From 'May' to 'Is': Certainty Distortion in Language Model Rewriting

Catarina G Belem[1], Shang Wu[1], Hongyu Yao[3], Mark Steyvers[1,2], Sameer Singh[1], Padhraic Smyth[1]

[1]Department of Computer Science, University of California Irvine
[2]Department of Cognitive Science, University of California Irvine
[3]Massachusetts Institute of Technology

Correspondence: cbelem@uci.edu

Abstract

Humans increasingly turn to Language Models (LMs) in ways that shape beliefs and drive decisions, including discussing, rewriting, and summarizing information from scientific articles, news, and medical reports. However, in these domains, where it often matters how confidently a claim is expressed, little is known about whether LMs faithfully preserve the degree of confidence. In this work, we investigate *certainty distortion* in LMs, defined as meaningful changes in expressed certainty during transformations intended to preserve meaning. We propose an LM-based evaluation metric that is consistent with population-level judgments of certainty. Using this metric, we characterize certainty distortion across different sizes and families of models in the context of scientific and medical communication tasks. Our results show that certainty distortion affects up to 75% of LM outputs and is systematically asymmetric in rewriting tasks with most LMs being 1.5-2× more likely to increase the expressed certainty than to decrease it. These effects can compound over repeated paraphrasing: in the medical domain, `Claude 4.5 (H)` increases certainty in 20% of examples after a single iteration, increasing to 40% after five iterations. Prompt-based interventions reduce overall certainty distortion but do not eliminate it. Together, these findings reveal a general bias toward inflating expressed certainty, with direct implications for users who rely on LMs in high-stakes domains.

1 Introduction

The use of language models (LMs) for summarizing, simplifying, and rewriting complex information is rapidly growing. In domains like medicine, law, and science, claims are carefully worded to reflect the degree of certainty supported by the evidence (Vincze et al., 2008; Lee and Whitehead, 2017; Shinagare et al., 2019; Fore, 2019). Authors use epistemic markers, such as hedges (*e.g.*, *may*, *could*), non-factive verbs (*e.g.*, *suggest*, *appear*), and source attribution (*e.g.*, *according to*) to signal what is established versus what is tentative (Fischhoff and Davis, 2014; Dhami and Mandel, 2022). When these markers are modified, the epistemic status of a claim shifts, and readers may form beliefs or make decisions that go beyond what the evidence supports (Wu et al., 2025; Guingrich et al., 2026). Prior work suggests that around 30% of human-written scientific findings are inflated when communicated across media outlets (Brechman et al., 2009; Sumner et al., 2016; Wuehrl et al., 2024), indicating that certainty distortion is a real phenomenon in human-mediated communication. However, whether LMs exhibit similar behavior in rewriting tasks, and to what extent this may be a systematic aspect of LM behavior, remains an open question. Consider the motivating example in Figure 1 concerning a radiology report from MIMIC-CXR (Johnson et al., 2019): the report describes the increased empty space behind the sternum as

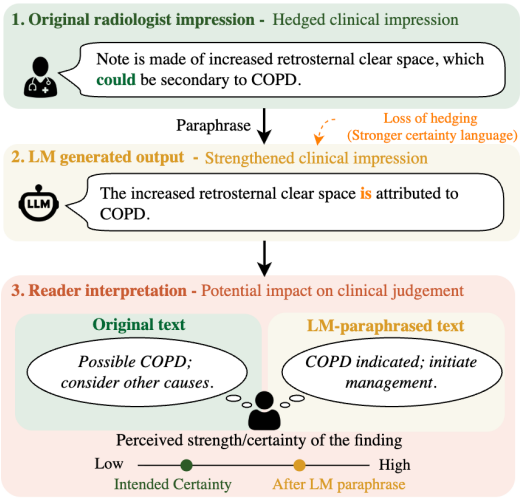


Figure 1: A real example of certainty inflation: `Qwen3 (8B)` paraphrases a hedged radiology impression into a confident attribution, shifting the reader's belief from a tentative diagnosis to a definite clinical conclusion.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 21426–21464
October 24-29, 2026 ©2026 Association for Computational Linguistics

possibly explained by a chronic obstructive pulmonary disease (COPD), among other diagnoses. The LM paraphrase drops the hedge, attributing the empty space to COPD. Such a shift, if relied on clinically, could prematurely push the clinician toward treating COPD.

In this work, we investigate *certainty distortion* in LMs.[1] We define certainty distortion as the likelihood that the degree of certainty expressed in source text is not preserved after an LM rephrases it. We make three contributions. First, we propose an LM-based evaluation metric for measuring certainty distortion, which we validate via a user study with multiple human annotations. We find that the proposed LM evaluation metric aligns with population-level consensus judgments as well as any individual human annotator does. Second, we use this metric to systematically characterize certainty distortion across model sizes and families for scientific communication and medical communication tasks. We find that certainty distortion affects up to 75% of LM outputs and that it is asymmetric, with most models being 1.5-2× more likely to amplify rather than attenuate expressed confidence. When evaluating across repeated LM rewriting, we observe a contrast: in scientific communication tasks, most certainty shifts occur after the first LM output, whereas in medical communication tasks certainty shifts compound across iterations. Specifically, in the medical communication task, we observe that models' tendency to increase expressed certainty grows by an additional 13-20% points by the fifth paraphrase. Third, we evaluate the robustness of certainty distortion to decoding strategy and test simple prompt-based mitigation approaches. We find that certainty distortion remains largely unchanged under greedy decoding, suggesting that the phenomenon is not primarily driven by stochastic sampling, while prompt-based interventions can reduce distortion but do not eliminate it.

Together, these findings show that the evaluated LMs systematically distort expressed certainty when rewriting text. Most exhibit a recurring tendency toward certainty inflation, which can compound across repeated medical paraphrases in particular. Combined with growing evidence that more confident model outputs drive stronger belief updates (Wu et al., 2025; Guingrich et al., 2026) and overreliance in users (Kim et al., 2024; Steyvers et al., 2025), this points to a real risk in everyday LM use, one that current mitigation strategies fail to address. We argue that *epistemic faithfulness*, *i.e.*, preserving not just what is said, but the degree of certainty with which it is said, deserves attention alongside factuality.

2 Related Work

Certainty Distortion in Human Communication. Epistemic markers (*e.g.*, hedges, plausibility shields and non-factive verbs, evidential markers) encode a claim's evidential status and the speaker's commitment to it, distinguishing, for example, *she might have committed fraud* from *she definitely committed fraud* (Zhou et al., 2023, 2024). Social science research further shows that epistemic markers (*e.g.*, *likely*, *doubtful*) correspond to measurable differences in inferred certainty (Budescu et al., 1988; Wallsten et al., 2008). Studies of scientific dissemination through press releases, news, and social media find that such epistemic markers are not faithfully preserved (Brechman et al., 2009; Sumner et al., 2016; Wuehrl et al., 2024). However, these studies focus on communication across human intermediaries and do not establish whether LMs exhibit analogous patterns.

Information Distortion in LM Transformations. Recent work raises concerns that LMs can distort human-generated language, including by altering textual meaning (Abdulhai et al., 2026). Prior work on LM transformations primarily evaluates whether source semantic meaning and factual claims are preserved (Maynez et al., 2020; Pagnoni et al., 2021; Wang et al., 2025; Ramprasad and Wallace, 2025). Closer to our setting, recent work adopts the *broken telephone* paradigm to study cumulative distortion: Perez et al. (2025) measures drift in surface attributes, such as toxicity, sentiment, and length under iterative paraphrasing and open-ended generation, while Mohamed et al. (2025) examines factual and semantic drift in translation chains. Both examine drift along axes distinct from epistemic certainty, focusing on factual or semantic preservation rather than shifts in expressed certainty during rewriting intended to preserve meaning. We treat expressed certainty as a separate axis from factuality (whether a claim is true) and from calibration (whether a model's predicted probabilities or confidence estimates are statistically aligned with the empirical frequency of correctness) (Kadavath et al., 2022) and further show that this drift is di-

[1] All data and code are publicly available at https://github.com/UCIDataLab/certainty-distortion-in-llm-rewriting.

rectional, with a systematic bias toward increasing certainty.

Systematic Biases in LM Outputs. Prior work documents biases such as sycophancy, where models align responses with user beliefs rather than evidence (Perez et al., 2023; Sharma et al., 2024), and brittleness to epistemic markers, which can sway LM discriminative judgments of accuracy and misinformation (Lee et al., 2025; van de Sande et al., 2025). We extend this literature to generative tasks that require preserving source certainty.

3 Measuring Certainty Differences

In this paper, we focus on information-preserving tasks like rephrasing and simplification. In these tasks, an LM is given a source text x_o and is asked to generate an output x_M that preserves all or part of the information in the source text. We are particularly interested in whether the *level of certainty* in the input text remains the same in the transformed output text generated by the model. Here "level of certainty" is associated with semantic propositions being expressed in the text. For example, the text "From our experiments, there is evidence to suggest that X may cause Y" contains some uncertainty about the relationship between X and Y, whereas the text "From our experiments we conclude that X causes Y" contains no uncertainty, despite preserving the same semantic meaning. To characterize such *certainty differences*, we consider text pairs: the original text x_o and its model-generated version x_M. Assuming a function $\mathrm{c}(x) \in [0, 1]$ that reliably estimates the expressed certainty of a text, we define *per-example certainty difference* as

$$\delta(x_o, x_M) = \mathrm{c}(x_M) - \mathrm{c}(x_o), \quad (1)$$

where $\delta = 0$ indicates preservation of certainty (no distortion), $\delta > 0$ indicates increased certainty (*overconfidence*), and $\delta < 0$ indicates decreased certainty (*underconfidence*).

To assess whether certainty distortion is a systematic property of model behavior rather than an artifact of individual examples, we define **Certainty Distortion (CD)** as the fraction of examples in a dataset $\mathcal{D}$ in which a shift occurs:

$$\mathrm{CD}(\mathcal{D}) = \mathbb{E}_{(x_o,x_M)\sim\mathcal{D}}[\mathbb{1}_{[\delta(x_o,x_M)\neq 0]}]. \quad (2)$$

Additionally, we can further decompose CD into upward (CD-↑) and downward (CD-↓) shifts to capture the distortion direction, where we track $\mathbb{1}_{[\delta(x_o,x_M)>0]}$ and $\mathbb{1}_{[\delta(x_o,x_M)<0]}$, respectively.

3.1 Operationalizing Certainty Estimates

To operationalize CD, we require a reliable estimator $\delta(x_o, x_M)$. One approach is to estimate certainty $c(x)$ independently for each text x and then compute the difference. However, this *reference-less* approach could fail to capture subtle directional shifts in certainty, which could render it unreliable for estimating CD. Instead, we treat $c(x)$ as a latent quantity and estimate δ directly through pairwise comparison.

Human annotations of pairwise certainty differences would be ideal, but they are inherently noisy and subjective for this type of task (Pei and Jurgens, 2021; Wuehrl et al., 2024). As a result, a large number of human annotations would be necessary to obtain reliable estimates, which is not scalable.

As an alternative, prior work has explored automated proxies based on fine-tuned BERT models (Pei and Jurgens, 2021; Wuehrl et al., 2024). However, because these models operate in a reference-less manner, we hypothesize that they have limited sensitivity to subtle pairwise differences in certainty. We validate this hypothesis empirically by constructing controlled datasets in which certainty shifts are systematically manipulated (Appendix A.2). We then evaluate whether these metrics align with the intended shift across subsets, finding that they fail to reliably detect these shifts. In fact, we observe a saturation effect where the BERT-based metrics assign uniformly high scores, leaving little headroom to distinguish moderately certain from highly certain texts. We further corroborate these findings by analyzing cases where the models disagree (Tables 6 and 7).

With these limitations in mind, in this work we pursue a pairwise evaluation approach that can directly compare two texts without relying on absolute scores. Specifically, we propose to estimate CD directly through an LM-as-a-judge approach (Liu et al., 2023). We ask `gpt-5.4-mini` to determine which of two texts communicates the main finding with greater certainty. To avoid biases associated with the use of labels such as *Original* and *Modified*, we present the two texts x_o and x_M blindly as *Text A* and *Text B*, and prompt the model to select from a 5-point scale ranging from *Clearly A* to *Clearly B*, with *No Clear Difference* as the midpoint (full prompt in Figure 24).[2] To encour-

[2]The scale is ordinal, so its labels represent ordered differences rather than calibrated thresholds. We validate it for source–transformation pairs expressing the same underlying claim (Appendix A). Other uses would require separate vali-

age the model to explicitly formalize a certainty estimate for each text (*i.e.*, $c(x_o)$, $c(x_M)$) prior to making a comparative judgment, we elicit chain-of-thought (CoT) reasoning (Wei et al., 2022) over the certainty sources present in each text.[3] To further marginalize over positional bias (Zhao et al., 2021; Stureborg et al., 2024; Pezeshkpour and Hruschka, 2024), we query the model twice reversing the input order, remapping each prediction into a canonical form–*Clearly Original*, *Slightly Original*, *No Clear Difference*, *Slightly Modified*, or *Clearly Modified*–accounting for the input order of each call. When both calls produce the same canonical label, that label is adopted as the final prediction; when they agree in direction but differ in magnitude, we select the label closer to *No Clear Difference*. When they disagree in direction, we flag the pair as order-inconsistent and treat it as a tie when computing directional-shift rates. A moderator-based analysis indicates that most order-inconsistent pairs are best characterized as ties (Appendix D.1). For analysis, we map the canonical labels to a numerical scale $\{-2, -1, 0, 1, 2\}$, where negative values indicate greater certainty in the original text x_o, positive values indicate greater certainty in the LM-generated text x_M, and 0 indicates no perceived difference.

3.2 Human Validation

To verify that `gpt-5.4-mini` serves as a reliable proxy for human certainty judgments, we conduct a human annotation study comparing LM-as-a-judge predictions against human pairwise assessments of certainty. Specifically, we select 240 text pairs from the paraphrasing and news tasks in Section 4. We recruited 138 annotators from Prolific.[4] After applying pre-specified quality filters, we excluded 9 annotators (3 for excessive off-screen time, 1 for fast uniform responses, and 5 for failing the attention check) leading to 129 total annotators.[5]

As in the LM-as-a-judge setup, the two texts x_o and x_M are presented anonymously as *Text A* and *Text B* in randomized order. Annotators judge which text communicates greater certainty on a 5-point Likert scale, subsequently remapped to the same numerical scale $\{-2, -1, 0, 1, 2\}$ introduced in Section 3.1, facilitating direct comparison between human and LM-as-a-judge predictions.

We measure inter-annotator agreement using Krippendorff's α, finding $\alpha \approx 0.25$ across annotator groups. The observed agreement is consistent with prior work on certainty annotation (Rubin, 2007; Wuehrl et al., 2024), reflecting the inherent subjectivity of certainty judgments.

Results. To assess whether `gpt-5.4-mini` can serve as a reliable proxy for human judgment, we evaluate it in terms of its correlation with population-level perceptions of certainty. Specifically, we compute Kendall τ_B correlations between each annotator (and `gpt-5.4-mini`) and the *human consensus*, defined via leave-one-out as the mean prediction of the remaining annotators. If `gpt-5.4-mini` correlates more strongly with the consensus than the average annotator, it is a more reliable proxy for the central tendency of human certainty perception than any single annotator. As a complementary metric, we also report in Appendix B, the average consistency between `gpt-5.4-mini` and *individual* human annotators, defined as the pairwise rank correlations between all `Human-LM` and `Human-Human` annotator pairs. If `gpt-5.4-mini` agrees with the individual humans at least as well as two humans agree with each other, it can be considered a reliable substitute for a randomly sampled annotator. As reference points, we include a random baseline and the 3-class classifier from Wuehrl et al. (2024).

Figure 2 shows that `gpt-5.4-mini` correlates with the human consensus more strongly than the average individual annotator ($\tau_B = 0.47_{\sigma=0.09}$ vs $\tau_B = 0.34_{\sigma=0.19}$), indicating that it captures the central tendency of human certainty perception at least as well as a randomly sampled annotator. At the individual-annotator level, `gpt-5.4-mini` also achieves consistency comparable to Human-Human consistency (Appendix B), confirming its reliability as a substitute annotator.

Compared with the prior BERT-based estimator, `gpt-5.4-mini` correlates more strongly with the human consensus. Across 10,000 paired bootstrap samples, the mean correlation gap was $+0.16$ in favor of `gpt-5.4-mini` (95% CI $[0.07, 0.25]$). In addition, `gpt-5.4-mini` provides a textual rationale alongside each judgment, facilitating qualitative inspection and debugging of how certainty differences are identified, and does not require supervised training on a task-specific data distribution. We therefore use `gpt-5.4-mini` as our certainty

dation.

[3]The improved alignment of using CoT with certainty shifts is validated in a controlled setting in Appendix A.2.

[4]https://www.prolific.com/

[5]Compliant with IRB protocols; see Appendix B.

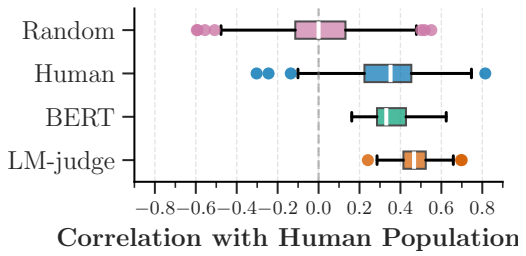


Figure 2: Correlation of different CD estimators with the mean prediction of the human population, measured using Kendall τ_B correlations. Overall, the LM-as-a-judge correlates better with the human population than the average individual human or the strongest BERT-based estimator (Wuehrl et al., 2024).

estimator throughout the remainder of the paper.

4 Experiments

We now investigate the degrees to which LMs preserve or distort the epistemic certainty of original content. To characterize this behavior, we report the fraction of examples in which a model-generated output differs in certainty from the source (CD), and further decompose this distortion into upward (CD-↑) and downward (CD-↓) shifts to capture the distortion direction.

4.1 Experiment Setup

Datasets. We begin with sentence-level tasks as our first testbed. At this granularity, models are less likely to introduce hallucinated or unrelated content (Zhang et al., 2024; Zhao et al., 2025; Yang et al., 2026), and, as a result, CD is more likely to faithfully capture changes in epistemic certainty rather than differences in content. We use two datasets: SPICED (Wright et al., 2022) and MIMIC-CXR (Johnson et al., 2019). SPICED is a dataset of scientific findings paired with corresponding news and social media versions, spanning various disciplines (*e.g.*, biology, computer science, and psychology). We focus specifically on the paired *News* subset and filter out examples in which the original finding spans multiple sentences, resulting in 397 sentence pairs. Additionally, we use the Findings section of the radiology reports available in MIMIC-CXR. We restrict the analysis to a subset of 800 sentences across reports, 400 of which are selected as containing at least one epistemic marker (*e.g.*, *suggests*, *compatible with*). Additional details are available in Appendix E.

Tasks. We evaluate LMs on two tasks within each domain. In the scientific domain, we instruct models either to paraphrase a scientific finding or to rewrite it as if reporting it in The New York Times. In the medical domain, we similarly instruct models either to paraphrase a sentence from the findings section of a radiology report or to rewrite it in simplified language for a non-expert patient. All generation prompts are provided in Appendix C. In this initial study, we intentionally keep prompts simple to better reflect how everyday users are likely to interact with these systems.

Models. We evaluate five instruction-tuned LMs from different model families that are widely used in practice: `Gemini 3.1 (fl)` (Google DeepMind, 2026), `gpt-5.4-nano` (OpenAI, 2026), and `Claude 4.5 (H)` (Anthropic, 2024) as representative closed-source models, and `Llama 3.3 (70B)` (Meta AI, 2024) and `Qwen3 (8B)` (Qwen Team, 2025) as representative open-source models. Except when explicitly stated otherwise, all experiments use top-p sampling (`T=1`, `top_p=0.9`), reflecting how LMs are commonly used in practice (Holtzman et al., 2020; Li et al., 2023; Scholten et al., 2025). Across all experiments, models are given a generation budget of 3.5k tokens. Finally, we evaluate the CD metric for each pair of original and generated texts, using `gpt-5.4-mini` as the LM-as-a-judge, as described in Section 3.

4.2 Certainty Distortion Results

LMs exhibit non-negligible certainty distortion rates across tasks and domains. Table 1 summarizes the CD metric across tasks. Average CD rates range from 30% for `gpt-5.4-nano` to 49.4% for `Llama 3.3 (70B)`. Even in sentence-level paraphrasing of scientific findings, a task with minimal incentive for models to alter content or length, *37.5% to 64% of LM-generated outputs distort the source text's expressed certainty*.

News rewriting amplifies certainty distortion more than paraphrasing. In the scientific domain, news rewriting pushes distortion up to 74.8% (`Claude 4.5 (H)`) and 75.3% (`Gemini 3.1 (fl)`), increases of 26.7 and 15.1 points respectively over paraphrasing for the same models. The difference is smaller for `gpt-5.4-nano` and `Llama 3.3 (70B)`, with only a 5.3 to 6.8-point increase. Manual inspection (see Table 20) shows that `Claude 4.5 (H)` and `Gemini 3.1 (fl)` adopt the evidential framing typical of science journalism (*e.g.*, "Researchers have", "Scientists have") and stronger verbs (*e.g.*, "discovered"), both of which

	SPICED		MIMIC		
Method	**Para**	**News**	**Para**	**Simpl**	**Avg**
`gpt-5.4-nano`	37.5	44.3	25.8	12.5	30.0
`Claude 4.5 (H)`	48.1	74.8	29.1	20.1	43.0
`Gemini 3.1 (fl)`	60.2	75.3	24.3	26.8	46.7
`Qwen3 (8B)`	43.1	57.7	31.1	18.2	37.5
`Llama 3.3 (70B)`	64.0	69.3	38.3	26.0	49.4

Table 1: Model certainty distortion (in %) on paraphrasing and rewriting of single-sentence scientific and radiological findings. Lower is better. All 95% CIs ≤ 0.1.

	SPICED			MIMIC		
Model	**CD-↑**	**CD-↓**	**R**	**CD-↑**	**CD-↓**	**R**
`gpt-5.4-nano`	19.0	21.9	0.9	14.4	4.8	3.0
`Claude 4.5 (H)`	37.5	24.2	1.5	16.0	8.6	1.9
`Gemini 3.1 (fl)`	42.9	24.8	1.7	15.5	10.1	1.5
`Qwen3 (8B)`	33.4	17.0	2.0	16.1	8.6	1.9
`Llama 3.3 (70B)`	33.5	33.1	1.0	12.3	19.9	0.6

Table 2: Directional certainty distortion rates (%) and inflation-to-deflation ratio, averaged across tasks within each domain. Ratios R ≥ 1 indicate models inflate certainty more often than they deflate it.

strip hedges and present results as established.[6] Conversely, `gpt-5.4-nano` mostly preserves the source's voice ("We find", "Our results show") and the same verbs, leaving less room for confidence shifts to occur. Since omitting hedges may also shorten outputs, these qualitative patterns raise the possibility that the observed inflation partly reflects compression. We investigate this possibility in Appendix F.2, where regression controls and a matched-length analysis show that differences in length do not fully explain the effect.

Certainty distortion is lower in medical communication than in scientific communication. While distortion rates range between 37.5%-75% in scientific communication, the distortion rates in the medical domain are much lower, 12.5%-38.3%. One possible explanation lies in how each type of content relates to the model's prior knowledge. Scientific findings discuss general claims about the world that LMs may have encountered during pretraining, which may lead them to pattern-match to familiar phrasings rather than preserve the source's hedging. Radiology findings, by contrast, are observations about a specific patient about whom the model has no prior information, potentially leaving less room for the model to substitute more certain formulations learned from similar texts during pretraining.

Most models inflate certainty more than they deflate it. Table 2 decomposes certainty distortion into upward (CD-↑) and downward (CD-↓) shifts. Across both domains, most models inflate the certainty of the source more often than they deflate it, with inflation-to-deflation ratios ranging from 1.5 to 2.0. We note two exceptions: `gpt-5.4-nano` and `Llama 3.3 (70B)` are roughly balanced across SPICED tasks (0.9 and 1.0, respectively), and `Llama 3.3 (70B)` further stands out on MIMIC, where it is the only model to significantly deflate more than it inflates (0.6).

Figure 3 breaks the analysis down by the expressed certainty of the original text. Although certainty distortion varies with source certainty, we find **directional biases** in four of five models that cannot be explained by simple floor or ceiling effects alone. CD-↑ should decrease, and CD-↓ should increase, as source certainty rises: since texts that are already highly certain have little room to become more certain, and vice versa. However, because the *low* and *high* subsets are equally sized, these mechanical constraints alone do not account for the observed imbalance in directional changes. In SPICED, all models except `gpt-5.4-nano` increase expressed certainty for **50–75% of low-certainty source texts**, whereas they decrease expressed certainty for **20–50% of high-certainty source texts**. A similar asymmetry holds in MIMIC, where models increase the certainty of uncertain source texts in 25–30% of examples but decrease the certainty of certain source texts in fewer than 16%. In both datasets, `Llama 3.3 (70B)` decreases certainty in certain texts at least as often as it increases certainty in uncertain texts. Surprisingly, for already-confident texts, `Claude 4.5 (H)` and `Gemini 3.1 (fl)` still inflate certainty in a non-trivial fraction (20%) of SPICED examples. These findings suggest that for some LMs, the bias towards "overconfident" outputs is not only associated with hedged inputs but reflects a more general tendency to strengthen the claims

[6] The use of journalistic framing suggests that genre conventions may partly contribute to the higher distortion observed in this task. As a diagnostic magnitude check, we apply the same evaluator to the corresponding human-written SPICED news texts, which exhibit 23.9% CD, compared with 44.3–75.3% for the LM rewrites. This is not a controlled human–LM comparison, as the human texts were naturally occurring and may reflect additional context, editorial goals, and other genre conventions. Nevertheless, the comparison suggests that LM distortion is not explained solely by reproducing the average distortion observed in human-written science news (Appendix F.3). Moreover, genre conventions cannot explain the distortion observed in non-news tasks, including scientific and medical paraphrasing.

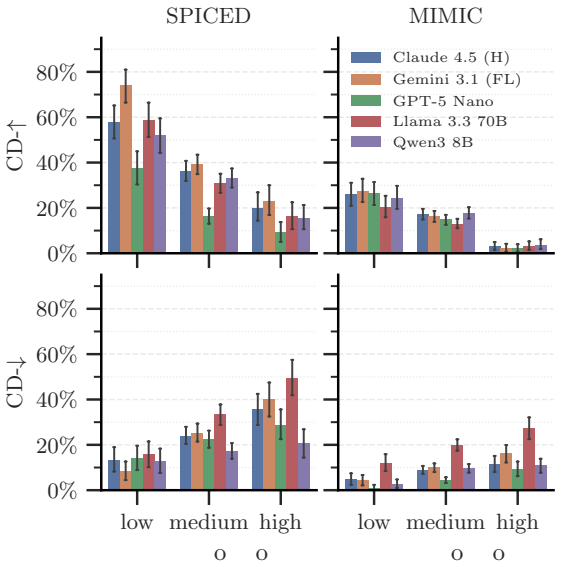


Figure 3: Direction of certainty distortion (CD-↑ top, CD-↓ bottom) as a function of the source's certainty level (low/medium/high) across SPICED (left) and MIMIC (right). As expected from floor/ceiling effects, CD-↑ decreases and CD-↓ increases with source certainty. However, the magnitudes are asymmetric: across most models and both domains, inflation of uncertain sources is more frequent than deflation of certain sources. Error bars show 95% CI.

regardless of the source's starting point.

Certainty Distortion is not a sampling artifact. Our main results use top-p sampling (`top-p = 0.9` and `T = 1`), reflecting practical settings (Holtzman et al., 2020). To distinguish model behavior from stochastic decoding effects, we re-run experiments with greedy decoding (`T = 0`). Greedy decoding changes certainty distortion by only 1.1 points on average (95% CI $[-0.7, 2.7]$), and a paired McNemar test with Holm correction finds no significant difference for any model (Figure 28). The largest difference is a 4.2-point increase for `Llama 3.3 (70B)`. Increases and decreases in expressed certainty change by less than one point on average, so greedy decoding does not alter the direction of distortion either. Thus, unlike prior work showing that lower temperature reduces other forms of information distortion in content-constrained tasks (Mohamed et al., 2025; Perez et al., 2025), certainty distortion persists under deterministic decoding.

4.3 The Broken Telephone Effect

The results so far focus on certainty distortion after a single LM transformation. We next ask whether repeatedly applying the same task compounds the observed distortion (the "broken telephone" effect),

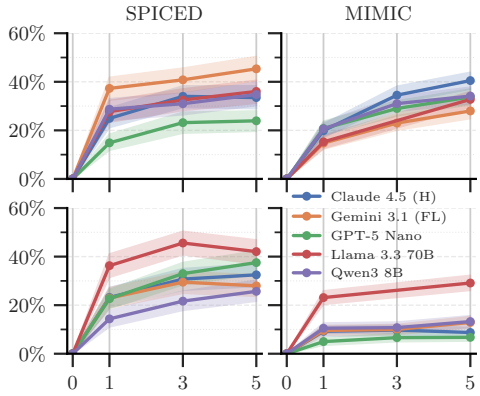


Figure 4: Certainty distortion across iterations (paraphrasing task) measured with respect to the certainty of the original text. In SPICED (left), most distortion is observed within one LM call, with little relative change afterwards. In MIMIC (right), CD-↑ continues climbing after 5 LM calls, while CD-↓ remains flat, indicating directional compounding towards inflated certainty.

or if it remains the same. This has been investigated for information distortion (Mohamed et al., 2025; Perez et al., 2025) but not for certainty distortion. To examine this, we simulate a communication chain where each model rewrites its own previous output for up to five iterations, comparing each intermediate output to the original source text.

As shown in Figure 4, in the scientific domain, most certainty distortion occurs after the first call: both CD-↑ and CD-↓ rise sharply between turns 0 and 1, plateauing thereafter. On the other hand, **in the medical domain, CD-↑ continues to climb through turn 5**: `Claude 4.5 (H)`, for instance, goes from 20% CD-↑ after one iteration to 40% after five. Meanwhile, CD-↓ remains essentially flat for all models throughout, remaining under 15% for four of five models, with `Llama 3.3 (70B)` as the exception. In other words, repeated paraphrasing in the medical domain compounds certainty distortion directionally, producing progressively more outputs with greater expressed certainty than the original source. One possible contributor is the greater density of uncertainty-bearing language in the medical sources: a descriptive lexicon-based analysis finds 4.84 epistemic markers per 100 words in MIMIC, compared with 1.54 in SPICED (Appendix F.5). This difference provides more opportunities for such cues to be progressively weakened, although it does not establish the mechanism behind the divergent trajectories.

Overall, the observed LM chain results raise

the question of whether the stabilization and directional compounding patterns observed here would also emerge in human rewriting chains. Prior work documents certainty distortion across stages of science communication but does not isolate fixed-task sequential rewriting chains (Brechman et al., 2009; Sumner et al., 2016; Wuehrl et al., 2024), while studies of human communication chains show that repeated retelling can preserve some dimensions while altering others (Breithaupt et al., 2018), consistent with formal accounts of iterated learning (Griffiths and Kalish, 2007). Future work could address this gap by studying expressed certainty in fixed-task human rewriting chains, comparing instructions to preserve meaning alone with instructions to preserve both meaning and certainty.

4.4 Distortion as a Function of LM Scale

So far, our analyses have focused on single models across 5 different model families. To assess whether distortion is an artifact of model scale or persists in larger, more capable models, we extend the analysis to bigger variants of `gpt-5.4-nano`, `Claude 4.5 (H)`, and `Gemini 3.1 (fl)`. For cost reasons, we focus on the two tasks with the highest distortion rates in each domain: news rewriting in SPICED and paraphrasing in MIMIC. Table 3 summarizes the results.

Overall, ***even the largest evaluated LMs distort certainty***, and the effects of model size vary across families and tasks. Among the largest variants, CD ranges from 46.1–69.5% on SPICED and from 18.6.0–27.4% on MIMIC. On news rewriting, CD drops from 74.8% to 55.7% across the Claude family and from 75.3% to 69.5% across Gemini variants, whereas the GPT family shows no overall improvement (44.3% to 46.1%). On medical paraphrasing, increasing model size produces no consistent improvement across families.[7] Across all eight model variants, no model achieves distortion below 39.5% or 11% in SPICED and medical tasks, respectively. These results indicate that certainty distortion is a phenomenon that persists across LM scales.

	SPICED - News		MIMIC - Para	
Model	**CD**	**CD-↑**	**CD**	**CD-↑**
`gpt-5.4-nano`	44.3	23.2	25.8	20.8
`gpt-5.4-mini`	39.5	15.1	11.1	9.8
`gpt-5.4`	46.1	15.1	19.0	16.2
`claude-haiku-4-5`	74.8	49.6	29.1	19.9
`claude-sonnet-4-6`	66.8	40.1	27.0	11.4
`claude-opus-4-7`	55.7	31.7	27.4	17.4
`gemini-3.1-flash-lt`	75.3	48.6	24.3	14.8
`gemini-3.5-flash`	69.5	49.9	18.6	15.5

Table 3: Effect of model scale on certainty distortion (%) for news rewriting (SPICED) and paraphrasing (MIMIC). Effects vary across model families and tasks, and no variant achieves below 11% distortion.

5 Beyond Single Sentence Tasks

To test whether certainty distortion generalizes beyond sentence-level transformation, we extend our analysis to document-level tasks. We introduce a new dataset of 500 academic abstracts spanning five disciplines across social and life sciences, and use the full Findings section of 800 MIMIC-CXR reports, restricted to reports containing at least one uncertainty-bearing sentence to enrich the sample for certainty variation (Appendix E). We introduce a summarization task to simulate a literature review task, while extending the other tasks (paraphrasing and simplification) to the whole document. Because outputs can rephrase, drop, or merge claims, we extend our LM-as-a-judge with an alignment step that locates output counterparts (if any) for finding-level claims in academic abstracts and sentences in medical reports before comparing certainty. We conduct a controlled sanity check of this evaluator using certainty-varied rewrites (Appendix G) and leave full validation for future work.

Results. Certainty distortion persists in document-level transformations, with patterns that echo our sentence-level findings: distortion is higher in the scientific domain than in the medical domain, and increases in expressed certainty generally occur more often than decreases. Table 4 shows that, across models, increases in expressed certainty occur at 3.1–11.3× the rate of decreases for academic abstracts. The asymmetry is weaker for medical reports, where ratios range from 0.9× to 4.6×, with increases occurring at more than twice the rate of decreases for three of the five models, although overall distortion is lower (2–5%). We view these results as initial evidence of document-level certainty distortion; independent validation of the alignment-based evaluator and investigation of the cross-domain gap are promising directions for future work.

[7] A BERT-based estimator recovers the GPT-variant ordering, reducing concerns about judge self-preference (Appendix F.6).

Model	Academic Papers		MIMIC	
	CD	CD-↑	CD	CD-↑
gpt-5.4-nano	15.0	$11.3_{3.1}$	2.0	$1.4_{2.1}$
Claude 4.5 (H)	22.8	$21.0_{11.3}$	3.4	$2.8_{4.6}$
Gemini 3.1 (fl)	24.8	$22.7_{10.8}$	4.1	$3.3_{4.0}$
Qwen3 (8B)	18.4	$15.8_{6.3}$	5.0	$2.7_{1.1}$
Llama 3.3 (70B)	15.8	$12.3_{3.6}$	4.5	$2.1_{0.9}$

Table 4: Certainty distortion (%) on document-level scientific and medical tasks, micro-averaged over aligned claims and sentences from 500 abstracts and 800 reports, respectively. CD-↑ denotes increased expressed certainty; subscripts report inflation-to-deflation ratios.

Intervention	Paraphrase		News	
	CD	CD-↑	CD	CD-↑
original	50.6	$26.7_{1.4}$	64.3	$39.7_{1.6}$
+ *preserve semantics*	47.4	$25.8_{1.4}$	51.0	$33.1_{1.9}$
+ *do-not-hallucinate*	42.8	$24.6_{1.4}$	47.3	$30.7_{1.9}$
+ *preserve-certainty*	40.1	$19.4_{1.0}$	41.9	$18.5_{1.1}$

Table 5: Cumulative prompt interventions, averaged across models in SPICED. Subscripts on CD-↑ report the inflation-to-deflation ratio. Only *preserve-certainty* comes close to neutralizing the directional bias.

6 Prompt-Based Mitigation

Now, we test whether standard, low-cost prompting interventions that require no access to model weights or training pipelines are sufficient to mitigate certainty shifts. Specifically, we test restricting the original prompts with additional constraints, including (1) *preserve-semantics*, which instructs to keep meaning intact; (2) *do-not-hallucinate*, which instructs against unsupported content; and (3) *preserve-certainty*, which explicitly instructs against shifting the certainty of the original text. All three are appended to the base prompt; full text in Appendix C. We test the efficacy of these prompts on the scientific domain tasks because these exhibited the highest CD.

Results. Prompt interventions reduce distortion partially but do not eliminate it (Table 5). Stacking all three constraints lowers CD from 50.6% to 40.1% on paraphrase and from 64.3% to 41.9% on news. However, we find that adding *preserve certainty* brings the inflation-to-deflation ratio close to or below 1, while *preserve semantics* and *do not hallucinate* reduce both directions proportionally and leave the asymmetry intact.

These results distinguish between correcting an aggregate directional bias and preserving certainty in individual rewrites. Although *preserve certainty* approximately balances inflation and deflation, around 40% of rewrites remain distorted, hence suggesting that zero-shot prompting rebalances errors rather than preventing them. More effective mitigation may require instance-level feedback that explicitly compares rewrites with their source. Few-shot demonstrations could better specify certainty-preserving behavior, while verifier-guided generation or reranking (Ichihara et al., 2025) could reject candidate rewrites that alter the source's epistemic stance. Training-based approaches could instead use supervised certainty-preserving rewrites or preference (Rafailov et al., 2023) and reinforcement-learning objectives (Ouyang et al., 2022) that penalize such shifts.

7 Conclusion

LMs can distort human-generated text in ways that change the information conveyed (Mohamed et al., 2025; Abdulhai et al., 2026). We focused on certainty distortion: shifts in expressed certainty that occur during rewriting intended to preserve meaning. Using an LM-based metric validated against human annotations, we characterized this phenomenon across five model families on scientific and medical domains. We find that certainty distortion affects a non-negligible fraction of LM outputs and is directionally asymmetric, with most models being more likely to inflate the certainty of the source than to attenuate it. This directional tendency compounds under repeated LM paraphrasing in the medical domain. Larger model variants and changes in decoding strategy do not consistently reduce distortion, while prompt-based interventions reduce but do not eliminate it. These results suggest that preserving the expressed certainty of aligned claims deserves attention as a first-class evaluation target alongside factuality and semantic fidelity.

Limitations

Cross-Domain and Cross-Lingual Generalization. We evaluate certainty distortion only in English and within scientific and medical communication tasks. Although we observe distortion across both domains, the differences in magnitude suggest that domain characteristics, such as whether claims are general statements about the world versus patient-specific observations, play a meaningful role. Extending the analysis to additional domains (*e.g.*, legal, financial) and task types would help clarify how broadly these patterns generalize. Ad-

ditionally, performing cross-lingual analysis would be a valuable direction, as hedging conventions and epistemic markers vary considerably across languages (Lauwereyns, 2002; Yagız and Demir, 2014; Shafqat et al., 2019; Mur-Dueñas, 2021).

Document-Level Evaluation Validity. While our sentence-level evaluation is validated against human judgments, the document-level experiments rely on claim-alignment procedures that have not been comparably validated through human studies. To partially address this limitation, we assessed the medical evaluator using 160 controlled source–rewrite pairs that preserved, increased, or decreased certainty. Although this does not constitute an independent quantitative validation, its judgments generally followed the intended direction of the certainty shifts, providing preliminary evidence that it captures directional distortion (Appendix G). We therefore view the document-level results as initial evidence and leave more rigorous human validation to future work. Techniques such as claim decomposition (Wanner et al., 2024; Javaji et al., 2025), explicitly isolating individual propositions within a sentence before comparing certainty, could improve document-level evaluation in future studies.

Mechanisms of Certainty Distortion. Finally, although we observe lower one-turn distortion in the medical domain and greater directional compounding under repeated paraphrasing, our experiments do not identify the mechanism behind this cross-domain difference. Our analyses nonetheless constrain the space of explanations: distortion persists under greedy decoding (ruling out a sampling artifact), varies systematically by task and domain (implicating the communication setting), and exceeds what floor/ceiling effects alone predict. One hypothesis is that a model's internal belief about a claim, *i.e.*, whether it supports or opposes it, drives certainty inflation or deflation during rewriting. To test this, we conducted a belief elicitation experiment on the SPICED dataset: for each source claim, we collected model beliefs by normalizing certainty and stance into a single alignment score reflecting whether the model supported or opposed the claim, and compared this against the observed direction and magnitude of certainty distortion. We found no meaningful correlation (Kendall's $\tau_B \approx -0.1$), suggesting that prior belief alone does not explain the observed distortion patterns. An alternative explanation is that models internalize domain- and task-specific writing conventions during training. Science journalism, for instance, is known to conventionally favor confident, attention-grabbing language that strips hedges from original findings (Sumner et al., 2016; Dempster et al., 2022), whereas radiology reports follow well-established conventions for expressing graded uncertainty (Lee and Whitehead, 2017; Shinagare et al., 2019), and academic scientific writing treats hedging itself as a disciplinary norm (Hyland, 1998). If models learn these register-specific patterns during pretraining, they may apply them during rewriting regardless of the source text's original certainty level. A third, non-exclusive hypothesis concerns post-training, as instruction tuning and preference optimization may favor responses that appear fluent, helpful, and decisive over preserving epistemic uncertainty. Prior work comparing base, instruction-tuned, and RLHF models on question answering tasks finds that a preference for certainty strengtheners emerges after RLHF, while reward models and human preference data penalize expressions of uncertainty (Zhou et al., 2024). Related tendencies of post-trained models toward sycophancy provide broader evidence that preference optimization can systematically alter how models communicate claims (Perez et al., 2023; Sharma et al., 2024). However, these studies do not examine meaning-preserving rewriting and therefore do not establish post-training as the cause of the certainty distortion observed here. Disentangling these potential mechanisms remains an important direction for future work.

Societal Impact

Certainty distortion may have societal consequences even when rewritten text remains factually and semantically close to its source. In particular, certainty inflation can make tentative or weakly supported claims appear established, potentially amplifying misleading information and encouraging overreliance in high-stakes domains. Certainty deflation may obscure well-supported conclusions.

Acknowledgments

We thank the members of MADLAB and UCINLP for their valuable feedback and discussions throughout this project. We are especially grateful to Yu Fei, Preethi Seshadri, Lukas Mayer, and ZhaoBin Li for their insightful comments. This work was supported by the National Science Foundation under NSF CAREER award IIS-2046873 and award NSF 2505006, by the Hasso Plattner Institute (HPI)

Research Center in Machine Learning and Data Science at UCI, and by funding support from Google and SAP. The views and conclusions expressed in this paper are those of the authors and should not be interpreted as representing the official policies, views, or endorsements of the funding agencies or organizations.

References

Marwa Abdulhai, Isadora White, Yanming Wan, Ibrahim Qureshi, Joel Leibo, Max Kleiman-Weiner, and Natasha Jaques. 2026. How LLMs distort our written language. *Preprint*, arXiv:2603.18161.

Anthropic. 2024. Claude Haiku 4.5 model snapshot (claude-haiku-4-5-20251001). https://docs.anthropic.com/en/docs/about-claude/models/all-models. Last accessed: 2026-08-30.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345.

Jean Brechman, Chul-joo Lee, and Joseph N. Cappella. 2009. Lost in translation?: A comparison of cancer-genetics reporting in the press release and its subsequent coverage in the press. *Science Communication*, 30(4):453–474.

Fritz Breithaupt, Binyan Li, Torrin M. Liddell, Eleanor B. Schille-Hudson, and Sarah Whaley. 2018. Fact vs. affect in the telephone game: All levels of surprise are retold with high accuracy, even independently of facts. *Frontiers in Psychology*, 9.

David V. Budescu, Shalva Weinberg, and Thomas S. Wallsten. 1988. Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, 14(2):281–294.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Georgia Dempster, Georgina Sutherland, and Louise Keogh. 2022. Scientific research in news media: a case study of misrepresentation, sensationalism and harmful recommendations. *Journal of Science Communication*, 21(01):A06.

Mandeep K Dhami and David R Mandel. 2022. Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*, 26(6):514–526.

Baruch Fischhoff and Alex L. Davis. 2014. Communicating scientific uncertainty. *Proceedings of the National Academy of Sciences*, 111:13664–13671.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Joe Fore. 2019. "A court would likely (60–75%) find . . .": Defining verbal probability expressions in predictive legal analysis. *Legal Communication & Rhetoric: JALWD*, 16:49–84.

Google DeepMind. 2026. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-05-03.

Thomas L. Griffiths and Michael L. Kalish. 2007. Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31(3):441–480.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on LLM-as-a-judge. *Preprint*, arXiv:2411.15594.

Arnav Gudibande, Eric Wallace, Charlie Victor Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2024. The false promise of imitating proprietary language models. In *The Twelfth International Conference on Learning Representations*.

Rose E. Guingrich, Dvija Mehta, and Umang Bhatt. 2026. Belief offloading in Human-AI interaction. *Preprint*, arXiv:2602.08754.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins Publishing Company.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Eiji Uchibe. 2025. Evaluation of Best-of-N sampling strategies for language model alignment. *Transactions on Machine Learning Research*.

Shashidhar Reddy Javaji, Yupeng Cao, Haohang Li, Yangyang Yu, Nikhil Muralidhar, and Zining Zhu. 2025. Can AI validate science? benchmarking LLMs on claim → evidence reasoning in AI papers. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th*

Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2355–2379, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm not sure, but…": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA. Association for Computing Machinery.

Shizuka Lauwereyns. 2002. Hedges in Japanese conversation: The influence of age, sex, and formality. *Language Variation and Change*, 14(2):239–259.

Bonmyong Lee and Matthew T. Whitehead. 2017. Radiology reports: What YOU think you're saying and what THEY think you're saying. *Current Problems in Diagnostic Radiology*, 46(3):186–195.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2025. Are LLM-Judges robust to expressions of uncertainty? investigating the effect of epistemic markers on LLM-based evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8962–8984, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ge Luo, Hebi Li, Youbiao He, and Forrest Sheng Bao. 2022. PrefScore: Pairwise preference learning for reference-free summarization quality assessment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5896–5903, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Meta AI. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Amr Mohamed, Mingmeng Geng, Michalis Vazirgiannis, and Guokan Shang. 2025. LLM as a broken telephone: Iterative generation distorts information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7493–7509, Vienna, Austria. Association for Computational Linguistics.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. spacy.

Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. 2024. MT-Ranker: Reference-free machine translation evaluation by inter-system ranking. In *The Twelfth International Conference on Learning Representations*.

Pilar Mur-Dueñas. 2021. There may be differences: Analysing the use of hedges in English and Spanish research articles. *Lingua*, 260:103131.

OpenAI. 2026. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`. Accessed: 2026-05-01.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2021. Measuring sentence-level and aspect-level (un)certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9959–10011, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Jérémy Perez, Grgur Kovač, Corentin Léger, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier. 2025. When LLMs play the telephone game: Cultural attractors as conceptual tools to evaluate LLMs in multi-turn settings. In *The Thirteenth International Conference on Learning Representations*.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sanjana Ramprasad and Byron C. Wallace. 2025. Do automatic factuality metrics measure factuality? A critical evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

P. V. Rao and L. L. Kupper. 1967. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317):194–204.

Victoria L. Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144, Rochester, New York. Association for Computational Linguistics.

Yan Scholten, Stephan Günnemann, and Leo Schwinn. 2025. A probabilistic perspective on unlearning and alignment for large language models. In *The Thirteenth International Conference on Learning Representations*.

Asmara Shafqat, Rafique Ahmed Memon, and Huma Akhtar. 2019. Cross-cultural analysis of the use of hedges in European and Pakistani English newspaper: A corpus-based study. *International Journal of English Linguistics*, 9(5):126.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna M. Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Atul B. Shinagare, Ronilda Lacson, Giles W. Boland, Aijia Wang, Stuart G. Silverman, William W. Mayo-Smith, and Ramin Khorasani. 2019. Radiologist preferences, agreement, and variability in phrases used to convey diagnostic certainty in radiology reports. *Journal of the American College of Radiology*, 16(4 Pt A):458–464.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andrew Williams, Lewis Bott, Rachel Adams, Christos A. Venetis, Leanne Whelan, Bethan Hughes, and Christopher D. Chambers. 2016. Exaggerations and caveats in press releases and health-related science news. *PLOS ONE*, 11(12):e0168217.

Yana van de Sande, Gunes Açar, Thabo van Woudenberg, and Martha Larson. 2025. On fact and frequency: LLM responses to misinformation expressed with uncertainty. *Preprint*, arXiv:2503.04271.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S11):S9.

Robin Vujanic and Thomas Rückstieß. 2026. LEAF: Knowledge distillation of text embedding models with teacher-aligned representations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43362–43383, San Diego, California, United States. Association for Computational Linguistics.

Thomas S. Wallsten, Yaron Shlomi, and Hisuchi Ting. 2008. Exploring intelligence analysts' selection and interpretation of probability terms: Final report for research contract 'expressing probability in intelligence analysis'.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Qipeng Guo, Xiangkun Hu, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Xuming Hu, Zehan Qi, Wenyang Gao, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2025. Survey on factuality in large language models. *ACM Comput. Surv.*, 58(1).

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. 2022. Modeling information change in science communication with semantically matched paraphrases. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1783–1807, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zekun Wu, Mayank Jobanputra, Vera Demberg, Jessica Hullman, and Anna Maria Feit. 2025. How AI responses shape user beliefs: The effects of information detail and confidence on belief strength and stance. *Preprint*, arXiv:2511.09667.

Amelie Wuehrl, Dustin Wright, Roman Klinger, and Isabelle Augenstein. 2024. Understanding fine-grained distortions in reports of scientific findings. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6175–6191, Bangkok, Thailand. Association for Computational Linguistics.

Oktay Yagız and Cuneyt Demir. 2014. Hedging strategies in academic discourse: A comparative analysis of Turkish writers and native writers of English. *Procedia - Social and Behavioral Sciences*, 158:260–268.

Joonho Yang, Seunghyun Yoon, Hwan Chang, Byeongjeong Kim, and Hwanhee Lee. 2026. Hallucinate at the last in long response generation: A case study on long document summarization. *Preprint*, arXiv:2505.15291.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59670–59684. PMLR.

James Xu Zhao, Jimmy Z.J. Liu, Bryan Hooi, and See-Kiong Ng. 2025. How does response length affect long-form factuality. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3102–3125, Vienna, Austria. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Markus Zopf. 2018. Estimating summary quality with pairwise preferences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1687–1696, New Orleans, Louisiana. Association for Computational Linguistics.

A Automatic Metric Validation

Automatic metric validation proceeded in two stages. First, we began with a validation of different certainty estimators (*i.e.*, LM-as-a-judge approach, and fine-tuned BERT models (Pei and Jurgens, 2021; Wuehrl et al., 2024)) in a set of synthetically generated datasets, controlled for levels of certainty distortion. Once we were comfortable with the proposed approaches, we conducted the user study to validate our findings. The remainder of this section is organized as follows:

- Appendix A.1 provides concrete examples where BERT (Wuehrl et al., 2024) and our proposed LM-as-a-judge approach disagree.
- Appendix A.2 describes the controlled experiments conducted earlier on in this project that led to the final proposed LM-as-a-judge metric, using `gpt-5.4-mini`.

A.1 Qualitative Analysis

Tables 6 and 7 represent examples where the proposed LM-as-a-judge metric and two existing BERT-based metrics disagree. In addition to listing the original (or source text) and an LM-generated text, we also report the predicted scores by the two evaluators.

A.2 Controlled Evaluation

With the goal of better understanding how to detect differences in certainty, we conduct a control experiment where for each scientific finding in SPICED, we create three controlled subsets:

- `Same-SciFi`: pairs of texts where the confidence in the modified text x_M is equivalent to the original text x_o, *i.e.*, $\delta(x_o, x_M) \approx 0$;
- `Less-SciFi`: pairs of texts where the confidence in the modified text x_M is explicitly lower than the original text x_o, *i.e.*, $\delta(x_o, x_M) < 0$;
- `More-SciFi`: pairs of texts where the confidence in the modified text x_M is explicitly higher than the original text x_o, *i.e.*, $\delta(x_o, x_M) > 0$.

A.2.1 Synthetic Control Dataset Construction

This section describes the creation process for the controlled datasets. We restrict this analysis to the examples in the SPICED dataset (Wright et al., 2022), whose final information matching score is less than 3. Note that this represents the complement of the subset used in the main paper, which concerned the subset whose score was ≥ 4. Additionally, we filter out incomplete sentences (*e.g.*, "As illustrated in the scatter plot (Fig.") using spaCy's constituency parser (Montani et al., 2023). Finally, we randomly sample 400 scientific findings, which we use to create the synthetic datasets.

To obtain the three variants, we prompt `GPT-5.1` and `Gemini 3.1 Pro (Preview)` to perform targeted minimal edits s.t. only the epistemic language associated with the main reported finding is affected. Both models are called with `T=0.5`, providing a good trade-off between diversity and prompt-adherence.[8] All implementation details will be shared in a public repository.

The resulting controlled datasets provide pairwise rankings of epistemic certainty rather than absolute scores. This choice is motivated by the subjectivity of certainty judgments, which often leads to low inter-annotator agreement (Pei and Jurgens, 2021; Wuehrl et al., 2024). Additionally, prior work has shown that relative comparisons yield more reliable judgments than absolute scoring across tasks such as machine translation (Callison-Burch et al., 2007; Moosa et al., 2024), summarization (Zopf, 2018; Luo et al., 2022), and other generative evaluation settings (Zheng et al., 2023). With these controlled datasets we can evaluate whether a certainty scoring function can detect *directional changes in certainty*, assigning higher scores to the more certain variant and lower scores to the less certain variant within each pair, while being able to distinguish true paraphrases that preserve the certainty.

A.2.2 Dataset Statistics and Validation

This section validates that the generated variants (1) preserve the semantics of the original sentences and (2) modify the epistemic language such that the text is perceived as more certain in `More-SciFi` and less certain in `Less-SciFi`.

Length Analysis. We begin by measuring differences in average number of sentences between the original scientific finding and the modified variant. Using `nltk.sent_tokenize` to partition the text, we find all outputs to have matching length, suggesting that modifications remain at the sentence-level.

Word Analysis. The number of sentences provides only a coarse signal about the extent of the modifications. To better capture how much the text changes, we also evaluate the Word Edit Distance (WED), which measures the minimum number of word-level operations–additions, removals, or replacements–required to convert one sentence into another. In our datasets, models are instructed to produce minimally edited versions of the original sentence. Therefore, we expect WED values to be low, meaning that only a few operations are needed to transform the original

[8] Experiments were conducted throughout March 2026.

Original	LM-generated	BERT	LLM
'In particular, we find that a strong social media presence **may be** more important than being under contract with a top agency, or than the aesthetic standards sought after by the industry.'	'**Notably**, having a strong presence on social media **can matter** more than being signed with a leading agency or meeting the fashion industry's preferred look.'	-1.0	1.5
'To the best of our knowledge, this is the first convincing demonstration of a professional group showing higher accuracy on face matching tasks.'	'A notable breakthrough has been achieved, with a particular profession exhibiting unparalleled precision in identifying and matching facial features.'	-0.97	2.0
'Yield increase is therefore shown to be more effective in reducing land use when combined with land-use regulation and consumption changes.'	'The results suggest that pairing higher yields with policies that regulate land use and shift consumption patterns more effectively lowers land demand.'	0.62	-1.0
'As exercise capacity improves, progression towards longer sessions of at least 20 min duration on most days of the week is recommended.'	'It is suggested that as physical fitness increases, individuals should aim to engage in extended periods of exercise, ideally lasting 20 minutes or more, on a majority of days throughout the week.'	0.62	-1.0
'Across three studies, well-educated graduates who entered the workforce during economic downturns were more satisfied with their current jobs than those who entered during more prosperous economic times.'	'Three studies revealed that highly educated graduates entering the job market during economic recessions reported greater job satisfaction compared to those who joined during more prosperous periods.'	1.0	0.0
'Teens starting school at 8:30 AM or later are the only group with an average time in bed permitting 8 hours of sleep, the minimum recommended by expert consensus for health and well-being.'	'Experts agree that teenagers need at least eight hours of sleep for optimal health, and only those whose school day begins at 8:30 AM or later typically get enough rest.'	1.0	0.0
The towns and villages that constituted research sites for our study are **likely** similar to hundreds or even thousands of sites throughout Asia, where humans live alongside synanthropic macaques.	Our study's research sites, comprising towns and villages, **likely** resemble hundreds or thousands of similar locations across Asia where human populations coexist with synanthropic macaques.	0.89	0.5
Study anticoagulant was discontinued because of bleeding in 242 patients (4.7%) in the otamixaban group and in 95 patients (1.7%) in the UFH-pluseptifibatide group (P < .001).	The study found that significantly more patients in the otamixaban group had to stop treatment due to bleeding compared to those in the UFH plus eptifibatide group.	0.99	-0.5

Table 6: Qualitative analysis of edge cases where the reference-less BERT-based model proposed by Pei and Jurgens (2021) disagrees with the pairwise-ranking LM-as-a-judge approach proposed in the paper. The first four examples were selected from extreme cases, and the remaining four were randomly picked.

sentence into the LM-generated variants. WED is computed on lowercase, non-punctuated text using `rapidfuzz.distance.Levenshtein`. Figure 5 breaks down the WED statistics by dataset and models. Considering `Less-SciFi`, we observe that the WED generally ranges from 2 to 6 operations. These edits include replacing assertive language such as "is", "has", "showed", "significantly", "improved", and "strongly" with modal verbs like "may", "might", and "could", as well as other less certain expressions such as "appeared", "appears", "suggest", and "potentially". We also manually examine examples with WED > 6 and find that higher WED values are typically associated with replacing multiple occurrences of epistemic language within the same sentence or replacing a single word with a multi-word phrase (*e.g.*, "point to" is replaced with "seem to point to", "fits" is replaced with "appears to be consistent with"). When considering `More-SciFi`, we observe slightly lower WED values, indicating that fewer modifications are required compared to `Less-SciFi`. In many cases, it suffices to add adverbs such as "definitely", "clearly", or "strongly" to increase the expressed certainty. We also observe that modal verbs (*e.g.*, "may", "can", "might", "should") and hedges (*e.g.*, "possible", "believe", "appears", "suggests") are often replaced with more assertive verbs such as "confirms", "are", "demonstrate", and "establish". As before, we find higher WED values to be associated with replacing a word or short phrase with a longer expression (*e.g.*, "strong majorities" replaced with "It is a demonstrated fact that").

Semantic Analysis. The word-level analysis indicates that LM-generated sentences are minimally edited relative to the originals, but it provides limited insight into whether their semantics are preserved. To quantify semantic preservation, we also measure pairwise cosine similarity in the control datasets using `MongoDB/mdbr-leaf-mt` (Vujanic and Rückstieß, 2026), a compact, fast text embedding model that is competitive with OpenAI's text-embedding-3-small. Intuitively, if the generated

Original	LM-generated	BERT	LLM
'Underlying consolidation would be difficult to exclude.'	'It is hard to rule out lung consolidation that is not seen on imaging.'	2.0	-0.5
'The cardiac silhouette **appears** top normal in size.'	'The heart is a normal size.'	0.0	2.0
'Please note that pericardial fluid would be difficult to exclude on this study.'	'It is difficult to determine whether or not there is fluid around the heart based on this scan.'	-1.0	2.0
'Increased interstitial markings which **could** represent a component of interstitial edema or chronic lung disease.'	'The radiographic image shows enhanced interstitial patterns, **suggestive** of either interstitial fluid accumulation or a chronic pulmonary condition.'	2.0	-0.5
'Some focal opacity in the right upper zone laterally is noted – **if real**, this is new, but it **could** represent artifact due to the presence of the subcutaneous emphysema.'	'A small area of abnormal density is visible in the outer part of the right upper lung zone, although it is **unclear** whether this is a genuine finding or an artifact caused by air trapped under the skin.'	2.0	-0.5
'Airspace consolidation is noted within the right lower lobe **concerning** for pneumonia.'	'The chest X-ray shows a cloudy area in the lower right lung, **indicating a strong likelihood** of pneumonia.'	0.0	2.0

Table 7: Qualitative analysis of edge cases where the reference-less fine-tuned BERT-based model proposed by Wuehrl et al. (2024) disagrees with the pairwise-ranking LM-as-a-judge approach proposed in this paper. The first three examples are drawn from the MIMIC-CXR (Johnson et al., 2019) paraphrasing task, and the remaining three are drawn from the simplification task.

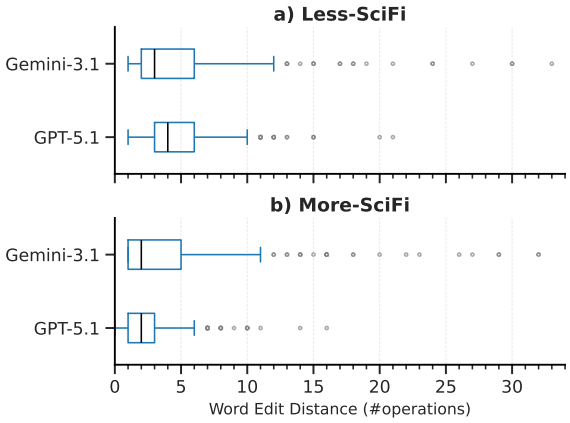


Figure 5: Word Edit Distance (WED) for the two SPICED control datasets. We discriminate the number of operations by model. We observe that the median number of modifications ranges between 3 and 4 in `Less-SciFi` and 2 for the `More-SciFi`.

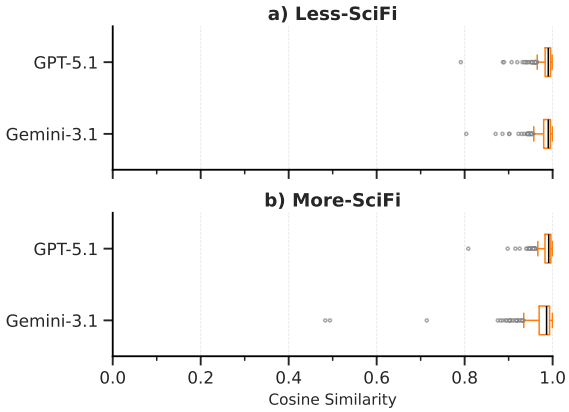


Figure 6: Semantic similarity analysis for `Less-SciFi` and `More-SciFi`. The median cosine similarity is close to 1 across models and dataset variants, indicating high average similarity between the source and transformed texts.

outputs retain the meaning of the original SPICED sentence, the cosine similarity is expected to be close to 1. Figure 6 shows high cosine similarity across both control datasets. Upon manually inspecting examples with lower cosine similarity (<0.9; n = 6), we find that these cases are typically either shorter sentences, which are more sensitive to small word changes (*e.g.*, "It thus does not fit our definition." versus "It thus may not fully fit our definition.") or examples where the register differs, such as a shift from informal spoken language to formal written style. Interestingly, we observe similar patterns for the examples in `More-SciFi` with lower cosine similarity (<0.9; n = 10). We also find that the lower similarity scores in 4 out of 10 examples are associated with incomplete sentences. This can be attributed to our prompting design, which generates both certainty variants in a single model call while restricting the number of output tokens. In the analysis that follows, we remove examples with lower cosine similarity from our control datasets, as they seem to correspond to incomplete or otherwise sensitive generations, totaling n=794 and n=790 examples for `Less-SciFi` and `More-SciFi`, respectively.

Manual Validation. To complement the automated evaluation, three authors of this paper manually annotated 150 paired examples, equally bal-

anced across the three variants. We observe substantial agreement among annotators, as indicated by a Fleiss' kappa of $\kappa = 0.68$ (Fleiss, 1971). A more fine-grained analysis across subsets shows that agreement is higher for `Less-SciFi` and `More-SciFi` ($\kappa \approx 0.74$), but substantially lower for `Same-SciFi` ($\kappa \approx 0.22$). This lower agreement is expected, as examples in `Same-SciFi` exhibit minimal or no differences in certainty, making the annotation task inherently more ambiguous.

A.2.3 Evaluation of Certainty Estimators

With these controlled datasets, we can now compare how well different measures of uncertainty capture the directional biases in the control data. Specifically, we evaluate previously proposed BERT-based fine-tuned models (Pei and Jurgens, 2021; Wuehrl et al., 2024) as well as various prompting approaches.[9]

A.2.3.1. Aspect-Level Certainty Estimates We start by examining the outputs of BERT-based models fine-tuned to estimate certainty in the context of scientific communication. Through several rounds of annotation, Pei and Jurgens (2021) collect human annotations of certainty in scientific findings across various fields, achieving an average Krippendorff's α of 0.57. The annotations cover 6 certainty aspects: Condition, Extent, Framing, Numerical, Probability, Suggestion. Using a subset of this (imbalanced) annotated dataset, the authors fine-tune a SciBERT model (Beltagy et al., 2019) to predict a categorical distribution over three classes–Uncertain, NotPresent, Certain–for each aspect. The proposed model outperforms four other baselines[10] across different aspects, observing lower F1-scores for less common aspects such as Condition or Suggestion. For that reason, we only consider the model predictions for four aspects: Extent, Framing, Number, and Probability.

In the main paper, we hypothesize the existence of an *oracle certainty scoring function* that maps any text to an absolute certainty score, $s : \mathcal{X} \mapsto [0, 1]$. However, this differs from the fine-tuned SciBERT model, which outputs a probabilistic distribution over three classes for each aspect (mathematically, $f : \mathcal{V} \mapsto [0, 1]^3$ where $\sum_i f[i] = 1$ and f is an aspect-specific classifier) and therefore requires an additional transformation to produce a scalar score.

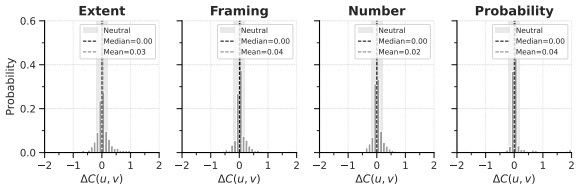
Methodology. One simple approach is to consider the class-wise difference. Given two sentence-level inputs u and v, we compute $\Delta C(u, v; i) = f(v)[i] - f(u)[i]$, where $[i]$ denotes the i-th class. However, by focusing on a single class (*e.g.*, Certain), this metric fails to account for the probability mass assigned to the other classes. To illustrate why this matters, consider the example where $f(u) = [0.4, 0.5, 0.1]$ and $f(v) = [0.7, 0.2, 0.1]$. If we only examine the difference in the Certain class (Class 2), we overlook the substantial shift in probability mass for the Uncertain class (Class 0), which may indicate that u expresses less certainty than v. An alternative approach is to aggregate the Certain and Uncertain scores into a continuous certainty measure that (1) is high when the Certain class dominates and (2) is low when the Uncertain class dominates. One possible formulation is $C(u) = f(u)[2] - f(u)[0]$, where $C(u) \in [-1, 1]$.

Results. Having defined a certainty score for each sentence-level input, we can estimate the certainty distortion for each dataset by averaging the example-level differences. Because of the way we constructed the controlled datasets, we expect a reliable certainty scoring function to correctly reflect the lower certainty in `Less-SciFi` and higher certainty in `More-SciFi`. Figure 7 shows the distortion gap distribution for both datasets across different aspects ($\Delta C(u, v; f) = C(v) - C(u)$). While the distortion gap is noticeably left-skewed for Framing and Probability in the `Less-SciFi` (with median values of -0.75 and -1.93), only negligible differences are observed for Framing and Probability in `More-SciFi` (with median values around 0.0 but mean values superior to 0.14).

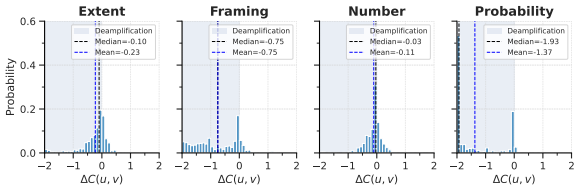
Analysis. To ensure that our findings are not specific to the scoring function $C(u) = f(u)[2] - f(u)[0]$, we repeat the analysis using a simpler certainty-specific function, $C(u) = f(u)[2]$. However, this metric also fails to accurately capture the increase in certainty in `More-SciFi`. To better understand this behavior, we examine the class boundaries of the Framing certainty classifier. We observe that, unlike in the `Less-SciFi`, there is little difference in the class distribution for `More-SciFi` (see Figure 8). This lack of sensitivity may stem from the nature of modifications in `More-SciFi`.

[9] We evaluate existing estimators, leaving the development of specialized pairwise discriminative models to future work.

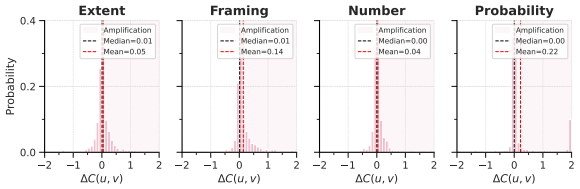
[10] These include bag-of-words and hedge-frequency baselines. Pei and Jurgens (2021) find that these baselines underperform model-based estimators and that hedges capture only part of expressed certainty. We therefore restrict our analysis to their stronger model-based estimator.



(a) `Same-SciFi`



(b) `Less-SciFi`



(c) `More-SciFi`

Figure 7: Distribution of distortion scores in sentence-level control datasets. Certainty scores are computed using the aspect-level certainty classifiers (Pei and Jurgens, 2021). Given the original scientific finding u and its modified version v, we report $\Delta C(u, v) = C(v) - C(u)$. We find that scores from Framing and Probability classifiers can be useful to detect decrease in certainty (with medians ranging between -0.75 and -1.93) but that they are not able to detect increase in certainty (with medians close to 0).

As discussed earlier, these examples are often constructed by adding adverbs and adjectives that emphasize certainty (*e.g.*, strongly, definitely, significantly). We hypothesize that for sentences already receiving high certainty scores ($f[u] > 0.2$) the addition of such terms does not further increase the predicted certainty. For instance, the original finding "These findings **suggested** the importance of liver pathogenesis in SMA." is assigned a certainty score of $f(u)[2] = 0.997$, whereas the modified version "These findings **strongly support** the importance of liver pathogenesis in SMA." is attributed a score of $f(v)[2] = 0.999$. Another example is "These examples **demonstrate** that some introns significantly boost expression, even in the absence of prior promoter activity." ($f(u)[2] = 0.98862$) versus the modified version "These examples **definitively prove** that some introns significantly boost expression, even in the absence of prior promoter activity." ($f(v)[2] = 0.99762$). This saturation effect may arise from limitations in

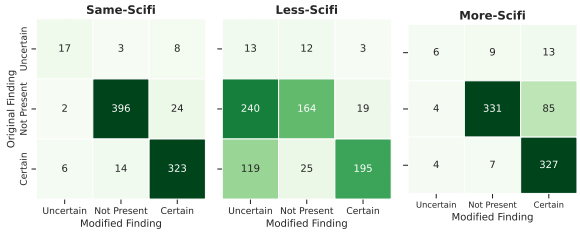


Figure 8: Class predictions using the Framing certainty classifier proposed in Pei and Jurgens (2021). Modified predictions are collected by considering the *maximum a posteriori*. Unlike the `Less-SciFi` where model classes become less certain (lower triangular), in `More-SciFi` there are no significant class shifts, which suggests the classifier lacks granularity to detect the increase in certainty.

the training data, which may not adequately capture the impact of these linguistic cues, and suggests that the classifier lacks the granularity needed to detect increases in certainty.

A.2.3.2. Global Certainty Estimates. In a similar vein, Wuehrl et al. (2024) fine-tune a SciBERT model on scientific findings annotated with a single certainty score. The annotations are collected in a 4-point Likert Scale ranging from (1) Uncertain to (4) Certain. Inter-annotator agreement is moderate (average Spearman Correlation Coefficient is 0.44), suggesting task difficulty and subjectivity of the task. Noticing the lower class prevalence of the Uncertain class, we combine the lower certainty classes into a single Uncertain class. The resulting class prevalence in training (n=1330) is 21:32:47, whereas in the test set (n=334 examples) it is 25:33:42 for (0) Uncertain, (1) Somewhat Certain, and (2) Certain classes, respectively.

Methodology. Since the original models are not publicly available, we fine-tune `allenai/scibert_scivocab_uncased` for certainty prediction. In addition to using the default hyperparameter configuration, we perform Bayesian hyperparameter optimization over the modeling objective (classification vs ordinal regression), learning rate, warm-up steps, and number of frozen layers, training a total of 80 models. Given the ordinal structure of the labels, we select the final model by **maximizing the quadratic weighted kappa (QWK)** on the test set. QWK measures the agreement between two annotators while accounting for the ordinal relationship between labels, penalizing larger disagreements more heavily than smaller ones. It takes values

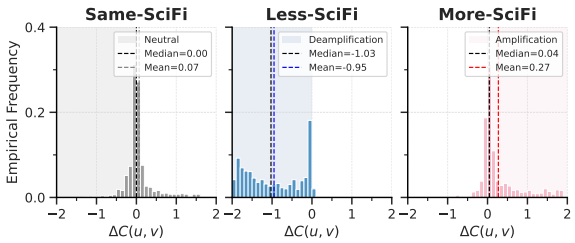
Metric	Classification BERT	Ordinal BERT
QWK	0.551	0.500
Loss	0.939	0.957
Accuracy	0.608	0.580
F1 (macro)	0.601	0.560
Precision (macro)	0.603	0.570
Recall (macro)	0.599	0.560
Per-class F1		
Class 0	0.621	0.540
Class 1	0.493	0.460
Class 2	0.688	0.680
Per-class Precision		
Class 0	0.641	0.580
Class 1	0.500	0.470
Class 2	0.669	0.650
Per-class Recall		
Class 0	0.602	0.510
Class 1	0.487	0.460
Class 2	0.707	0.710

Table 8: Comparison of performance between the classification model and the ordinal-head model. The reported macro F1 results are on par with the values reported in the original paper (mF1 = 0.53 in Table 3 in Wuehrl et al. (2024)).

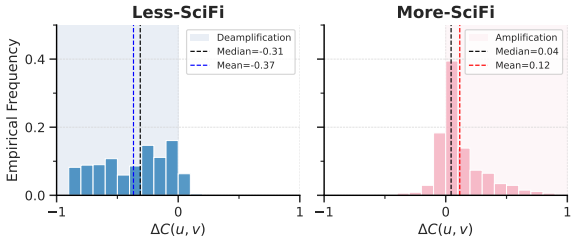
in $(-1, 1)$, where higher values indicate stronger agreement, a value of 0 corresponds to chance agreement, and negative values indicate systematic disagreement. Table 8 shows the final performance metrics on the test set for the classification and ordinal models with the highest QWK across all runs.

Results. We now assess how well each of the two models can detect the directional certainty shifts in the controlled SPICED datasets. Because the classification model predicts a 3-class distribution, we compute the certainty score for a single sentence u as the class-wise probability difference in the extreme classes, *i.e.*, $C(u) = f(u)[2] - f(u)[0]$. Figure 9a shows the certainty distortion observed in both datasets. While this method allows us to correctly detect the decrease in certainty in `Less-SciFi` (median=-1.03), it is not a reliable measure of confidence amplification in the `More-SciFi` (median: 0.04). Similar findings are observed for the ordinal model (Figure 9b).

A.2.3.3. LM-as-a-Judge: Absolute Scoring. Having considered multiple fine-tuned variants of BERT-based models, we now turn to LM-as-a-judge approaches (Liu et al., 2023; Gu et al., 2025). While prior work suggests the superiority of fine-tuned models (Wuehrl et al., 2024), we do not observe the same trend in our controlled datasets.



(a) Classification SciBERT



(b) Ordinal Regression SciBERT

Figure 9: Distribution of certainty distortion scores in the `Less-SciFi` and `More-SciFi`. Certainty scores are computed using SciBERT models trained to predict a single absolute certainty score (Wuehrl et al., 2024). The score for a single sentence u is computed as $C(u) = f(u)[2] - f(u)[0]$, where the two terms correspond to the likelihood of the certain and uncertain classes for sentence u according to the classifier f.

These differences can be attributed to both model capabilities and prompting strategies: prior work evaluates older-generation chat models, such as `Llama 2 (Chat-13B)` and `Mixtral (8x7B)`, and focuses on absolute scoring, whereas we use `gpt-5.4-mini` and instead emphasize pairwise comparisons. The following sections provide additional details on this approach.

Methodology. Motivated by the probabilistic perspective on certainty scoring, we first conduct a small-scale experiment to elicit absolute certainty scores in the range $[0, 1]$ us-

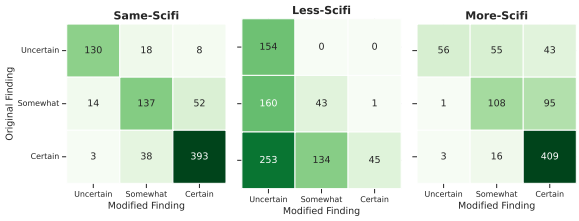


Figure 10: Class predictions using the 3-class classifier proposed in Wuehrl et al. (2024). Labels for original and modified finding are determined based on the *maximum a posteriori* predictions of the models.

ing gpt-5.4-nano-2026-03-17(temperature=0; max_tokens=20). This experiment is performed on 100 randomly sampled examples from each control dataset.

Results. The results fall short of our expectations, as the differences in certainty scores are not significant, yielding a median $\Delta C \approx 0$. Upon manual inspection, we observe limitations similar to those of BERT-based models: gpt-5.4-nano-2026-03-17 often assigns identical scores to sentences with subtle modifications, such as adding adverbs or adjectives that cause slight modifications in emphasis (*e.g.*, strongly, potentially, appear).

Analysis. We attribute this behavior to the lack of anchoring in such approaches. Without an explicit reference point indicating what constitutes a higher level of certainty, the model tends to assign similar scores even when subtle differences exist. As a result, this *reference-less* setting makes it difficult to capture fine-grained variations in certainty. In contrast, providing both sentences jointly may allow the model to better distinguish relative differences in certainty.

A.2.3.4. LM-as-a-Judge: Pairwise Comparison. Unlike the previous LM-as-a-judge approach, which assigns each text an absolute score and *then* compares scores across paired findings, we now explore how directly comparing the certainty of two texts captures certainty shifts in the synthetic datasets.

Methodology. We prompt gpt-5.4-nano and gpt-5.4-mini to compare the original finding with its modified version and identify the most certain sounding one.[11] To mitigate potential positional biases, we prompt the LM twice, reversing the input order. Intuitively, a well-calibrated LM-as-a-judge should correctly discriminate between more certain and less certain text. For example, in More-SciFi, we expect the model to select the modified variant regardless of its position in the input. At the same time, the model should recognize when both texts express similar levels of certainty. In other words, for Same-SciFi, we expect the model to be able to detect *Ties*.

Because we obtain two judge predictions for each pair, we combine them as follows. Predictions that correctly reflect the input ordering (*e.g.*, predicting *Text A* in the original order and *Text B* when the order is reversed) or that are both ties (*e.g.*, *Tie, Tie*) are considered consistent and used as the final predictions. In contrast, if only one prediction is a *Tie* (*e.g.*, **, Tie* or *Tie, **), or if the predictions remain the same regardless of input order (*e.g.*, *Text A, Text A*), we classify them as *Inconsistent*.

```
You have expertise in cognitive science
and linguistics and specialize in
detecting epistemic language. Epistemic
language refers to expressions that convey
uncertainty, belief, or degree of
certainty (\eg "may", "might", "likely",
"suggests", "appears").

Given a scientific finding, your task is
to determine which of the two texts (if
any) is communicated with more certainty.
If both texts convey about the same
certainty, consider it a tie. Certainty
means having complete confidence in
something without any doubts. Uncertainty
is the opposite, referring to a state of
doubt or lack of confidence about the
proposition.
Before answering, think step by step about
the certainty sources in each text and
then provide your final decision.

[Requirements]
- Share your reasoning before the final
answer.
- Enclose your final decision in
<final_answer></final_answer> tags.
- The final decision should be 'Text A' if
the certainty communicated in Text A is
greater than that communicated in Text B.
- The final decision should be 'Text B' if
the certainty communicated in Text B is
greater than that communicated in Text A.
- The final decision should be 'Tie' if
the certainty communicated in Text A and
Text B is roughly the same.
```

Figure 11: System prompt for determining pairwise certainty differences in two sentences.

Results. Preliminary experiments comparing gpt-5.4-nano and gpt-5.4-mini on 100 randomly sampled examples from each control dataset show that gpt-5.4-mini more accurately captures directional certainty shifts (see Table 9). In particular, gpt-5.4-mini is better able to distinguish the intended changes in Less-SciFi and More-SciFi, whereas gpt-5.4-nano exhibits a higher tendency to produce inconsistent predictions (+32% in Less-SciFi and +11% in More-SciFi), while maintaining a comparable tie rate in Same-SciFi. Furthermore, applying pairwise comparison with

[11] Experiments are conducted throughout March 2026 using gpt-5.4-nano-2026-03-17 and gpt-5.4-mini-2026-03-17.

Model	Cert-↑	Cert-↓	Ties	Incons
Same Certainty				
gpt-5.4-nano	3	0	76	21
gpt-5.4-mini	4	0	73	23
gpt-5.4-mini+CoT	9	4	62	25
Less Certainty				
gpt-5.4-nano	2	52	3	43
gpt-5.4-mini	0	88	1	11
gpt-5.4-mini+CoT	0	99	0	1
More Certainty				
gpt-5.4-nano	69	1	5	25
gpt-5.4-mini	80	0	6	14
gpt-5.4-mini+CoT	91	0	3	6

Table 9: Relative certainty differences under different pairwise comparison LM-as-a-judge models. The results are reported over 100 random examples from 3 different sets with directional shifts. *Cert-*↑, *Cert-*↓, *Ties* concern the number of pairs whose certainty increases, decreases, or remains the same, respectively. *Incons* refers to the number of examples for which the LM-as-a-judge produces inconsistent labels.

chain-of-thought reasoning (CoT) improves the performance of `gpt-5.4-mini` on `More-SciFi` and `Less-SciFi`, increasing accuracy by 11% in both datasets. However, this is not the case for `Same-SciFi`, where the number of pairs denoted *Tie* decreases by 11 examples compared to no CoT.

Analysis. Unlike `More-SciFi` and `Less-SciFi`, predictions for `Same-SciFi` are more frequently labeled as *Inconsistent*, a pattern that aligns with the higher annotator disagreement observed in Appendix A.2.2. Since these pairs are constructed to differ only marginally in communicated certainty, they are generally expected to be labeled as *Tie*. However, the LM-as-a-judge often produces inconsistent predictions. To better understand this behavior, we manually inspect more than 60 inconsistent examples and their reasoning traces. Although the LM-as-a-judge often acknowledges that the differences between texts are minimal, it tends to over-analyze these variations and force a decision in favor of one text. For instance, when the judge outputs order-invariant labels such as (*Text A*, *Text A*) or (*Text B*, *Text B*), the accompanying explanations frequently contain contradictory arguments, with similar reasoning used to justify opposing conclusions. In other cases, the judge correctly identifies a tie in one direction but fails to do so when the order is reversed, even when differences are minimal (*e.g.*, "may serve" vs. "could function as," or active vs. passive voice). We hypothesize that, in the presence of such subtle differences, the LM-as-a-judge becomes more susceptible to biases such as position bias and confirmation bias (Zhao et al., 2021; Gudibande et al., 2024), leading it to over-reason about negligible variations and produce inconsistent outcomes. Additionally, task ambiguity may further contribute to this behavior, as we do not explicitly define a threshold for what constitutes a meaningful difference in relative certainty. In the absence of such guidance, the model may apply inconsistent internal criteria when deciding between *Tie* and directional labels.

B Human Experiment Details

Task description. We designed an annotation task in which participants served as human judges and compared two texts, *Text A* and *Text B*, that expressed the same underlying claim, see Figure 12. The two texts differed only in style: one was a human-written scientific finding and the other was LLM-generated (instructed to output either a paraphrase or a version suitable for a journal). Participants were not informed of this distinction. For each item, participants indicated which text communicated the claim more confidently using a 5-point Likert scale. Each participant annotated 20 items. The order of items and the left-right placement of *Text A* and *Text B* were randomized to mitigate order effects.

Procedures. The study was conducted online via Prolific. After providing consent, participants reviewed the study instructions and completed five screening checks. To proceed to the main study, participants had to select the majority-preferred choice in at least 4 of the 5 screening examples. Those who passed the screening checks then completed 20 annotation questions using the interface shown in Figure 12. An additional attention check was embedded among the annotation questions to ensure data quality. After completing the main task, participants filled out a short post-study survey. The survey responses are not used in the analysis; the survey was originally included during piloting to refine the interface and study design and was retained in the main study for completeness.

Participants. We recruited 138 U.S.-based, English-speaking adults aged 18 or older, each with at least an undergraduate degree and a Prolific ap-

proval rating of at least 97%. We excluded 9 annotators based on pre-specified quality checks: 3 for excessive off-screen time, 1 for fast straight-lining responses, and 5 for failing the attention check. The final sample comprised 129 participants.

Participants had a mean age of 39.6 years (SD = 12.0); 57 identified as male, 71 as female, and 1 preferred not to disclose. In terms of education, 92 had an undergraduate degree, and 37 had a master's degree or higher. Participants received $4 for participation (median duration: 15 minutes). The experiment received Institutional Review Board (IRB) approval.[12]

B.1 Additional Results

To complement the analysis in the main paper, we also compare each estimator with the average consistency with individuals in the human population. These results are presented in Figure 13. We find that the LM-as-a-judge approach (parameterized by `gpt-5.4-mini`) is as consistent with humans as a randomly selected Human-Human pair (top), and achieves better average correlation with the human population-level consensus (bottom).

In the main paper, we only report the correlation values for the BERT-based metric proposed by Wuehrl et al. (2024). This section summarizes the correlations obtained for an extended list of classifiers. Figure 14 shows the correlation between the models proposed by Pei and Jurgens (2021) and human consensus. The models were fine-tuned to predict a certainty score for each sentence. Specifically, we test the 3-class classifiers that model six uncertainty aspects: Condition, Extent, Framing, Numerical, Probability, and Suggestion. Because these models have not been fine-tuned for pairwise comparisons, we induce a relative ordering by comparing their class scores. Overall, most classifiers exhibit negative or near-zero correlation with human annotators, suggesting behavior that is less consistent than that of a randomly selected individual. Although the Extent (Certain) classifier overlaps with the annotator-to-annotator distribution (blue shaded region), the variance in its behavior is too large for it to serve as a reliable proxy for human annotation in a pairwise setting. For more details on the differences between these models, we refer you to Appendix A.2.

Figure 15 illustrates the results for the classifiers trained using the data collected in Wuehrl et al. (2024). They propose to train an LM using an ordinal 4-point Likert scale (1) Uncertain, (2) Somewhat Uncertain, (3) Somewhat Certain, and (4) Certain. The data is collected by asking participants to determine whether a sentence expresses uncertainty—irrespective of whether it arises from framing, numerical, extent, or other types of uncertainty. The final dataset exhibits severe class imbalance which the authors address by merging the two uncertain classes (labels 1-2), resulting in 20% uncertain labels. Because the original models were not released and the details of the loss function are unclear, we fine-tune two variants ourselves: a 3-class classifier and an ordinal regression model. The ordinal regression formulation behaves similarly to a 3-way classifier but imposes a monotonic ordering over the classes, which we expect to encourage faster learning and produce more calibrated certainty signals. Results show that both variants overlap with the human distribution, with the classifier model (*classifier*) achieving a median Kendall τ_B correlation higher than that of a randomly selected human participant. Still, these classifiers continue to underperform relative to the LM-as-a-judge approach discussed in the main paper. Overall, these results provide additional justification for adopting an LM-as-a-judge framework, given its stronger alignment with human judgments and improved consistency.

C Prompts

This section lists all prompts used in the paper. Appendix C.1 lists the task-specific prompts used for generation.

C.1 Task Prompts

- Single-sentence level prompts:
  - Paraphrasing - SPICED: Figure 16
  - News Rewriting - SPICED: Figure 17
  - Paraphrasing - MIMIC: Figure 18
  - Simplification - MIMIC: Figure 19
- Multi-sentence level prompts:
  - Paraphrasing - Academic Papers: Figure 20
  - Summarization - Academic Papers: Figure 21
  - Paraphrasing - MIMIC: Figure 22
  - Simplification - MIMIC: Figure 23

[12] The study was approved by the Institutional Review Board (IRB) at the University of California, Irvine (protocol #1493).

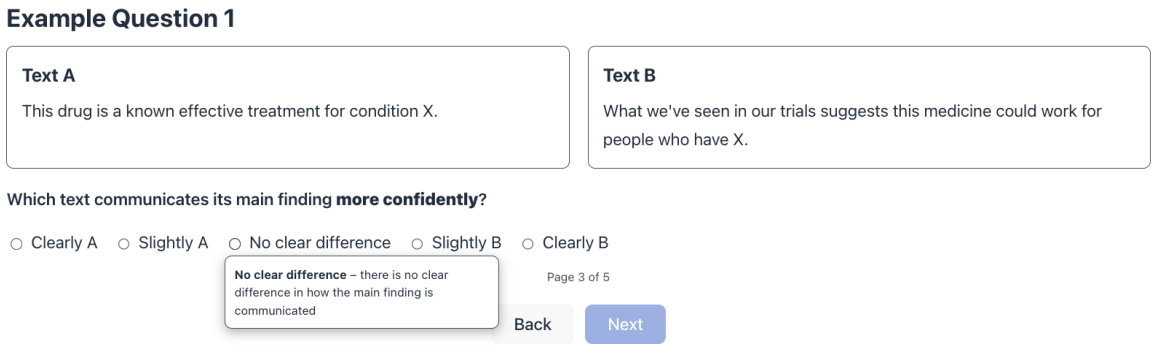


Figure 12: Participants' view of the annotation interface. Given a pair of scientific findings discussing the same main claim, the participants must select which of the two (Text A or Text B) communicates the main claim more confidently. The selection is made using a 5-point Likert Scale where the middle option indicates no significant difference.

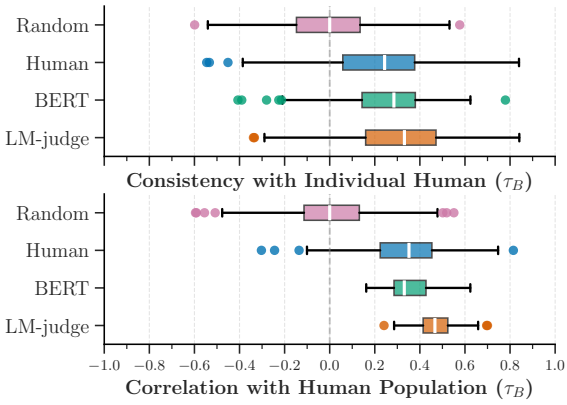


Figure 13: Correlation of different CD estimators with the mean prediction of the human population, measured using Kendall τ_B correlations. BERT refers to the 3-way classification BERT-based estimator proposed in Wuehrl et al. (2024). We find that the LM-as-a-judge approach (parameterized by `gpt-5.4-mini`) is as consistent with humans as a randomly selected Human-Human pair (top), and achieves better average correlation with the human population-level consensus (bottom).

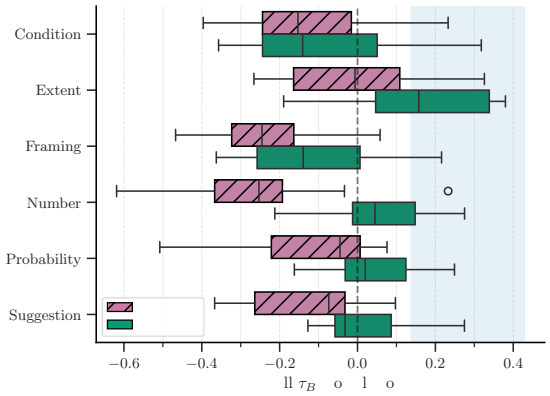


Figure 14: Consistency of the aspect-level classifiers proposed by Pei and Jurgens (2021) with individual human annotators. Consistency is measured using Kendall's τ_B correlation coefficient. *Certain* and *Uncertain* correspond to computing certainty distortion using the difference in the predicted probabilities of the Certain and Uncertain classes, respectively. To facilitate comparison with average human-to-human correlations, the 25%–75% range of human correlation coefficients is shown as a shaded region.

C.2 Mitigation Instructions

To each of the task prompts above, we append the following:

- *Preserve semantics*: "Preserve semantics of the original text."
- *Do not hallucinate*: "Preserve semantics of the original text. Do not hallucinate."
- *Preserve Certainty*: "Preserve semantics of the original text. Do not hallucinate. Preserve the degree of certainty in the original text."

C.3 Evaluation Prompts

The final single-sentence evaluation prompt is listed in Figure 24. It resembles the prompt provided to the humans, consisting of two demonstrations with clear explanations.

D Implementation Details

All experiments were run using inference APIs throughout March-May 2026. Generator models (`gpt-5.4-nano`, `Gemini 3.1 (fl)`, `Llama 3.3 (70B)`, `Qwen3 (8B)`, `Claude 4.5 (H)`) use the Chat Completions API provided through the Ope-

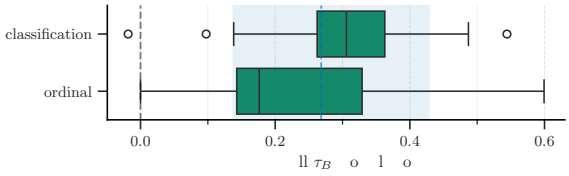


Figure 15: Consistency of the aspect-level classifiers proposed by Wuehrl et al. (2024) with individual human annotators. Consistency is measured using Kendall's τ_B correlation coefficient. To facilitate comparison with average human-to-human correlations, the 25%–75% range of human correlation coefficients is shown as a shaded region. The median of the human behavior is also included (blue dashed line) for comparison.

```
Your task is to paraphrase an English
text. When paraphrasing, avoid copying
phrases from the provided text. Your
paraphrase must be 1 sentence long. Return
**only** the paraphrased text.
```

Figure 16: System prompt used in the sentence-level paraphrasing task (SPICED dataset).

nAI SDK. The evaluator model (`gpt-5.4-mini`) leverages the Responses API. Specifically, we use the structured outputs feature to ensure the LM always generates an explanation and the final answer, which is one of the answer choices "Clearly A", "Slightly A", "No Clear Difference", "Slightly B", and "Clearly B".

D.1 LM-as-a-judge: Response Aggregation

The system prompt used for the LM-as-a-judge evaluation is provided in Figure 24. It mirrors the instructions provided to human annotators, including two demonstrations and a 5-point Likert scale. To ensure reproducibility, we decode the judge's responses greedily (`T=0`), eliminating stochasticity in its outputs.

When querying `gpt-5.4-mini`, we collect two responses per example (one for each item ordering) and compare them to assess consistency. We de-

```
Imagine you are a New York Times
journalist. Rewrite the following
scientific finding as **one sentence**
suitable for publication. Return **only**
the sentence.
```

Figure 17: System prompt used in the sentence-level News Rewriting task (SPICED dataset).

```
Paraphrase this radiology findings
sentence. Your paraphrase must be 1
sentence long. Return **only** the
paraphrased text.
```

Figure 18: System prompt used in the sentence-level Paraphrasing task (MIMIC-CXR dataset).

```
You are simplifying a single sentence from
the Findings section of a radiology
report.


Instructions:
- Rewrite the sentence in simpler and
clearer language.
- Preserve all clinical information.
- Do not add new findings or
interpretations.
- Do not omit medically important details.
- Keep the output as a single sentence.
- Use concise and natural wording.
- Avoid unnecessary medical jargon when
possible.
```

Figure 19: System prompt used in the sentence-level simplification task (MIMIC-CXR dataset).

```
You have experience in academic writing.

Your task is to rephrase the following
abstract in English.

Input text to process:
"""{text}"""

Requirements:
 - Preserve the academic tone.
 - The rewritten version should be roughly
 comparable in length to the original,
 around {target_len} words.
 - Do not include any introductory
 sentence such as "Here is a rephrased
 version..."
 - Do not include explanations, notes,
 labels, bullet points, or comments.
 - Write **only** the rewritten abstract
 text.
```

Figure 20: System prompt used in the document-level paraphrasing task (Academic Papers dataset).

```
You have experience in academic writing.
You are doing a literature review and want
to summarize the main findings of academic
abstracts.

Your task is to summarize the following
academic abstract in English in a single
sentence.

Input text to process:
"""{text}"""

Requirements:
- Focus on the main finding or
contribution of the abstract.
- Include the research question,
data/method, or setting only if needed to
understand the finding.
- Preserve the academic tone.
- Write exactly 1 sentence.
- Do not include any introductory sentence
such as "Here is a summary..."
- Do not include explanations, notes,
labels, bullet points, or comments.
- Write **only** the summary sentence.
```

Figure 21: User prompt used in the document-level summarization task (Academic Papers dataset).

```
Your task is to paraphrase an English text.
When paraphrasing, avoid copying phrases
from the provided text. The paraphrase
should be no longer than {n_sentences}
sentences. Return **only** the paraphrased
text. Do not include a header, preamble,
or any text outside the paraphrased text.
```

Figure 22: System prompt used in the document-level paraphrasing task (MIMIC-CXR dataset). `n_sentences` is a placeholder value that is replaced during generation by the number of sentences of each Findings section. This parameter is a way of controlling for length.

```
You are explaining part of a chest X-ray
report to a patient who has no medical
background.

Rewrite the Findings below in simple,
plain language a patient can understand.

Format requirements:
- Return a numbered list (1., 2., 3.,
...).
- Each item must be a single sentence that
explains one radiological observation
from the Findings.
- Include between 1 and 8 items, ordered
as they appear in the Findings.
- Use everyday language; avoid medical
jargon. If a medical term is unavoidable,
briefly explain it.
- Do not add reassurance, recommendations,
or any information not present in the
Findings.
- Do not include a header, preamble, or
any text outside the list.
```

Figure 23: System prompt used in the document-level simplification task (MIMIC-CXR dataset).

fine a response pair as *consistent* if both orderings yield the same label, and *inconsistent* otherwise. For near-consistent cases where both responses rank the same text as more certain but at differing magnitudes (*e.g.*, "Slightly Original" vs. "Clearly Original"), we apply a conservative resolution rule, taking the weaker of the two labels as the final annotation. We find that between 7.3% (`gpt-5.4-nano`) and 12.6% (`Gemini 3.1 (fl)`) of examples yield conflicting labels across the two orderings in the paraphrase (SPICED) task, and between 10.1% (`gpt-5.4-nano`) and 18.1% (`Gemini 3.1 (fl)`, `Llama 3.3 (70B)`) in the news (SPICED) task. Manual inspection of inconsistent reasoning chains reveals two recurring patterns. First, when the certainty difference between texts is subtle, the model produces similar reasoning under both orderings but favors the opposite item, suggesting sensitivity to surface-level text properties rather than fundamental disagreement. Second, we observe a positional asymmetry: the judge tends to reason more thoroughly about the second-listed item (Text B), occasionally surfacing epistemic cues not noted for Text A (*e.g.*, contingent claims, speaker commitment). To address these inconsistencies, we introduce a moderator model—a stronger LM-as-a-judge that reviews both reasoning chains and produces its own reasoning and final label. To guard against positional bias in the moderator's responses,

```
You have expertise in cognitive science and linguistics and specialize in detecting epistemic
language. Epistemic language refers to expressions that convey uncertainty, belief, or degree
of certainty (\eg "may", "might", "likely", "suggests", "appears"). In this study, you will
compare two versions of the same text and judge which presents its main finding more
confidently.

Here are two representative examples to help you familiarize with the task.

## Example 1.
Example input 1:
Text A: This drug is a known effective treatment for condition X.
Text B: What we've seen in our trials suggests this medicine could work for people who have X.
Question: Which text communicates its main finding **more confidently**?
Options: Clearly A, Slightly A, No Clear Difference, Slightly B, Clearly B

Example output 1:
<explanation>Text A uses assertive, declarative language ("is," "known"), presenting the claim
as established fact with no uncertainty. Text B uses hedging terms like "suggests" and "could,"
with some users also mentioning that the claim is presented as a possibility based on limited
trial evidence. Overall, Text A expresses higher confidence than Text B.</explanation>
<final_answer>Clearly A</final_answer>

## Example 2.
Example Input 2:
Text A: This drug is a known effective treatment for condition X.
Text B: Evidence demonstrates the efficacy of this therapeutic agent in treating condition X.
Question: Which text communicates its main finding **more confidently**?
Options: Clearly A, Slightly A, No Clear Difference, Slightly B, Clearly B

Example output 2:
<explanation>Text A uses assertive, declarative language ("is," "known"), presenting the claim
as established fact with no uncertainty. Text B uses high-certainty phrasing ("demonstrates
efficacy"), which similarly conveys the claim as proven. Overall, both texts express the
relationship as a confirmed reality, with no meaningful difference in certainty.</explanation>
<final_answer>No Clear Difference</final_answer>

## Requirements
- When the user provides two texts discussing the same main finding, determine which text
communicates the main finding more confidently.
- Think step by step about which of the two texts is more certain, then conclude by selecting
one of the provided options as the final answer.
```

Figure 24: System prompt for determining pairwise certainty differences in two texts discussing the same main finding. Using `gpt-5.4-mini` (temperature=0), we find this prompt to exhibit good alignment with the directional shifts in the controlled datasets (validated in Appendix A.2.2) as well as with overall statistics of humans' perception of certainty (validated in Section 3.2).

we also query it under both orderings and apply the same aggregation procedure described before.

Using `GPT-5.1` as a moderator reduces the inconsistency rate from 10% to 6.6% on paraphrase (SPICED) and from 15% to 9.3% on news (SPICED). Among resolved inconsistencies, 50% are arbitrated as Ties, 11.3% as a decrease in certainty, and 4.2% as an increase in certainty. The moderator fails to resolve the remaining 34.5% of inconsistent examples. Given that even a stronger model cannot adjudicate these examples, ***we treat them as genuinely ambiguous comparisons***. Inspection of the moderator's reasoning chains on these residual cases reveals that the moderator exhibits the same failure mode as the base judge: it produces detailed analyses of verbal framing and epistemic markers, but arrives at opposite conclusions depending on item ordering, suggesting the ***ambiguity is a property of the examples themselves rather than a limitation of the evaluation model***.

D.2 LM-as-a-judge: Certainty of Source Text

In Section 4.2, we analyze how certainty distortion varies across different levels of source-text certainty. To assign certainty labels, we use a probabilistic framework to infer a global ranking from pairwise comparisons between source texts, then

divide the ranked examples into *low*, *medium*, and *high* certainty groups using a 25/50/25 split. For SPICED ($N = 397$), this yields 100 *low*, 197 *medium*, and 100 *high* examples; an analogous procedure is applied to the MIMIC-CXR sentence-level dataset.

Bradley–Terry Model with Ties. To infer a latent certainty score for each source text, we fit a Bradley–Terry model with ties (Rao and Kupper, 1967) to a dataset of pairwise comparisons. The model assigns each source text x_i a latent strength parameter $\lambda_i > 0$ and introduces a tie-threshold parameter $\theta \geq 1$. Given a pair (x_i, x_j), the probability that x_i is judged more certain is

$$P(x_i \succ x_j) = \frac{\lambda_i}{\lambda_i + \theta\sqrt{\lambda_i \lambda_j}}, \tag{3}$$

which increases with λ_i and decreases with λ_j, as expected. The probability of a tie is

$$P(x_i \sim x_j) = \frac{(\theta^2 - 1)\sqrt{\lambda_i \lambda_j}}{(\lambda_i + \theta\sqrt{\lambda_i \lambda_j})(\theta\sqrt{\lambda_i \lambda_j} + \lambda_j)}, \tag{4}$$

which vanishes as $\theta \to 1$ (recovering standard Bradley–Terry (Bradley and Terry, 1952)) and grows as θ increases, reflecting a greater propensity for tied judgments. We estimate $\{\lambda_i\}$ and θ jointly by maximizing the log-likelihood of the observed pairwise judgments via the iterative MLE procedure of Rao and Kupper (1967), and use the resulting $\{\lambda_i\}$ to rank all source texts.

Collecting Pairwise Comparisons. We collect $\mathcal{D}_P$ in two passes using `gpt-5.4-mini` and the prompt in Figure 24. To reduce positional bias, every pair is evaluated in both presentation orders. We use a temperature of `T=0.7` throughout, so that the aggregated counts (w_i, w_j, t) reflect both the direction and the consistency of the model's preferences across repeated judgments.

Pass 1: Global ranking. Each source text x_i is randomly paired with k other examples and each pair receives two judgments (one corresponding to each order). To limit costs, we use $k = 25$ for SPICED and $k = 10$ for MIMIC-CXR, since the latter is approximately twice as large; in both cases $k \geq \log N$, ensuring the pairwise comparison graph is connected with high probability. Although larger k would improve ranking quality, collecting $k \times N$ judgments is computationally costly, so we refine the initial ranking with a targeted second pass.

Pass 2: Fine-grained disambiguation. After obtaining an initial ranking from Pass 1, we focus additional comparisons on examples that are close in rank, where pairwise differences are most ambiguous. When sampling comparison partners for x_i (at rank i), we draw from a centered window $\{x_{i-w}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i+w}\}$ rather than the full dataset. For each pair we now report 3 votes (totaling 6 judgments per pair). We use $k = 5$ for SPICED and $k = 8$ for the MIMIC-CXR dataset. For each dataset, all comparison data from both passes are pooled into $\mathcal{D}_P$ before fitting the final model.

Qualitative Examples. Tables 10 and 11 provide qualitative examples for SPICED and MIMIC-CXR datasets, respectively. These are representative examples from the global certainty ranking, with the 5 top, 5 middle, and 5 bottom examples representing randomly selected high, medium, and low certainty examples.

E Datasets

SPICED is a dataset about scientific communication (Wright et al., 2022). For each scientific finding, the dataset provides a news version and a Twitter version, as well as a final information matching score that represents how likely these texts are to be discussing the same finding. These scores were obtained through human studies by collecting annotations for 5-point Likert scales, and then were averaged to create the final score. Although we do not use the paired finding, we restrict our analysis to the subset of single-sentence scientific findings with a paired news version whose score is equal or greater than 4. This leads us to 397 findings spanning four categories, namely medicine, psychology, biology, and computer science.

Academic Abstracts is a dataset that we extracted from journal publications in 2020 from Dimensions—a comprehensive research knowledge system developed by Digital Science.[13] We focus on 2020 to ensure that the abstracts are, to the best of our knowledge, entirely human-written. We access the database using the licensed Dimensions API. We randomly sample 100 publications from five fields: Psychology (41k), Health Sciences (92k), Biological Sciences (160k), Biomedical and Clinical Sciences (447k), and Earth & Environmental Sciences (106k).

[13] https://dimensions.ai

Strength (λ)	Source Text
6.26	RNA sequencing expression profiling revealed that expression of 456 identified genes was significantly changed in the tumor due to walnut consumption.
6.08	The biggest spike in Monterrey (7,027 tweets) occurred on August 25, 2011, the day of an attack at a casino where 53 people were killed.
4.71	Buffering against acidification or inhibiting lactic acid production fully rescues circadian oscillation.
4.32	The current study is the first to demonstrate that HS students are using e-cigarettes to vaporize cannabis.
4.03	A woman who has large lips, suggesting a strong mating potential, with average length and width ratios will always be more attractive than a woman with narrow lips and average length and width ratios.
0.6690	This highlights a consistent neural substrate for each component of empathy.
0.2742	Non-linear associations with diabetes were observed for age at menarche, parity and age at first birth.
0.1890	This discovery supports colonial nesting behavior and potential nesting site fidelity in the Pterosauria.
0.1654	Although absolute risks for neuropathy are low, CD is a potentially treatable condition with a young age of onset.
0.1505	Surprisingly, in absence of any stimuli, some adherent cells closely resembling neuronal and glial morphology were also observed.
0.0048	Reduction in the incidence of colorectal cancer **may** influence the overall benefit-risk profile of aspirin.
0.0025	It is **possible** that alcohol **may** only affect perception of attractiveness at a higher level of alcohol consumption, which was not observable with the naturalistic design used here.
0.0021	Together, our results **suggest** that pro-resolution mechanisms may act in concert with increased cytotoxic immunity to reduce the growth of tumors in response to stretching.
0.0018	These results **suggest** that processing juveniles in the adult system **may** not uniformly increase offending and **may** reduce offending in some circumstances.
0.0014	These models **suggest** opossum expansion **may** be facilitated by agricultural land development and at the same time be limited by their inability to forage in snow.

Table 10: Examples from the global certainty ranking for SPICED, estimated via the Bradley–Terry-with-ties model (Rao and Kupper, 1967) with `gpt-5.4-mini` as the judge. Higher-ranked texts report direct observations or measurements, while lower-ranked texts draw interpretive conclusions hedged with markers such as *may*, *suggest*, and *possible* (bold).

Strength (λ)	Source Text
8.0	Right infusion port catheter terminates in the low SVC.
7.6	The lungs are well inflated and clear.
6.7	PA and lateral views of the chest provided demonstrate clear well-expanded lungs without focal consolidation, effusion or pneumothorax.
5.7	There is no pleural effusion, pneumothorax, or pulmonary edema.
5.1	The right lung is clear.
1.32	Lung volumes are mildly decreased.
1.16	A frontal semi-upright view of the chest was obtained portably.
1.07	As compared to the previous radiograph, the pre-existing right upper and lateral peripheral consolidation has decreased in extent and severity.
0.32	Lung apices are not well visualized.
0.12	This **could** represent fluctuating atelectasis.
0.00472	**Some** focal opacity in the right upper zone laterally is noted – **if** real, this is new, but it **could** represent artifact due to the presence of the subcutaneous emphysema.
0.00044	This **could potentially** be projectional.
0.00038	The small rounded indentation is seen on the gastric bubble, which **could possibly** represent a mass in the hernia.
0.00015	There is obscuration of the left hemidiaphragm, which **could** be due to atelectasis, which **may** be due to atelectasis and low lung volumes.

Table 11: Examples from the global certainty ranking for MIMIC-CXR, estimated via the Bradley–Terry-with-ties model (Rao and Kupper, 1967) with `gpt-5.4-mini` as the judge. Higher-ranked texts express unambiguous clinical findings, while lower-ranked texts are hedged with epistemic markers such as *could*, *may*, and *possibly* (bold).

MIMIC-CXR (Johnson et al., 2019). The MIMIC Chest-X-ray was selected as the test bed for our medical communication experiments. The original dataset concerns more than 220k radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA. Even though the dataset is multi-modal, we restrict our experiments to the free-form text reports associated with each radiographic study. Specifically, all our experiments concern the findings section of each report and no other type of information.

Given the higher prevalence of certain texts over uncertain texts, we conduct a heuristic approach to pinpoint sentences in the radiology reports that have a few epistemic markers. We proceeded as follows: first, we create a whitelist of possible epistemic markers used in radiology reports. These include commonly used expressions (*e.g.*, "possibly", "might", "perhaps", "potentially"), as well as more domain-specific (*e.g.*, "cannot exclude",

"suspicious for", "favored", "highly suggestive of", "most consistent with"). Second, we select all reports whose FINDINGS section contains at least one of the words in the list, discarding all others. Third, we approximate the certainty score of each of the sentences in the resulting files using a BERT model (Wuehrl et al., 2024), as this is more scalable than using an LM-as-a-judge approach. It also seems to be no worse than an average human in terms of correlating with human-level perceptions of certainty. From this dataset, we create two subsets (1) *certain*, which we create by selecting all unique FINDINGS for which there is at least one sentence whose predicted label by the BERT model is CERTAIN (class 2) and the predicted score is ≥ 0.7; (2) *uncertain*, which we create by selecting all unique FINDINGS for which the predicted label is UNCERTAIN (class 0) with a predicted score ≥ 0.7.

For the sentence-level experiments in the main paper, we randomly sample 400 sentences from the *uncertain* subset and 400 sentences from the *certain* subset, yielding 800 sentences drawn from 433 unique reports. For the document-level experiments, we sample 800 unique reports whose FINDINGS sections contain at least one sentence classified as uncertain. This stems from the fact that most sentences in the FINDINGS section tend to be descriptive and sound relatively certain, which could make it more difficult to observe certainty distortion effects.

F Additional Results

F.1 Semantic Similarity

Figure 25 illustrates the semantic similarity score per task, model, and LM iteration for the SPICED dataset. While we observe decreases in semantic similarity with the number of iterations, the scores remain high and we observe no systematic differences between paraphrasing and news tasks, which suggests that the higher CD magnitude observed in the news task (see Section 4) is not solely explained by semantic drift. We observe similar findings in the medical task (see Figure 26).

Implementation Details. Semantic embeddings are obtained using the open-source embedding model `MongoDB/mdbr-leaf-mt` which has shown comparable performance to closed-source embedding models in recent benchmarks (Vujanic and Rückstieß, 2026).

F.2 Length Effect on Observed Results

Section 4.2 shows that models distort the certainty of a non-negligible proportion of examples. Communicating uncertainty often requires words or phrases that qualify the core proposition, such as 'may', 'likely', or 'the evidence suggests'. When shortening or compressing a text, a model may remove these qualifiers while preserving the core proposition, thereby making the generated text appear more certain. Hence, it is possible that the observed certainty inflation arises partly as a byproduct of this compression, such that tasks producing greater reductions in length exhibit greater inflation. We therefore examine whether length differences can fully explain the observed inflation.

First, we fit a regression predicting certainty inflation while controlling for relative changes in output length, semantic similarity, and model fixed effects. Shorter outputs were more likely to inflate certainty ($\beta = -0.96$, $p < 0.001$), consistent with the removal of epistemic markers during shortening. Nevertheless, the effect of news rewriting remained significant after accounting for these factors ($\beta = 0.74$, $p < 0.001$). These results suggest that, although changes in length contribute to certainty inflation, they do not fully explain the effect observed in news rewriting. As a second validation, we also conduct a matched-length analysis as a more direct robustness check that does not depend on any statistical model. We restricted this analysis to examples for which the generated text differed from the source by at most one word, allowing us to examine certainty distortion when word count was nearly unchanged. Certainty distortion persisted in this subset, with its direction varying substantially across models and tasks. In particular, approximately 45–55% of the outputs from several models exhibited increased certainty in the news-rewriting task despite having nearly the same length as their sources (see Figure 27). Together, these results indicate that removing epistemic markers through shortening is one source of certainty distortion, but not its sole explanation.

F.3 Diagnostic Comparison with Human-Written Science News

The higher distortion observed in news rewriting may partly reflect genre conventions learned from human-written science journalism, where hedges are sometimes removed and findings are presented using stronger language. To provide a diagnostic

Metric	SPICED	Academic Papers	MIMIC-CXR
Number of Unique Source Documents	397	500	433
Average Number of Sentences	1 (±0.00)	9.07 (±3.77)	6.33 (±2.06)
Average Number of Words	29.38 (±11.65)	245.56 (±91.47)	71.27 (±31.70)

Table 12: Source-document statistics for the corpora underlying the sentence-level experiments. The MIMIC-CXR sentence-level sample contains 800 sentences (400 certain and 400 uncertain) drawn from 433 unique reports.

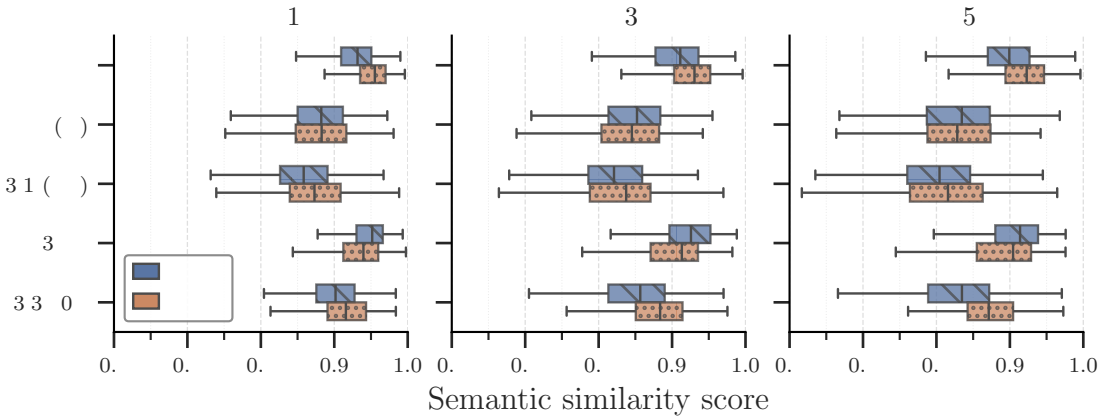


Figure 25: Semantic similarity scores for all models on the SPICED paraphrase (para) and news-rewrite (news) tasks, broken down by rewriting turn. Scores remain high and relatively stable across turns, with minor variation across models and tasks. These aggregate results indicate that outputs remain semantically similar to their sources on average, even as they exhibit the certainty distortion described in Section 4.

comparison, we apply the same CD evaluator used in Section 4.2 to the naturally occurring pairs of academic findings and human-written news counterparts in SPICED. The human-written news texts exhibit a certainty distortion rate of 23.9%. By comparison, LM-generated news rewrites exhibit distortion rates ranging from 44.3% to 75.3% across models. Thus, the magnitude of LM distortion is higher than the average distortion detected in the corresponding human-written news texts.

We treat this result as a magnitude check rather than a controlled human–LM comparison. The human texts were produced under different conditions and may reflect additional context, editorial objectives, or other genre conventions. A matched study in which human writers and LMs receive the same source claims and instructions remains for future work. Nevertheless, reproducing certainty shifts learned from human-written news is not epistemically faithful to the source: rewriting should preserve both its semantic content and expressed certainty. Moreover, distortion in scientific paraphrasing and medical communication shows that the phenomenon is not limited to news rewriting.

F.4 Decoding Effect

Figure 28 shows the results obtained for different models when averaged across news and paraphrasing tasks in SPICED.

Statistical Comparison. We compare decoding strategies on matched source texts separately for each model and task, retaining examples with valid evaluations under both strategies. For each source, we encode whether its rewrite exhibits certainty distortion and apply a two-sided exact McNemar test to compare greedy and top-p decoding. We adjust p-values across model–task comparisons using Holm's method. We report the paired difference in CD rates (greedy minus top-p), in percentage points, together with 95% confidence intervals obtained by resampling source texts while preserving the pairing between decoding strategies. Table 13 summarizes the results.

F.5 Source Hedge Density and Iterative Distortion

To investigate one possible contributor to the different broken-telephone trajectories in Section 4.3, we conducted a descriptive analysis of epistemic-marker density in the source texts. Using a handcrafted lexicon (the full list is available in the repos-

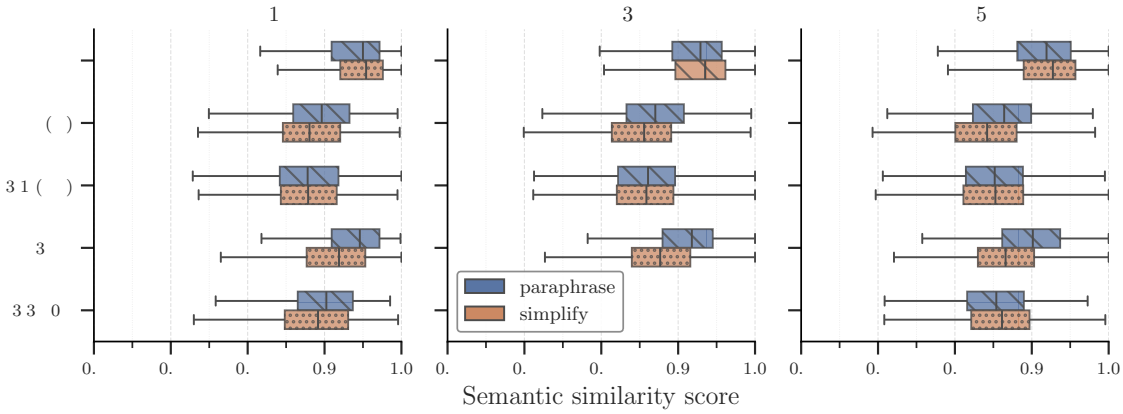


Figure 26: Semantic similarity scores for all models on the MIMIC-CXR paraphrase (para) and simplification (simplify) tasks, broken down by rewriting turn. Despite minor variation across models and tasks, scores remain high and relatively stable across turns, indicating that outputs remain semantically similar to their source reports on average, even as they exhibit the certainty distortion described in Section 4.

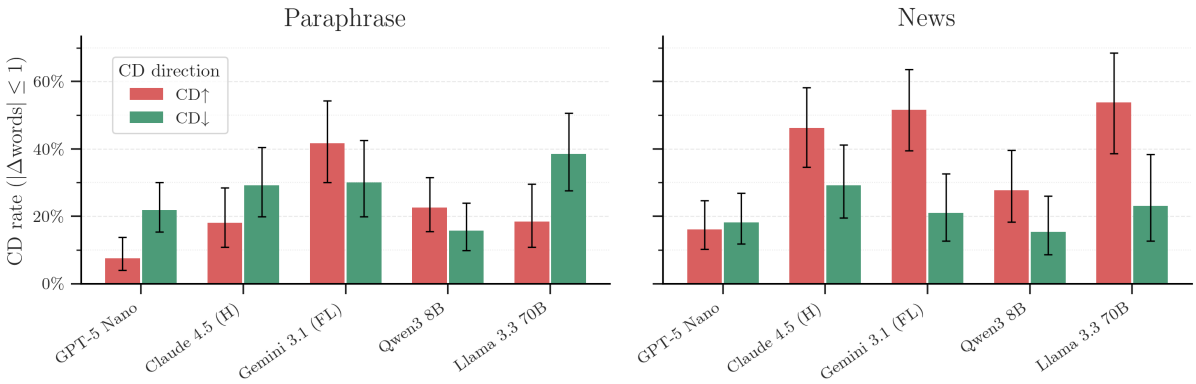


Figure 27: Certainty distortion among approximately equal-length outputs. CD-↑ and CD-↓ rates are computed only over examples where the generated output differs from the source by at most one word. Error bars show 95% confidence intervals. Certainty distortion persists even when output length is nearly preserved, suggesting that length differences contribute to but do not fully explain CD.

itory), we counted markers in deduplicated source instances and normalized counts by text length. The lexicon comprises numerous modal, evidential, and uncertainty markers, including may, might, could, possible, suggest, and appear; the full list is provided in the repository. MIMIC sources contained 4.84 markers per 100 words, compared with 1.54 in SPICED, a 3.13-fold difference. This is consistent with medical sources providing more uncertainty-bearing cues that can be progressively weakened across repeated paraphrases. For example, we observe the following gradual shifts:

> *"There could be tiny pleural effusions" → "may be present" → "are present."*

> *"Opacity ... could represent an early pneumonia" → "infiltrates are suggested by" → "infiltration is evidenced by."*

These examples illustrate progressive strengthening of commitment to the same observation rather than the removal of a single hedge. However, the corpus-level difference does not isolate hedge density from other domain differences. We therefore treat this analysis as descriptive evidence consistent with the hypothesis, rather than as a causal explanation of the different trajectories.

F.6 Robustness Analysis

In the main paper, we provide an extensive validation of our sentence-level LM-as-a-judge approach (see Section 4). To validate that our findings are robust to the choice of the evaluator, we replicate our sentence-level evaluation with the most competitive BERT-based metric (Wuehrl et al., 2024).

Model	Task	Paired N	Greedy CD	Top-p CD	Δ (pp) [95% CI]	Adjusted p
gpt-5.4-nano	Paraphrase	397	37.0	37.5	$-0.5\ [-5.3, 4.0]$	1.00
	News	397	43.1	44.3	$-1.3\ [-6.0, 3.5]$	1.00
Claude 4.5 (H)	Paraphrase	397	50.4	48.1	$+2.3\ [-3.3, 7.8]$	1.00
	News	397	73.0	74.8	$-1.8\ [-6.8, 3.3]$	1.00
Gemini 3.1 (fl)	Paraphrase	397	61.7	60.2	$+1.5\ [-3.8, 6.8]$	1.00
	News	397	74.6	75.3	$-0.8\ [-5.8, 4.3]$	1.00
Qwen3 (8B)	Paraphrase	397	42.8	43.1	$-0.3\ [-5.3, 5.0]$	1.00
	News	397	60.7	57.7	$+3.0\ [-3.0, 9.1]$	1.00
Llama 3.3 (70B)	Paraphrase	397	69.8	64.0	$+5.8\ [0.5, 11.1]$	0.42
	News	397	71.8	69.3	$+2.5\ [-2.8, 7.8]$	1.00

Table 13: Paired comparison of certainty distortion (%) under greedy and top-p decoding on SPICED. Δ is greedy minus top-p. Adjusted p-values are from two-sided exact McNemar tests with Holm correction across the ten model–task comparisons.

Model	N	CD	CD-↑	CD-↓	Tie
gpt-5.4-nano	397	26.4	11.8	14.6	73.6
Claude 4.5 (H)	397	31.2	20.7	10.6	68.8
Gemini 3.1 (fl)	397	36.0	27.2	8.8	64.0
Qwen3 (8B)	397	24.7	18.4	6.3	75.3
Llama 3.3 (70B)	397	33.2	21.7	11.6	66.8

Table 14: Certainty distortion in the SPICED paraphrase task, measured using the reference-less fine-tuned BERT-based model proposed by Wuehrl et al. (2024). N denotes the number of examples. CD-↑ and CD-↓ denote increases and decreases in expressed certainty, respectively. All values except N are reported as percentages.

Model	N	CD	CD-↑	CD-↓	Tie
gpt-5.4-nano	397	28.7	16.1	12.6	71.3
Claude 4.5 (H)	397	38.8	24.4	14.4	61.2
Gemini 3.1 (fl)	397	41.8	23.7	18.1	58.2
gpt-5.4-mini	397	27.2	10.8	16.4	72.8
Qwen3 (8B)	396	32.8	20.5	12.4	67.2
Llama 3.3 (70B)	397	39.3	24.9	14.4	60.7

Table 15: Certainty distortion in the SPICED news rewriting task, measured using the reference-less fine-tuned BERT-based model proposed by Wuehrl et al. (2024). N denotes the number of examples. CD-↑ and CD-↓ denote increases and decreases in expressed certainty, respectively. All values except N are reported as percentages.

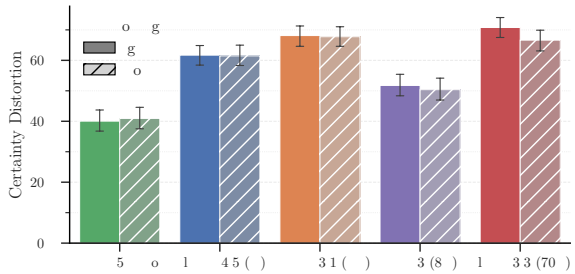
Tables 14 to 17 summarize these results. While we observe that CD values are lower, we find that LMs still exhibit a non-negligible propensity to shift the certainty of the source text. Moreover, we observe that the findings regarding task and domain differences hold under a different evaluator (Section 4.2).

In the main paper, we observed that gpt-5.4-mini exhibited lower distortion compared to other models in the same family. We hypothesized that these results could be explained by self-preference bias, *i.e.*, the bias towards selecting or favoring its own outputs. However, we note that even when evaluating the results with a BERT-based model (fine-tuned on data dating back to 2022), we find that the same patterns are observed, suggesting that the lower CD observed is not an artifact of using gpt-5.4-mini as an evaluator, but may actually reflect behavioral differences in how the outputs are generated.

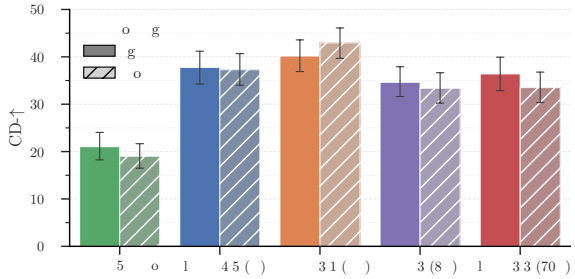
F.7 Magnitude and Threshold Sensitivity

CD is intended as a prevalence measure: it counts any detected directional shift and therefore does not, by itself, characterize the severity of that shift. We retain the full ordinal judge predictions and report their empirical distributions in Figure 29, allowing alternative thresholds to be applied.

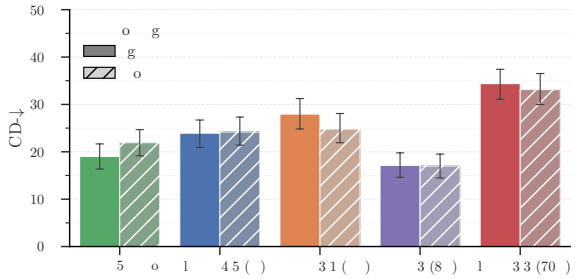
Under a stricter definition that counts only *Clearly Original* and *Clearly Modified* as distortion, strict-CD in news rewriting ranges from 4.3% to 15.6% across models. Although substantially lower than the corresponding CD rates, these results show that clear shifts remain present. Moreover, four of five models still exhibit more upward than downward clear shifts, with inflation-to-deflation ratios between 1.4 and 1.8. Thus, absolute distortion prevalence depends on the selected threshold, while the directional tendency toward certainty inflation remains under this stricter criterion. The appropri-



(a) Certainty Distortion Measurements

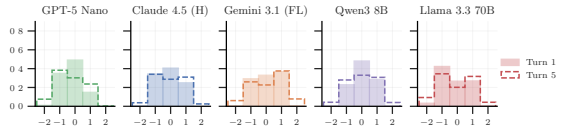


(b) Certainty Inflation Measurements

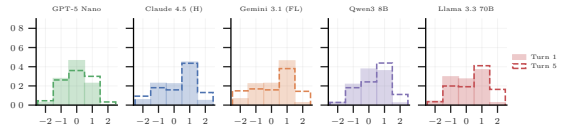


(c) Certainty Deflation Measurements

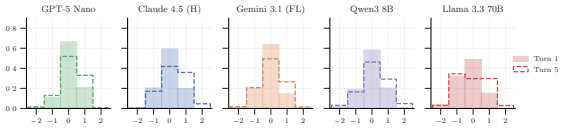
Figure 28: Impact of decoding on different certainty measurements, averaged across tasks in SPICED. Greedy decoding implies LM outputs were generated using `T=0` and top-p decoding concerns the generations with `top-p=0.9` and `T=1`.



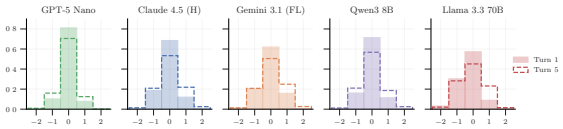
(a) Paraphrase (SPICED)



(b) News Rewrite (SPICED)



(c) Paraphrase (MIMIC-CXR)



(d) Simplification (MIMIC-CXR)

Figure 29: Empirical frequency of LM-as-a-judge scores across all four task–dataset combinations. Solid bars show scores after a single rewrite (Turn 1); dashed outlines show scores after five successive rewrites of the same source text (Turn 5). Scores range from -2 (source text is more certain) to +2 (model output is more certain), with 0 representing "No Clear Difference". Across all conditions, distributions shift rightward with repeated rewriting, indicating that certainty distortion accumulates as the model iterates over its own output rather than the original source.

ate threshold for an acceptable change is ultimately application-dependent.

F.8 Multi-turn Effect

Figure 29 shows the empirical LM-as-a-judge score distributions at Turn 1 and Turn 5 for each task and model, illustrating how the distortion profile shifts after five successive rewrites.

Figure 30 shows per-example LM-as-a-judge trajectories across rewrites for models with strong (`Claude 4.5 (H)`) and weak (`gpt-5.4-nano`) distortion trends on SPICED, across both tasks. Each row represents a single source text; colors indicate the LM-as-a-judge score at that rewrite iteration. Note that rows are not aligned across panels. Each cell corresponds to a different set of source texts, so the comparison is between aggregate trajectory pat-

Model	N	CD	CD-↑	CD-↓	Tie
gpt-5.4-nano	800	7.2	7.2	0.0	92.8
Claude 4.5 (H)	800	13.0	12.8	0.2	87.0
Gemini 3.1 (fl)	732	6.0	5.5	0.5	94.0
Qwen3 (8B)	799	12.6	11.0	1.6	87.4
Llama 3.3 (70B)	800	13.0	11.8	1.2	87.0

Table 16: Certainty distortion in the MIMIC-CXR paraphrase task, measured using the reference-less fine-tuned BERT-based model proposed by Wuehrl et al. (2024). N denotes the number of examples. CD-↑ and CD-↓ denote increases and decreases in expressed certainty, respectively. All values except N are reported as percentages.

Model	N	CD	CD-↑	CD-↓	Tie
gpt-5.4-nano	800	3.1	3.0	0.1	96.9
Claude 4.5 (H)	800	5.1	4.2	0.9	94.9
Gemini 3.1 (fl)	800	6.4	5.8	0.6	93.6
Qwen3 (8B)	799	6.4	5.9	0.5	93.6
Llama 3.3 (70B)	800	5.8	4.4	1.4	94.2

Table 17: Certainty distortion in the MIMIC-CXR simplification task, measured using the reference-less fine-tuned BERT-based model proposed by Wuehrl et al. (2024). N denotes the number of examples. CD-↑ and CD-↓ denote increases and decreases in expressed certainty, respectively. All values except N are reported as percentages.

terns rather than individual examples. Comparing across columns reveals whether drift patterns are task-dependent within a model; comparing across rows shows whether the two models differ in how quickly and consistently distortion accumulates. Figure 31 shows the same analysis on MIMIC-CXR. Unlike SPICED, the clinical domain shows a directional asymmetry in label transitions: generations judged as more certain than the source (red) tend to remain so across rewrites, while the reverse is less stable: blue labels (source text more certain) are more likely to transition to red (LM-generated text more certain) than the other way around. The two models differ in how neutral examples (no clear difference between texts) evolve: for `Claude 4.5 (H)`, the proportion of ties (white) diminishes steadily across rewrites, whereas for `gpt-5.4-nano` it remains approximately stable.

F.9 Additional Qualitative Examples

Tables 18 to 20 provide various qualitative examples of LM generations.

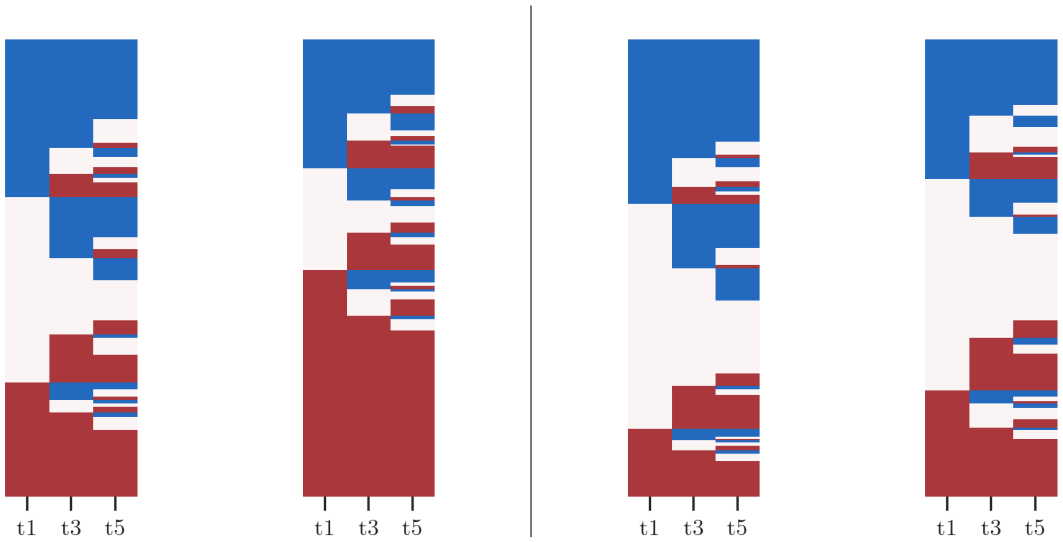
G Document-Level Experiments

Datasets. Similarly to the sentence-level experiments, we restrict these experiments to the scientific and medical domains. Specifically, we consider a dataset of 500 academic abstracts, spanning 5 different life and social sciences disciplines, and extend the sentence-level experiments to the full FINDINGS section of 800 MIMIC-CXR reports. Each FINDINGS section in MIMIC was drawn from a pool of uncertain documents, where at least one of the sentences contains an epistemic marker. See Appendix E for additional details.

Tasks. We extend the medical tasks to the document-level, prompting LMs to paraphrase or simplify the whole findings section. To facilitate the evaluation under stylistic differences, we further instruct the LMs to preserve the ordering when simplifying the language (see prompt in Figure 23). Similarly, we keep the paraphrasing of full abstracts as one of the document-level tasks and introduce targeted paraphrasing, where we instruct the LM to summarize the main findings in an abstract. With these two tasks, we aim to simulate real-world tasks where users (or authors) may resort to these LMs for help polishing their manuscripts and/or performing a literature review.

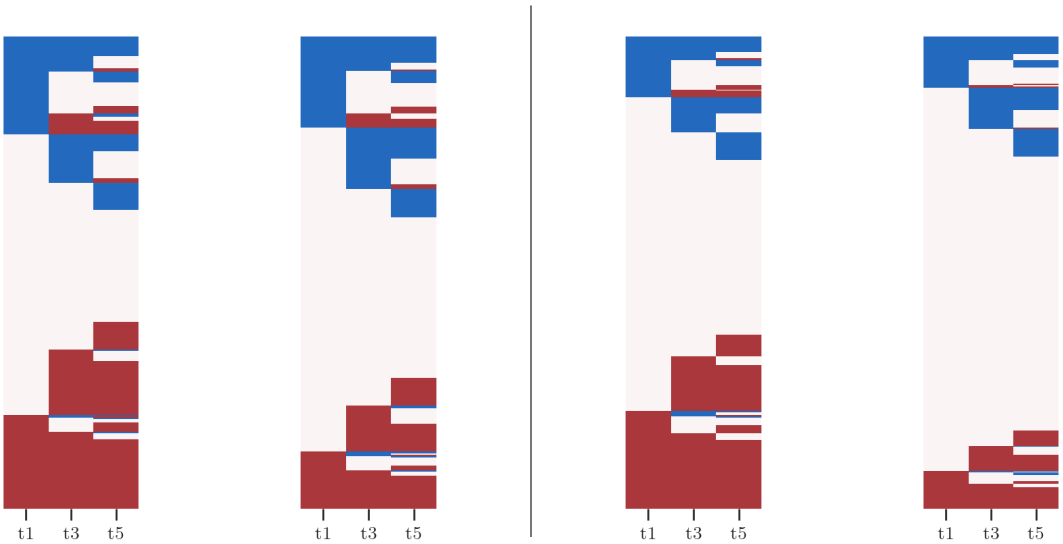
LM-as-a-judge Approach. One challenge in expanding the experiments to the multi-document evaluation is to ensure that the LM-as-a-judge bases its certainty judgment decision on semantically equivalent information. In other words, we want to ensure that semantic difference is not biasing our evaluator in unexpected ways. To encourage the LM to focus on the common claims across pairs of documents, we tweak the LM-as-a-judge prompt by providing it with examples of expected behavior across the academic and clinical contexts.

Given a pair of documents (the original document and the modified document), the LM-as-a-judge identifies a list of shared scientific (or radiological) observations for each pair of documents. For each identified claim, the LM-as-a-judge has to reason through the sources of certainty used to communicate the claim in both documents and, finally, provide its final output. To differentiate certainty distortion from hallucination, we further instruct the LM (explicitly through a requirement as well as implicitly via example) to produce N/A instead of an ordinal score when a claim in the modified text is not grounded in the original text. To en-



(a) Claude 4.5 (H) Paraphrase (SPICED) (b) Claude 4.5 (H) News Rewrite (SPICED) (c) gpt-5.4-nano Paraphrase (SPICED) (d) gpt-5.4-nano News Rewrite (SPICED)

Figure 30: Per-example LM-as-a-judge score trajectories across successive rewrites for Claude 4.5 (H) and gpt-5.4-nano on the SPICED paraphrase and news rewrite tasks. Claude 4.5 (H) and gpt-5.4-nano were selected as models with high and low certainty distortion at Turn 1, respectively. Each row is a source text; blue indicates the source text was judged more certain than the source, red indicates the LM-generated text was judged more certain, and white indicates no clear difference. Rows are sorted by trajectory pattern but are not aligned across panels, each panel represents a different set of source texts. Across all conditions, examples rarely oscillate between red and blue across rewrites; instead, labels tend to stabilize or drift monotonically toward one polarity. gpt-5.4-nano shows a larger proportion of neutral examples at Turn 1 that subsequently shift toward a definite label, while Claude 4.5 (H) exhibits fewer neutral examples from the outset.



(a) Claude 4.5 (H) Paraphrase (MIMIC-CXR) (b) Claude 4.5 (H) Simplification (MIMIC-CXR) (c) gpt-5.4-nano Paraphrase (MIMIC-CXR) (d) gpt-5.4-nano Simplification (MIMIC-CXR)

Figure 31: Per-example LM-as-a-judge score trajectories across successive rewrites for Claude 4.5 (H) and gpt-5.4-nano on the MIMIC-CXR paraphrase and simplification tasks. Both models exhibit a directional asymmetry in label transitions: examples that reach a red label rarely revert, while blue labels are comparatively less stable, making blue-to-red transitions more common than the reverse. However, the two models differ in how neutral examples evolve: for Claude 4.5 (H), the proportion of neutral examples decreases steadily across rewrites, whereas for gpt-5.4-nano it remains approximately stable.

Model	Original	LM-generated
Claude 4.5 (H)	There was a **statistically significant difference** in the global score of SwQoR (SD) between inadequate (37.4 [34.7]) and sufficient FHL (17.7 [21.0]) (P < .001).	Individuals with adequate functional health literacy **demonstrated substantially superior** sleep quality ratings compared to those with inadequate functional health literacy (P < .001).
	Overall, three in four participants **perceived** the advertisements to breach at least one alcohol-specific rule in the BCAP Code.	Three-quarters of study participants **identified** alcohol advertisements as violating at least one provision of the industry's advertising standards code.
Gemini 3.1 (fl)	Together, we **provide sound evidence** of a disease-independent unfortunate relationship between well-defined pre-adult environmental hits and violent aggression, **calling for** more efficient prevention.	Our findings **demonstrate a clear link** between early-life trauma and later violent behavior regardless of underlying health conditions, **underscoring the urgent need** for improved preventative strategies.
	We **show** that this single meal had an overall and long-lasting impact.	Our findings **demonstrate** that consuming this one dish resulted in a significant and enduring effect.

Table 18: Examples of inflation distortions by Gemini 3.1 (fl) and Claude 4.5 (H) when the degree of confidence in the source text is high, sampled from SPICED.

Model	Original	LM-generated
Claude 4.5 (H)	The impact of specialty level culture, and peer influence **appears to** be significant factors of antibiotic decision-making.	The decision to prescribe antibiotics **is** heavily influenced by the cultural norms within a medical specialty and the opinions of colleagues.
	[Conclusion] It was concluded that the 12 weeks of vitamin D supplementation and circuit training **would have positive effects** on abdominal fat and blood lipid profiles in T2D and vitamin D deficient elderly women.	A 12-week regimen involving vitamin D supplementation and circuit training **significantly reduced** abdominal fat and improved blood lipid profiles in elderly women with T2D and vitamin D deficiency.
Gemini 3.1 (fl)	If a birth is a positive life event, the happiness response to having a child **may be** strongest when the event is first experienced.	Considering birth as a positive life event, the happiness associated with parenthood **is** strongest during the first experience.
	We **show** that this single meal had an overall and long-lasting impact.	Our analysis **proves** that a single meal had a widespread and long-term impact.

Table 19: Examples of overconfident distortions in the *Paraphrasing* task in SPICED. LMs are prompted to generate paraphrases of an English text that are 1 sentence long.

sure the LM adheres to these constraints, we use Structured Outputs and obtain the response greedily (T=0). The LM-as-a-judge is operationalized using gpt-5.4-mini.

Note that due to the implicit nature of the medical data, we found it difficult to find a clear definition of a claim, especially given that the LM does not have full access to the radiological report and supporting images. As a result, in the medical communication tasks, our LM-as-a-judge compares each sentence in the modified text against the original text (*i.e.*, the semantic unit is now a sentence instead of a claim). Other than that, the evaluation process remains the same.

LM-as-a-Judge Quality Control. To assess the medical document-level evaluator, we constructed a controlled set pairing source sentences with rewrites that preserved, increased, or decreased certainty. We sampled 20 FINDINGS sections from MIMIC-CXR. From each section, we selected one sentence expressing uncertainty and generated four variants: a paraphrase and three rewrites expressing slightly, moderately, or clearly greater certainty. We repeated this process for one sentence expressing certainty from each section, generating a paraphrase and three rewrites expressing slightly, moderately, or clearly lower certainty. This process yielded 160 source–variant pairs whose intended certainty direction was known by construction. We used Claude Opus 4.8, Gemini 3 Pro, and GPT-5.4 to generate the variants, increasing stylistic diversity and reducing reliance on artifacts from a single generator. The first author inspected every variant for clinical plausibility, semantic faithfulness, and consistency with the intended certainty relation; unsuitable or duplicate variants were resampled. Borderline and extreme cases were additionally discussed with two other authors. This review served as quality control and calibration rather than formal human validation.

We used this controlled set to inspect and refine the two-step LM-as-a-judge described above, particularly the alignment step that isolates the shared observation before comparing certainty. The prompt demonstrations cover increased certainty, decreased certainty, hallucinated content, and no clear difference. Qualitative inspection suggested

that the evaluator generally followed the intended certainty shifts, particularly for clearer cases, while subtle differences remained ambiguous. Because the controlled set also informed prompt refinement and lacked independent human annotations, we treat this analysis as quality control rather than as formal validation.

This analysis is limited to the medical evaluator. The academic-abstract evaluator uses the same two-step procedure and author-curated demonstrations, but lacks an analogous controlled set; its systematic evaluation remains future work.

Metrics. CD, CD-↑, and CD-↓ are all computed using micro-averaging, *i.e.*, they provide an estimate of how likely a claim is to be distorted.

Prompts. The actual prompts used for evaluating the document-level experiments are quite long, so we refer the interested reader to our GitHub repo for the evaluation prompts for Academic Abstracts and Medical experiments.

H AI Usage

AI assistance was used to polish the writing and assist with styling the plots.

Model	Original	LM-generated
gpt-5.4-nano	**These studies provide evidence** that treatment of growth factors can be used to engineer a regeneration response from a non-regenerating amputation wound.	**The studies provide evidence** that growth-factor treatment can be used to engineer a regenerative response in non-regenerating amputation wounds.
Claude 4.5 (H)	**These studies provide evidence** that treatment of growth factors can be used to engineer a regeneration response from a non-regenerating amputation wound.	**Scientists have discovered** that growth factor treatment can trigger regenerative healing in amputation wounds that would normally fail to regrow.
Gemini 3.1 (fl)	**These studies provide evidence** that treatment of growth factors can be used to engineer a regeneration response from a non-regenerating amputation wound.	**Researchers have discovered** that applying growth factors can coax non-regenerating amputation wounds into regrowing lost tissue.
Qwen3 (8B)	**These studies provide evidence** that treatment of growth factors can be used to engineer a regeneration response from a non-regenerating amputation wound.	**Researchers have discovered that** applying growth factors can induce regeneration in non-regenerating amputation wounds, offering new hope for advanced healing therapies.
Llama 3.3 (70B)	**These studies provide evidence** that treatment of growth factors can be used to engineer a regeneration response from a non-regenerating amputation wound.	**Researchers have found** that manipulating growth factors can stimulate regeneration in wounds that would not normally regrow tissue, offering a potential new approach to healing amputations.
gpt-5.4-nano	The hydrophobic gating mechanism **is also consistent with** scanning mutagenesis studies **showing** that modulation of pore hydrophobicity is correlated with activation properties.	Independent scanning mutagenesis experiments **further support** the hydrophobic gating model, **showing** that changes in pore hydrophobicity **track** with the channels' activation behavior.
Claude 4.5 (H)	The hydrophobic gating mechanism **is also consistent with** scanning mutagenesis studies **showing** that modulation of pore hydrophobicity is correlated with activation properties.	**Researchers have found** that ion channels' gating behavior is controlled by the hydrophobicity of their pores, a **discovery confirmed** by genetic studies **showing a direct link** between pore water-repelling properties and channel activation.
Gemini 3.1 (fl)	The hydrophobic gating mechanism **is also consistent with** scanning mutagenesis studies **showing** that modulation of pore hydrophobicity is correlated with activation properties.	New research **reinforces** the theory that hydrophobic barriers **act** as a gate within cellular channels, **revealing** that altering the water-repelling properties of these pores **directly dictates** how they open and close.
Qwen3 (8B)	The hydrophobic gating mechanism **is also consistent with** scanning mutagenesis studies **showing** that modulation of pore hydrophobicity is correlated with activation properties.	**Scientists have found** that the way certain proteins control ion flow **is linked** to how altering their hydrophobicity affects their activation, as demonstrated by genetic experiments.
Llama 3.3 (70B)	The hydrophobic gating mechanism **is also consistent with** scanning mutagenesis studies **showing** that modulation of pore hydrophobicity is correlated with activation properties.	**Researchers have found** that the hydrophobic gating mechanism is **supported** by scanning mutagenesis studies, **which reveal a correlation** between pore hydrophobicity and activation properties.
gpt-5.4-nano	These results **point to the remarkable** efficacy of our nanoparticle-based drug delivery system and **suggest** the value of the chicken egg tumor model for testing novel therapies for ovarian cancer.	The findings highlight the remarkable efficacy of a nanoparticle-based drug delivery system and suggest that a chicken egg tumor model may provide a valuable platform for testing new ovarian cancer therapies.
Claude 4.5 (H)	These results **point to the remarkable** efficacy of our nanoparticle-based drug delivery system and **suggest** the value of the chicken egg tumor model for testing novel therapies for ovarian cancer.	**Researchers have** developed a nanoparticle-based drug delivery system that **shows promise** against ovarian cancer when tested in a novel chicken egg tumor model.
Gemini 3.1 (fl)	These results **point to the remarkable** efficacy of our nanoparticle-based drug delivery system and **suggest** the value of the chicken egg tumor model for testing novel therapies for ovarian cancer.	**A new study demonstrates** that a novel nanoparticle-based drug delivery system is highly effective in treating ovarian cancer when tested in a chicken egg model, **offering a promising** new platform for future therapeutic research.
Qwen3 (8B)	These results **point to the remarkable** efficacy of our nanoparticle-based drug delivery system and **suggest** the value of the chicken egg tumor model for testing novel therapies for ovarian cancer.	**The study highlights** the promising efficacy of a nanoparticle-based drug delivery system for ovarian cancer and underscores the utility of the chicken egg tumor model in evaluating novel therapies.
Llama 3.3 (70B)	These results **point to the remarkable** efficacy of our nanoparticle-based drug delivery system and **suggest** the value of the chicken egg tumor model for testing novel therapies for ovarian cancer.	**Researchers have** made a **breakthrough** in ovarian cancer treatment, **demonstrating** the potent effectiveness of a nanoparticle-based drug delivery system in an innovative chicken egg tumor model that **shows promise** for testing new therapies.

Table 20: Randomly selected examples from the News Rewriting task. Original texts are randomly selected.