model made errors. This raises concerns about integrating LLMs into real-world fact-checking systems.

To summarize, existing literature indicates that LLMs offer promising tools for truth identification but face critical challenges. Further research is needed to develop more robust safeguards and to better understand the interplay between LLM capabilities and human judgment. However, existing studies have primarily focused on comparing performance across different LLMs, while giving less attention to how prompting strategies and text characteristics shape model judgments.

A central difficulty in evaluating truth is that factual correctness and justifiability are conceptually different. In correspondence-based accounts, a statement is true when what it asserts matches the relevant state of affairs; epistemic evaluation instead concerns whether an agent has adequate reasons or evidence for accepting it (9). A proposition may therefore be true without appearing well supported, or appear credible while being false. This distinction is especially relevant to LLMs, which generate judgments from learned patterns rather than direct observation of external reality.

Related distinctions appear in contemporary LLM research. Truthfulness benchmarks assess whether model outputs are factually correct (10), whereas uncertainty research examines whether models appropriately represent the reliability of their answers (11). These targets are related but not interchangeable. However, they are rarely operationalized together for evaluating an LLM's judgments of externally supplied texts. The present study therefore adapts this established conceptual distinction into two experimental measures: Semantic Truth (ST), targeting factual correspondence, and Epistemic Truth (ET), targeting the credibility or justification conveyed by the text.

In response to this gap, this study systematically evaluates the capability of LLMs to classify the truthfulness of texts by comparing model judgments on original factual reports and carefully constructed variants. The findings help illuminate both the strengths and limitations of current LLMs in supporting trustworthy information ecosystems. To further examine why LLMs may overestimate unreliable texts, the study introduces a second phase based on a two-dimensional conception of truth evaluation. Drawing on Tarski's semantic conception of truth (9), it distinguishes semantic truth, which concerns whether a text's claims correspond to external facts or states of affairs, from epistemic credibility, which concerns whether a text appears coherent, supported, and knowledge-like in its presentation.

The significance of this study lies in its attempt to move beyond evaluating whether an LLM can produce correct truthfulness judgments and toward explaining what kind of evidence the model may rely on when making such judgments. Existing research has often focused on comparing the fact-checking performance of different models, datasets, or languages. In contrast, this study focuses on how the structure of the evaluation prompt, and the temporal status of the text influence the model's judgment.

This study makes three main contributions. First, it shows that a single reliability score can conceal semantic–epistemic misalignment, because fabricated texts may receive relatively high overall scores even when the model's explanation expresses doubt. Second, it operationalizes ST and ET as analytically distinct scoring dimensions and evaluates their relationship at the article level without assuming that they are independent. Third, it shows that semantic judgment varies strongly with source and temporal context, a pattern consistent with differences in model familiarity or internal knowledge coverage, and remains

affected by non-factual cues despite targeted prompting. The direct ST–ET representation is used to communicate this two-dimensional structure, while article-level ROC analysis evaluates the discriminative performance of each score.

2. Materials and Methods

2.1 Dataset construction

Articles were collected from reputable sources that are generally considered factually reliable. The initial collection contained a larger number of candidate texts, but unusable outputs, duplicated items, empty or incomplete scraped texts, and other low-quality materials were manually removed during later screening. After this filtering process, the final usable dataset contained 274 source texts: 26 newer BBC articles, 99 newer CNN articles, 100 older CNN articles, and 49 historical articles. Newer CNN and BBC articles were collected from the official CNN and BBC websites, older CNN articles were drawn from CNN's September 2023 archive, and historical texts were collected from history.com.

The BBC and CNN articles mainly consist of contemporary political headlines and reports, and the historical articles are informative texts about generally known historical events. While mainstream media can vary in editorial framing and presentation, these sources are widely accepted as being trustworthy and therefore suitable for constructing a dataset containing factual information.

Articles were obtained in full text. No additional preprocessing was applied; articles retained their original structure, wording, and length. To reflect the nature of the diversity of Internet content, no extra effort was taken to standardize variables such as article length, style, or lexical complexity. Variation across articles was instead treated as part of the natural diversity present in real-world news reporting.

2.2 Construction of text variants

In Phase 1, each usable base article was edited into six additional variations to systematically modify factual content and style. Variant #1 retained factual accuracy but paraphrased the original article. Factual inaccuracies were introduced in Variants #2, #3, and #4, progressing from a single false sentence to approximately 50% false information and finally to an entirely fabricated article. Variants #5 and #6 maintained factual accuracy but altered writing style. More precise details are shown in Table 1.

All variations were created using Gemini-2.5-Flash with controlled prompting and parameters. Variant #0 is the original text. All AI-altered texts were manually checked to ensure that the introduced inaccuracies were indeed false, and that factual elements intended to remain unchanged were preserved. Stylistically manipulated versions were also reviewed for internal coherence.

Table 1. Construction and characteristics of phase 1 text variants

Variation #	Topic	Factual accuracy	Writing style
0	Unchanged	Truthful, original text	Formal, informative
1	Unchanged	Truthful, paraphrase	Formal, informative
2	Unchanged	One sentence is factually false.	Formal, informative
3	Unchanged	Approximately 50% information is factually false	Formal, informative
4	Unchanged	Entirely factually false	Formal, informative
5	Unchanged	Truthful	Informal, entertaining
6	Unchanged	Truthful	Informal, persuasive

2.3 Variables

The study includes several independent variables corresponding to controlled characteristics of each text: factual accuracy, writing style (original/formal vs. informal-entertaining or informal-persuasive), temporal context (articles referencing contemporary vs. historical events), and source category (CNN newer, BBC newer, CNN older, and historical articles).

The dependent variable is the LLM-generated reliability score, ranging from 1 (highly unlikely to be true) to 5 (highly believable). The model's written explanation for each score was also recorded for later qualitative examination.

In Phase 2, the dependent variable was operationally separated into two analytically distinct metrics: Semantic Truth Score and Epistemic Truth Score. The Semantic Truth Score was designed to measure the degree to which the propositional content of a text corresponds to external facts or states of affairs. This definition draws on Tarski's semantic conception of truth (9), in which the truth of a sentence depends on whether what it states is the case. In contrast, the Epistemic Truth Score was designed to measure the degree to which the text appears credible as a knowledge claim based on internal textual features, including logical coherence, evidential presentation, stylistic consistency, and plausibility. The two-score design tested whether the model could respond differently to prompts targeting factual correspondence and prompts targeting perceived credibility.

These two scores were applied to both authentic texts and synthesized fabricated variants across the available source categories. The fabricated variants used in Phase 2 corresponded to Variation #4 in Phase 1, where the articles were designed to be entirely factually false. This made it possible to compare authentic and fabricated texts within each source category while also examining whether the two-score system behaved differently across newer news, older news, and historical information.

2.4 LLM settings and prompting procedure

Texts were evaluated by OpenAI's GPT 4.1 model accessed via OpenAI's API.

In Phase 1, a prompt was designed for the model to generate an analysis of the factual

reliability of the article and an integer between 1 and 5 to quantify its perceived truthfulness. The prompt instructed the model to judge factual reliability from multiple perspectives, including grammar, style, and content. Furthermore, the prompt informed the model that the input texts could refer to events beyond its knowledge cutoff date.

In Phase 2, two structurally parallel prompts were designed to target semantic and epistemic truth separately. For the Semantic Truth Score, the model was instructed to evaluate the factual reliability of the text based on the claims and events described in the article, regardless of whether the writing style appeared professional or coherent. For the Epistemic Truth Score, the model was instructed to evaluate the text's internal credibility, including writing style, logical consistency, evidential presentation, and whether the article appeared like a plausible knowledge claim, without directly judging whether the described events were factually true. This prompt separation was intended to test whether the model could distinguish the two evaluation targets; it was not assumed to guarantee independent underlying judgments.

2.5 Statistical analysis

After data collection, reliability scores were grouped by source category and variant type. Descriptive statistics, including means and standard deviations, were calculated for each group. Because manual screening removed unusable, duplicated, incomplete, or low-quality items, final sample sizes differed across source categories and are reported separately. Comparisons were made across factual accuracy levels, sources, and original/formal versus informal-entertaining and informal-persuasive variants. Qualitative analysis was also conducted to examine whether the model's explanations aligned with its numerical scores. In addition, selected fabricated texts were inspected for textual features, such as negation patterns, headline-body consistency, and plausibility, that could affect how easily fabricated content was detected.

For Phase 2, descriptive statistics were calculated separately for ST and ET across all source categories and truth conditions. To compare authentic and fabricated texts within each source category, paired-samples t-tests and Wilcoxon signed-rank tests were conducted when observations were paired by base article. To compare ST and ET generated from the same texts, paired-samples tests were also used. These analyses examined whether the two-score system exposed distinctions that were obscured by the single reliability score used in Phase 1.

For the article-level analysis, all 548 Phase 2 observations were retained as individually labeled authentic or fabricated texts. A direct ST–ET visualization was used to examine the joint distribution of the two scores. Dashed boundaries at ST = 3 and ET = 3 represent the midpoint of the scoring scales and are descriptive rather than optimized classification thresholds. Receiver operating characteristic (ROC) analysis treated fabricated texts as the positive class. Because lower ST and ET values indicated a greater likelihood of fabrication, both scores were oriented accordingly when calculating the area under the curve (AUC).

For each metric, the threshold maximizing Youden's J statistic was used to report sensitivity, specificity, and accuracy. Because the thresholds were selected and evaluated on the same dataset, they are exploratory and require validation on independent data.

3. Results

Table 2. Phase 1 reliability scores across source categories and text variants

Source	Data	#0	#1	#2	#3	#4	#5	#6
BBC News Newer	Mean	4.15	4.08	4.08	3.88	3.73	3.73	3.54
BBC News Newer	SD	0.97	1.24	1.10	1.16	1.20	1.03	1.27
CNN News Older	Mean	4.88	4.86	4.70	4.66	4.08	3.76	4.70
CNN News Older	SD	0.33	0.35	0.55	0.60	0.88	0.94	0.46
CNN News Newer	Mean	4.17	3.94	4.04	4.02	3.37	3.56	3.75
CNN News Newer	SD	0.95	1.04	1.03	0.98	1.16	0.82	0.90
Historical Articles	Mean	4.90	4.90	4.44	4.44	3.16	4.10	4.88
Historical Articles	SD	0.37	0.31	0.82	0.82	1.29	0.86	0.39

Table 2 shows the average reliability scores (1-5 scale) assigned by the model to different text types and sources in Phase 1. Across all conditions, scores are relatively high, ranging from 3.37 to 4.90. This indicates that the model generally rated most texts as reliable under a one-dimensional truth-evaluation framework.

Scores generally declined as factual inaccuracy increased across Variants #2–#4 relative to the truthful conditions (#0 and #1), although the pattern was not strictly monotonic across all sources. For example, in the CNN News category, scores drop from 4.17 (#0) to 3.37 (#4). Groups #5 and #6, which contain factually accurate texts with informal-entertaining or informal-persuasive writing styles, generally received lower scores than the original factual texts (#0), although their relationship to the entirely fabricated condition (#4) varied by source.

Differences also emerge between sources. CNN News Older and Historical Articles generally received higher scores (4.08-4.90) than the newer BBC and CNN categories (3.37-4.17), although this pattern varied across text variants.

Table 3 presents the descriptive statistics for Phase 2, in which truth evaluation was divided into Semantic Truth Scores and Epistemic Truth Scores. Across all sources, authentic texts received a higher mean Semantic Truth Score than fabricated texts, 4.05 compared with 1.82. Authentic texts also received a higher mean Epistemic Truth Score than fabricated texts, 4.93 compared with 3.94. These results indicate that the two-score framework captured distinctions that were less visible in the original single reliability score.

Table 3. Semantic truth and epistemic truth scores for authentic and fabricated texts by source category

Source / Text Type	Truth Condition	Semantic Mean	Semantic SD	Epistemic Mean	Epistemic SD	n
BBC News Newer	True	2.62	1.86	4.81	0.63	26
BBC News Newer	Fabricated	1.19	0.80	3.27	1.12	26
CNN News Newer	True	3.19	1.91	4.93	0.26	99
CNN News Newer	Fabricated	3.01	1.92	4.87	0.51	99
CNN News Older	True	4.94	0.24	4.98	0.14	100
CNN News Older	Fabricated	1.19	0.77	3.63	1.10	100
Historical Articles	True	4.76	0.72	4.88	0.39	49
Historical Articles	Fabricated	1.02	0.14	3.06	1.13	49
All Sources	True	4.05	1.61	4.93	0.31	274
All Sources	Fabricated	1.82	1.56	3.94	1.18	274

The two-score framework also revealed strong source-level variation. For CNN News Older and Historical Articles, Semantic Truth Scores showed large authentic-fake differences. CNN News Older decreased from 4.94 in authentic texts to 1.19 in fabricated texts, and Historical Articles decreased from 4.76 to 1.02. These large decreases suggest that the model was better able to distinguish fabricated content when the relevant information was older or historical. BBC News Newer also showed a decrease, from 2.62 to 1.19, although the authentic BBC texts already received a relatively low Semantic Truth Score.

The major exception was CNN News Newer. In this category, authentic texts received a mean Semantic Truth Score of 3.19, while fabricated texts received a mean score of 3.01. The difference was only 0.18. The Epistemic Truth Score showed an even smaller difference, decreasing only from 4.93 to 4.87. This suggests that the model had difficulty distinguishing authentic and fabricated newer CNN texts, even when explicitly prompted to evaluate semantic truth.

Table 4 compares authentic and fabricated texts within each source category. The largest semantic differences appeared in CNN News Older and Historical Articles, with differences of 3.75 and 3.73, respectively. BBC News Newer showed a smaller but still noticeable semantic difference of 1.42. In contrast, CNN News Newer showed minimal separation between authentic and fabricated texts, with a semantic difference of only 0.18 and an epistemic difference of only 0.06.

Paired statistical tests (Table 5) support this pattern. The authentic-fabricated difference in Semantic Truth Scores was significant for BBC News Newer, CNN News Older, and Historical Articles. However, the CNN Newer ST comparison was test-dependent: it did not reach

significance by the paired t-test (p=.052) but did by the Wilcoxon signed-rank test (p=.034).

Table 4. Authentic–fabricated differences in semantic and epistemic truth scores by source category

Source category	Semantic M: True	Semantic M: Fabricated	Semantic Difference	Epistemic M: True	Epistemic M: Fabricated	Epistemic Difference
BBC News Newer	2.62	1.19	1.42	4.81	3.27	1.54
CNN News Newer	3.19	3.01	0.18	4.93	4.87	0.06
CNN News Older	4.94	1.19	3.75	4.98	3.63	1.35
Historical Articles	4.76	1.02	3.73	4.88	3.06	1.82
Overall	4.05	1.82	2.24	4.93	3.94	0.99

Table 5. Statistical comparison of authentic and fabricated semantic and epistemic truth scores

Source category	Score Type	Mean Difference	Paired t-test p	Wilcoxon p
BBC News Newer	Semantic	1.42	< .001	.002
BBC News Newer	Epistemic	1.54	< .001	< .001
CNN News Newer	Semantic	0.18	.052	.034
CNN News Newer	Epistemic	0.06	.259	.298
CNN News Older	Semantic	3.75	< .001	< .001
CNN News Older	Epistemic	1.35	< .001	< .001
History Articles	Semantic	3.73	< .001	< .001
History Articles	Epistemic	1.82	< .001	< .001
Overall	Semantic	2.24	< .001	< .001
Overall	Epistemic	0.99	< .001	< .001

Table 6 reports the difference between Epistemic Truth Scores and Semantic Truth Scores. This gap measures the degree to which a text appears more credible than it is factually reliable. Overall, authentic texts had a mean gap of 0.87, while fabricated texts had a much larger mean gap of 2.12. This means that fabricated texts often received much lower factual-correspondence scores than surface-credibility scores. In other words, fabricated texts could still appear credible in style, structure, and presentation even when their factual correspondence was low.

The smallest gaps appeared in authentic CNN News Older and authentic Historical Articles, with gaps of 0.04 and 0.12, respectively. This suggests that when the model was judging information more likely to be covered by its training data, semantic truth and epistemic credibility tended to align. By contrast, fabricated texts and newer news texts produced

much larger gaps. CNN News Newer Fake was especially notable because it retained both a high Epistemic Truth Score of 4.87 and a moderate Semantic Truth Score of 3.01, suggesting that the model treated many fabricated newer CNN texts as plausible and potentially true.

Table 6. Semantic–epistemic score discrepancy (ET−ST) by source category and truth condition

Source category	Truth condition	Epistemic M	Semantic M	Epistemic – Semantic Difference
BBC News Newer	True	4.81	2.62	2.19
BBC News Newer	Fabricated	3.27	1.19	2.08
CNN News Newer	True	4.93	3.19	1.74
CNN News Newer	Fabricated	4.87	3.01	1.86
CNN News Older	True	4.98	4.94	0.04
CNN News Older	Fabricated	3.63	1.19	2.44
Historical Articles	True	4.88	4.76	0.12
Historical Articles	Fabricated	3.06	1.02	2.04
Overall	True	4.93	4.05	0.87
Overall	Fabricated	3.94	1.82	2.12

Table 7 provides the requested article-level evaluation of ST and ET. With fabricated texts treated as the positive class, ST produced the higher AUC (.801). The rule ST < 2 yielded sensitivity of .766, specificity of .810, and accuracy of .788. ET produced an AUC of .760; the rule ET < 5 yielded sensitivity of .569, specificity of .938, and accuracy of .754. Thus, ST was the stronger direct classification signal, while ET captured a different aspect of the model's judgment: perceived credibility. Descriptive source-stratified ST AUCs were .717 for BBC News Newer, .529 for CNN News Newer, .984 for CNN News Older, and 1.000 for Historical Articles. This pattern reinforces the association between semantic truth judgment and source/temporal context.

Table 7. Article-level classification performance of semantic and epistemic truth scores

Metric	Fabricated classification rule	AUC	Sensitivity	Specificity	Accuracy	Youden J
ST	ST < 2	.801	.766	.810	.788	.577
ET	ET < 5	.760	.569	.938	.754	.507

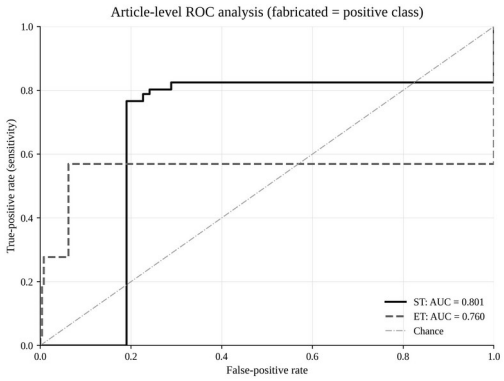

Figure 1. Receiver operating characteristic (ROC) analysis of Semantic Truth (ST) and Epistemic Truth (ET) scores for discrimination between authentic and fabricated texts. Fabricated texts were treated as the positive class, with lower ST and ET scores indicating greater likelihood of fabrication. ST showed greater overall discriminative performance (AUC = 0.801) than ET (AUC = 0.760). The diagonal reference represents chance-level discrimination (AUC = 0.50).

Figure 1 presents the ROC analysis and further clarifies the different roles of ST and ET. Both metrics performed above chance, indicating that each captures meaningful information related to the distinction between authentic and fabricated texts. However, ST achieved a higher AUC and a more balanced sensitivity–specificity trade-off, suggesting that semantic evaluation provides a stronger direct signal for factual discrimination. In contrast, ET showed higher specificity but lower sensitivity, indicating that credibility judgments may remain high for fabricated texts that preserve coherent structure and plausible presentation.

Table 8. Comparison of authentic–fabricated score separation between phase 1 and phase 2

Source category	Phase 1 #0 Mean	Phase 1 #4 Mean	Phase 1 Difference	Phase 2 Semantic Difference	Phase 2 Epistemic Difference
BBC News Newer	4.15	3.73	0.42	1.42	1.54
CNN News Newer	4.17	3.37	0.80	0.18	0.06
CNN News Older	4.88	4.08	0.80	3.75	1.35
Historical Articles	4.90	3.16	1.74	3.73	1.82

Table 8 compares Phase 1's single reliability score with Phase 2's dual-score system. For BBC News Newer, CNN News Older, and Historical Articles, Phase 2 produced larger authentic–fabricated separation than Phase 1. For CNN News Newer, its Phase 2 semantic difference was only 0.18, smaller than the Phase 1 reliability difference of 0.80.

4. Discussion

4.1 Phase 1

The high overall scores suggest that the language model tends to classify most texts as reliable, even when their factual accuracy varies. Although the scale is meant to range from 1 to 5, the average scores given for group #4, which consists of completely factually inaccurate articles, remained above 3. This implies that the LLM tends to overestimate factually unreliable articles, highlighting a potential limitation: the model may not be sufficiently discriminating in its reliability judgments.

Nevertheless, reliability scores generally decreased as factual inaccuracy increased from the truthful conditions (#0 and #1) through the progressively fabricated conditions (#2–#4), indicating some ability to detect false or misleading information. However, the relatively small range of scores implies that its numerical judgments do not scale strongly with accuracy. Moreover, standard deviations were generally larger in the more heavily fabricated conditions than in the original-text condition, although the pattern was not monotonic across all intermediate variants.

The lower scores for groups #5 and #6, relative to the original factual texts, indicate that stylistic presentation can influence reliability judgments even when factual content is preserved. The entertaining and persuasive variants therefore suggest that the model's numerical assessment is sensitive not only to factual accuracy but also to presentation

The difference between older sources and contemporary news points to the influence of the model's training data cutoff. Higher reliability scores for older content likely reflect greater familiarity with that information, while lower scores for recent articles suggest limitations in evaluating newer information. For CNN News Older and historical articles, whose contents existed on the internet before the LLM's knowledge cutoff, the standard deviation of the evaluation scores shows a more explicit increasing pattern from groups #1 to #4, compared with the weaker pattern found in the newer article sources. This further suggests that the LLM seems more confident when evaluating articles that are possibly represented in its training data.

Reviewing the LLM's justifications for its scoring reveals unexpected results. The model's scores were sometimes inconsistent with its explanations, especially for groups #3 and #4, where textual accuracy was low. In some cases, even though the model successfully identified major flaws in the input articles, it still assigned them a high reliability score, sometimes as high as 4 or 5. This observation suggests that the tested LLM may be overly lenient in its final numerical judgments. Such inconsistency may be related to response-tuning tendencies, although this explanation remains speculative.

The following is an example of such cases, where the LLM labeled the article (which belongs to group #4 of the 2025 CNN articles) with a relatively high score of 4. However, it can be seen that the model explicitly doubts the credibility and logical correctness of the article in its self-justification.

1. Likelihood of falsehood or deception: The text is likely to be misleading or partially fabricated.

2. Suspicious elements:
- The article claims a highly sensitive and consequential cyber-military operation ("Desert Echo") was conducted jointly by the US and Israel against Iran, which failed spectacularly and led to a strategic delay by President Trump. There has been no credible public reporting or confirmation of such an operation or failure, which would be a major intelligence and geopolitical event.
- The narrative relies heavily on anonymous sources and "leaked classified assessments," a common trope in unverified or speculative reporting.
- The description of Iranian cyber capabilities being "significantly more advanced than previously assessed" and successfully disrupting US satellite communications and commercial shipping lanes is dramatic and unprecedented, with no corroborating evidence.
- The article implies a covert, high-level conspiracy bypassing diplomatic channels, which is plausible in intelligence contexts but presented here without any supporting evidence beyond anonymous quotes.
- The tone and framing suggest a sensational leak meant to explain a political decision (Trump's delay) as a cover-up for military failure, which fits patterns of disinformation or politically motivated leaks.
- The timing and context (mentioning Trump as president) suggests this is a retrospective or fabricated story, as Trump left office in January 2021, and no such episode was publicly known during his tenure.

3. Believable elements:
- It is plausible that cyber operations are part of US-Israeli coordination regarding Iran.
- The notion of strategic delays to assess intelligence or prepare military options is consistent with real-world decision-making.
- Anonymous sources are common in intelligence reporting, though their presence alone does not guarantee truth.

Reasoning:
The article reads as a sensational exposé with dramatic claims but no verifiable evidence. The lack of any known credible reporting on such a significant event, the reliance on anonymous sources, and the extraordinary claims about Iranian cyber capabilities and their immediate impact on US assets raise serious doubts. The story fits a pattern of speculative or politically motivated leaks intended to cast doubt on leadership decisions and escalate tensions. Without corroboration, the text should be treated as unreliable.

4.2 Phase 2

The updated Phase 2 results provide stronger evidence that semantic truth evaluation varies with source and temporal context, consistent with an influence of model familiarity and internal knowledge coverage. In Phase 1, older CNN articles and historical articles generally received higher reliability scores than newer BBC and CNN articles, although this pattern varied across text variants. This pattern likely reflects the model's greater familiarity with information that existed before or within its training period. Since events occurring before 2024 were more likely to have been included in the model's training data, the model may have had stronger internal access to factual associations related to older news and historical events. In contrast, articles concerning newer events were less likely to be represented in the model's internal knowledge. As a result, the model's evaluation of newer texts relied more

heavily on indirect cues such as plausibility, style, and internal consistency.

The Phase 2 data further support this interpretation. Authentic older CNN articles and historical articles received very high Semantic Truth Scores, and their fabricated counterparts received very low Semantic Truth Scores. This produced large authentic-fabricated semantic differences of 3.75 for CNN News Older and 3.73 for Historical Articles. These results suggest that when the relevant events or facts are likely to be represented in the model's training data, the model can compare the input text against its internal factual knowledge more effectively. In such cases, semantic truth and epistemic credibility also tended to align for authentic texts, as shown by the very small semantic-epistemic gaps in older CNN and historical articles.

However, newer information remained more difficult. BBC News Newer authentic texts received relatively low Semantic Truth Scores, and CNN News Newer showed minimal difference between authentic and fabricated texts. Even under the semantic prompt, fabricated CNN News Newer texts received a mean Semantic Truth Score of 3.01, close to the authentic mean of 3.19. One possible explanation is that the model lacked direct knowledge of post-cutoff events and therefore relied on plausibility and journalistic style. Because many fabricated CNN texts were coherent and realistic, the model treated them as semantically possible.

Qualitative inspection provides an additional explanation for the difference between BBC News Newer Fake and CNN News Newer Fake. When being subjectively reviewed, some BBC fabricated texts appeared to contain more visible signals of fabrication, including explicit and unnatural negation, factual reversal, or illogical tension between the introduction and body text. These features may have made their unreliability easier for the model to recognize as a textual anomaly. In contrast, many CNN News Newer fabricated texts preserved a more natural journalistic structure with plausible institutional references, numerical details, background context, and quotation-style evidence. This suggests that LLM’s semantic truth judgment is, to some extent, also confounded by the style and appearance of text itself.

At the same time, Phase 2 provides a more informative evaluation framework than Phase 1. For older CNN and historical texts, the semantic dimension produced much stronger authentic-fabricated separation than the single reliability score. For CNN News Newer, however, the model failed to separate authentic and fabricated texts even when prompted to focus on semantic truth. This failure is informative because it shows that prompt design alone cannot overcome the absence of external factual access.

4.3 Epistemic credibility and Tarski’s semantic conception of truth

The relative stability of Epistemic Truth Scores is expected, given the design of the experiment. In constructing the dataset, the texts were intentionally kept close to the style of real news

articles. Even the fabricated variants were designed to remain coherent, plausible, and journalistic in presentation. Therefore, most texts preserved surface features normally associated with credible information. This explains why authentic texts received consistently high Epistemic Truth Scores and why many fabricated texts still retained moderate or high epistemic scores. Epistemic credibility, in this sense, reflects how knowledge-like the text appears, not whether its claims are actually true.

The difference between Semantic and Epistemic Truth Scores is therefore central to the interpretation of this study. Overall, fabricated texts had a much larger gap than authentic texts, meaning that they appeared more credible than they were factually reliable. This pattern helps explain the overestimation observed in Phase 1. When the model was asked to assign one overall reliability score, it likely combined semantic cues and epistemic cues into a single judgment. A fabricated article could therefore receive a moderately high score if it were written in a professional style, contained plausible details, and followed a coherent journalistic structure.

This distinction is closely related to Tarski's semantic conception of truth. Tarski connects the truth of a sentence to whether what the sentence states is actually the case. For example, Tarski explains the classical conception through the idea that the sentence "snow is white" is true if and only if snow is white, and he also describes truth as involving correspondence between a sentence and reality or an existing state of affairs (9). Applied to this study, the Semantic Truth Score attempts to approximate this correspondence-based conception of truth, while the Epistemic Truth Score captures a different dimension: whether the text appears credible as a piece of discourse.

This theoretical distinction clarifies the main limitation of using LLMs for truth evaluation without external retrieval. A model can judge whether a text sounds coherent, whether its claims are plausible, and whether its style resembles trustworthy journalism. It may also compare the text with factual patterns that already exist in its training data. However, when the relevant event is newer than its training cutoff, the model does not have direct access to the external state of affairs required for a strict semantic truth judgment. Therefore, its Semantic Truth Score is not an objective verification of truth but a prompted estimate based on internal knowledge.

To further illustrate how the model can be applied to truth evaluation, Figure 2 plots ST directly against ET. Each marker represents one article-level observation, and the dashed lines at the midpoint of each scale create four descriptive regions.

The four regions provide a conceptual vocabulary for interpreting the joint scores. High ST and high ET correspond to grounded information; low ST and high ET to plausible falsehoods; low ST and low ET to obvious falsehoods; and high ST and low ET to poorly supported truths.

The observed distribution supports two conclusions. Authentic older CNN and historical texts cluster mainly in the high-ST/high-ET region, while their fabricated counterparts appear predominantly at lower ST values. In contrast, authentic and fabricated CNN News Newer observations overlap extensively. This source-dependent pattern is consistent with an influence of temporal information familiarity and internal knowledge coverage on semantic judgment. The persistence of high ET values and overlapping ST values for newer, plausible texts also shows that targeted semantic prompting did not fully isolate factual correspondence from plausibility, style, and familiarity.

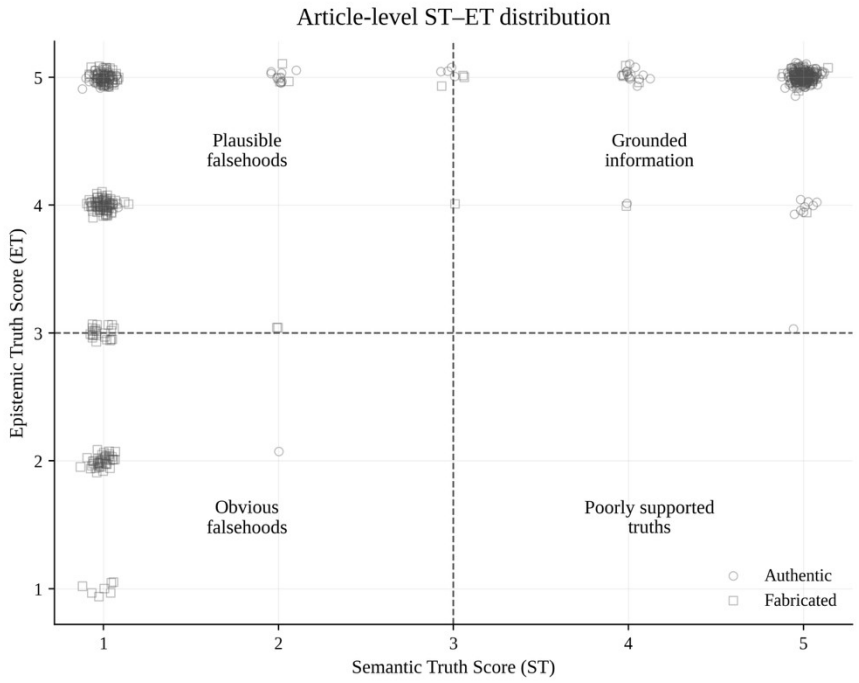

Figure 2. Article-level semantic–epistemic truth distribution. Semantic Truth (ST) represents the model's assessment of factual correspondence, whereas Epistemic Truth (ET) represents perceived credibility based on coherence, plausibility, and evidential presentation. Each point represents an individual authentic or fabricated text. Dashed lines at ST = 3 and ET = 3 denote the descriptive midpoints of the 1–5 scoring scales and divide the plot into four interpretive regions: high ST/high ET (grounded information), low ST/high ET (plausible falsehoods), low ST/low ET (apparent falsehoods), and high ST/low ET (poorly supported truths). These boundaries are descriptive and were not optimized as classification thresholds.

Overall, Phase 2 shows that separating semantic truth from epistemic credibility makes model behavior more interpretable. The clearer performance on older CNN and historical information and the failure on CNN News Newer indicate that semantic judgment remains dependent on internal knowledge coverage and non-factual cues. Without external evidence, ST is a prompted estimate rather than an objective verification of truth.

4.4 Limitations

Several limitations should be acknowledged. First, this study focuses only on written English

texts and one LLM. Images, video, audio, multilingual misinformation, and other models may produce different patterns. Second, sample sizes differed across source categories. Third, the ST–ET quadrant boundaries were fixed at the midpoint of the 1–5 scales for descriptive interpretation and were not empirically optimized.

Another limitation concerns the construction of fabricated texts. Although fabricated variants were designed to be entirely false while preserving a realistic news style, qualitative inspection suggests that the subtlety of fabrication varied across sources. Some degree of inconsistency is present when Gemini-2.5-flash constructs the factually false texts, as demonstrated previously by the slight difference between the factually false versions of "CNN News New" and "BBC News New". Therefore, confounding factors may not be fully eliminated when assessing how effectively GPT-4.1 evaluated the truth scores of each text.

4.5 Perspective

The distinction between semantic truth (ST) and epistemic truth (ET) opens a broader research direction beyond binary classification of authentic and fabricated information. A particularly valuable research question is how long epistemic credibility can remain elevated as semantic correspondence is progressively degraded through increasing fabrication. Rather than comparing only completely authentic and completely fabricated texts, future studies could systematically introduce increasing degrees of factual distortion while preserving the style, coherence, and plausibility of the original text. Such experiments would identify a *truth-stretching threshold*: the degree of semantic deviation that can be introduced before the model's judgment changes substantially. This would transform misinformation detection from a binary true–false problem into the study of a continuous transition from factual information to plausible falsehood.

The ST–ET framework provides a natural basis for examining this transition. As factual content is progressively altered, ST should ideally decline in proportion to the loss of factual correspondence. ET, however, may remain relatively stable if the altered text continues to appear coherent, authoritative, and well supported. The divergence between these trajectories could therefore reveal a form of truth inflation, in which the apparent credibility of a text remains disproportionately high as its factual grounding deteriorates. The important question would no longer be simply whether an LLM detects a false article, but how much factual distortion the model tolerates before credibility and semantic grounding become measurably disconnected.

Testing this hypothesis will require more controlled manipulation of fabrication subtlety than was possible in the present study. Future datasets could introduce factual changes incrementally—for example, altering individual quantities, dates, entities, causal relationships, quotations, and eventually entire events—while holding writing style and structure as constant as possible. Such designs could determine whether truth degradation produces a gradual response or whether LLM

judgments exhibit thresholds at which ST changes abruptly. They could also establish whether different forms of distortion have different truth-stretching limits; changing a numerical value, for example, may be treated differently from inventing an event that remains contextually plausible.

Relative measures such as ET/ST could be investigated within this framework as candidate measures of truth inflation. The value of such a ratio should not be assumed, however, but tested against the direct ST and ET scores and alternative discrepancy measures. Article-level ROC analysis, sensitivity, specificity, precision, F1-score, calibration analysis, and independent validation could establish whether ET/ST identifies progressively distorted information more effectively than ST alone or the absolute ET−ST difference. This would allow the data, rather than the definition of the metric, to determine which representation best captures semantic–epistemic discordance.

A further opportunity is to determine what controls the truth-stretching threshold. Hierarchical or mixed-effects models could quantify the contributions of source category, temporal distance from the model's training cutoff, fabrication magnitude, textual plausibility, writing style, and type of factual alteration. The present finding that newer CNN texts were considerably more difficult to discriminate than older or historical material suggests that the tolerance for distortion may itself depend on the model's familiarity with the underlying information. The same experiment should therefore be repeated across multiple LLMs, model generations, subject domains, languages, and levels of training-data familiarity, with human judgments and retrieval-based factual verification serving as external reference points.

Ultimately, this research direction could produce a more informative benchmark than asking whether an LLM can simply distinguish truth from falsehood. Quantifying how epistemic credibility changes as semantic correspondence is progressively degraded could establish a threshold at which plausible presentation can no longer sustain perceived credibility. Such a benchmark could be particularly relevant to real-world misinformation, where misleading information often does not replace truth completely but instead preserves a credible factual scaffold while progressively altering selected details. In this setting, truth inflation and the truth-stretching threshold may provide useful ways to characterize not only whether an LLM fails, but how much distortion it will tolerate before it fails.

5. Conclusion

This study distinguishes two components of LLM truth judgment that are often conflated in a single reliability assessment. Semantic Truth (ST) refers to whether the factual claims in a text correspond to the relevant external state of affairs, whereas Epistemic Truth (ET) refers to whether the text appears credible or justified on the basis of features such as coherence, plausibility, evidential presentation, and consistency. Operationally separating these dimensions provides a way to examine not simply whether an LLM labels information as

true or false, but what aspects of a text may contribute to that judgment.

Phase 1 demonstrated why this distinction is necessary. When asked to provide a single reliability score, the LLM generally rated texts highly even as their factual content was progressively altered, and completely fabricated articles still received relatively high scores. In some cases, the model's written explanation correctly identified substantial evidence of fabrication while its numerical reliability score remained high. Thus, a single score can obscure disagreement between the model's recognition of factual problems and the apparent credibility or plausibility of the text.

Phase 2 exposed this structure by separately eliciting ST and ET. Authentic and fabricated texts were much more strongly separated by ST than by ET overall, indicating that targeted semantic prompting can improve factual discrimination. However, the separation was strongly dependent on the information being evaluated. Older CNN and historical texts showed large authentic–fabricated differences, whereas newer CNN texts showed extensive overlap. Consequently, the experiment reveals an important limitation of semantic prompting: instructing an LLM to judge factual correspondence does not make its judgment independent of its existing knowledge. ST remained influenced by training-data familiarity and by non-factual characteristics such as plausibility, journalistic style, and textual structure.

The article-level classification analysis quantifies this limitation. ST provided the stronger direct discriminator between authentic and fabricated texts, with an overall AUC of 0.801, sensitivity of 0.766, and specificity of 0.810, compared with an AUC of 0.760 for ET. More importantly, ST performance varied markedly by source, from near-chance discrimination for newer CNN articles to nearly perfect discrimination for older and historical material. The apparent ability of an LLM to recognize false information therefore cannot be interpreted as a general truth-detection capability; it is conditional on the relationship between the evaluated information, the model's internal knowledge coverage, and the textual cues available to it.

The principal contribution of this study is therefore not simply that ST performs better than ET, but that LLM truth judgment can be represented as a semantic–epistemic problem rather than a single reliability judgment. A text may appear coherent and credible while lacking factual grounding, and an LLM may recognize these two properties differently. At the same time, the observed overlap between ST and ET demonstrates that targeted prompts do not create completely independent semantic and epistemic judgments. The two-dimensional ST–ET representation makes this overlap visible and provides a framework for investigating when apparent credibility diverges from factual grounding.

These findings also identify a direction for further development. Relative measures such as ET/ST, as well as other normalized measures of semantic–epistemic discordance, could be tested against the direct ST and ET scores to determine whether they provide additional

discriminatory or calibration value. Such measures require independent validation rather than being assumed to improve classification. Ultimately, LLM-based truth assessment should be understood as an estimate conditioned by model knowledge and textual evidence rather than as direct verification of reality. External retrieval, source verification, or human evaluation remains necessary when factual grounding cannot be established from the model's internal knowledge.

6. References

1. Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3), 554–559. https://doi.org/10.1073/pnas.1517441113

2. Das, A., Liu, H., Kovatchev, V., & Lease, M. (2023). The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2), 103219. https://doi.org/10.1016/j.ipm.2022.103219

3. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55. https://doi.org/10.1145/3703155

4. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2), 1–38. https://doi.org/10.1145/3639372

5. Huang, S. S., Song, Q., Beiting, K. J., Duggan, M. C., Hines, K., Murff, H., Leung, V., Powers, J., Harvey, T. S., Malin, B., & Yin, Z. (2024). Fact check: Assessing the response of ChatGPT to Alzheimer's disease myths. *Journal of the American Medical Directors Association*, 25(10), 105178. https://doi.org/10.1016/j.jamda.2024.105178

6. Sallami, D., Chang, Y.-C., & Aïmeur, E. (2024). From deception to detection: The dual roles of large language models in fake news. *arXiv*. https://doi.org/10.48550/arXiv.2409.17416

7. Saju, L., Bleier, A., Lasser, J., & Wagner, C. (2025). Facts are harder than opinions—A multilingual, comparative analysis of LLM-based fact-checking reliability. *arXiv*. https://doi.org/10.48550/arXiv.2506.03655

8. DeVerna, M. R., Yan, H. Y., Yang, K.-C., & Menczer, F. (2023). Artificial intelligence is ineffective and potentially harmful for fact checking. *arXiv*. https://doi.org/10.48550/arXiv.2308.10800

9. Tarski, A. (1944). The semantic conception of truth and the foundations of semantics. *Philosophy and Phenomenological Research*, 4(3), 341–376. https://doi.org/10.2307/2102968

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

11. Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 625–630. https://doi.org/10.1038/s41586-024-07421-0