Epistemic Erosion: An Error Class Beyond Hallucination in LLM Text Generation

Micky C. Nnamdi[♠], **Shaunna Wang**[♠], **Chinaza Nnamdi**[♡], **Jieru Li**[♠], **Benoit L. Marteau**[♠], **J. Ben Tamo**[♠], **Yining Yuan**[♠], and **May D. Wang**[†♠]

[♠]Georgia Institute of Technology
[♡]The University of Texas at Austin
†Corresponding author: maywang@gatech.edu

Abstract

Expert writers calibrate their claims with qualifiers such as "likely benign" in a radiology report, "may constitute" in a court ruling, and "suggest" in a scientific paper. We ask whether this calibration survives when Large Language Models (LLMs) summarize and rewrite professional text. We define *epistemic erosion* as any change in a source's expressed certainty, including qualified claims that are dropped or introduced without basis. We measure it with the Epistemic Faithfulness Score (EFS), computed by aligning source and output claims one-to-one with rule-based matching and assigning each distortion to one of six types, namely *escalation*, *attenuation*, *omission*, *fabrication*, *scope shift*, and *severity warp*. On clinical notes, mean EFS for 11 open-weight LLMs ranges from 0.415 to 0.493; three of these models are also evaluated on legal and scientific text. Four hosted models (GPT-4o, GPT-5.4, Phi-4, Mistral-Large-3) reach at most 0.483 in any domain. The most frequent detected distortion is omission (66%), in which a qualified claim is dropped entirely; claims that survive tend to shift toward greater certainty. Model scale and medical adaptation bring no consistent gain, seven prompting strategies bring at most modest gains, and ROUGE-L and BERTScore correlate only weakly with EFS ($|r| \approx 0.1$).

1 Introduction

When a radiologist describes a renal lesion as "likely benign" and recommends a follow-up ultrasound, the qualifier carries the clinical reasoning. The lesion is probably harmless but not certainly so, and that residual doubt is the reason to image again. When a Large Language Model (LLM) summarizes the report as "the renal lesion is benign," the finding survives, but its epistemic status does not, and the reason for the follow-up is lost. We term this failure mode *epistemic erosion*, a change in the certainty a source expresses, whether its qualifiers are strengthened, weakened, dropped together with their claims, or introduced without basis.

Whereas hallucination (Ji et al., 2023; Huang et al., 2025) and factual inconsistency (Kryściński et al., 2020) concern whether propositional content is supported by or consistent with the source, epistemic erosion concerns changes in expressed certainty; these error classes can co-occur. A summary asserting "the defendant *is* liable" instead of "the evidence *suggests* liability," for example, may evade standard automated checks despite changing the source's epistemic stance.

The consequences extend across knowledge-intensive domains. In clinical practice, changing "possible malignancy" to "definitive malignancy" can alter how a downstream clinician interprets the status and urgency of a diagnosis. In legal contexts, stripping qualifiers from a brief can turn qualified reasoning into an unsupported assertion. In scientific communication, changing "our results suggest an association" to "our results show an association" strengthens a claim without changing its underlying subject.

Despite the growing use of LLMs in clinical medicine (Singhal et al., 2023; Nori et al., 2023), legal analysis (Cui et al., 2026), and scientific summarization, preservation of source-expressed uncertainty remains under-evaluated. Concurrent work (Du et al., 2026) reports similar erosion in clinical text. Less clear is how certainty is distorted, how consistently the behavior appears across models and domains, and whether generation-time preferences are associated with its persistence. A basic obstacle is measurement. Standard evaluation metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) and BERTScore (Zhang et al., 2020) primarily measure content overlap or semantic similarity. The sentences "the patient has pneumonia" and "the patient *may have* pneumonia," for example, differ in

Code: https://github.com/micknn/Epistemic-Erosion

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4998–5026
November 6-10, 2026 ©2026 Association for Computational Linguistics

epistemic status while remaining nearly identical in propositional content. To study these changes directly, we make three contributions:

1. We formalize epistemic erosion as a distinct error class and propose a six-type distortion typology describing how LLMs alter source uncertainty (§3).

2. We introduce the Epistemic Faithfulness Score (EFS) and a rule-based pipeline for quantifying six operational forms of epistemic distortion, and test them with controlled perturbations, radiologist-annotated uncertainty, and independent judge analyses (§4, §6). We also report a learned sentence-pair classifier as an auxiliary label-learnability experiment (Appendix N).

3. Across 15 models and three professional domains, we detect epistemic erosion in every tested system, with no consistent improvement associated with model scale, medical domain adaptation, or prompting. Overlap and embedding metrics (ROUGE-L, BERTScore) are largely insensitive to these distortions, while token-distribution probes provide evidence consistent with a persistent preference for assertive completions (§6, Appendix M).

2 Related Work

Hallucination and Faithfulness. Much of LLM reliability evaluation focuses on hallucination and unsupported generation (Ji et al., 2023; Huang et al., 2025). Faithfulness methods based on consistency classification (Kryściński et al., 2020), natural language inference (NLI) (Laban et al., 2022), and question answering (QA) (Fabbri et al., 2022) primarily assess whether generated claims are supported by, or entailed by, the source. Similarly, the intrinsic–extrinsic distinction introduced by Maynez et al. (2020) concerns errors in propositional content. The same emphasis appears in professional-domain evaluation, which commonly targets medical question answering (Singhal et al., 2023), medical licensing exam performance (Kung et al., 2023; Nori et al., 2023), and factual fidelity in medical-evidence (Tang et al., 2023) and legal (Cui et al., 2026) generation. Most of these approaches ask whether a generated claim is supported, but do not ask whether the source's degree of certainty is preserved. Among the exceptions, Tang et al. (2023) identify a "certainty illusion" error subtype in a human evaluation of medical-evidence summaries, and Kolagar and Zarcone (2024) measure how uncertainty expressions transfer to GPT-4 summaries of news articles and blogs. Work on uncertainty calibration and expression (Kadavath et al., 2022; Lin et al., 2022; Liu et al., 2025; Yona et al., 2024) addresses a different question again: whether a model's expressed confidence reflects its own accuracy or internal state. Our question, whether the source author's expressed certainty survives generation, is closer to work on preserving other non-propositional properties of text, such as affective content in opinion summarization (Attri and Attri, 2026).

Error Typologies in Text Transformation. Our six-type taxonomy follows a broader tradition of decomposing text transformation errors into operational categories. Devaraj et al. (2022) categorize factuality errors in text simplification, while Du et al. (2022) characterize edit intentions in document revision. We apply the same error-analysis perspective to changes in expressed certainty. This decomposition allows us to study not only whether epistemic fidelity fails, but how it fails.

Uncertainty in NLP. Computational modeling of textual uncertainty has a long history. Early work includes the BioScope corpus (Vincze et al., 2008) for biomedical speculation and the CoNLL-2010 shared task on hedge detection (Farkas et al., 2010), alongside clinical systems for assertion status (Uzuner et al., 2011; Velupillai et al., 2014) and negation and contextual status (Harkema et al., 2009). This literature asks whether uncertainty can be identified within a text; we ask whether uncertainty expressed in a source survives LLM-mediated transformation.

Concurrent Work. Du et al. (2026) independently study uncertainty preservation in clinical text, constructing a benchmark of 1,200 documents and evaluating three LLMs using a five-level uncertainty scale. They similarly find that uncertainty is frequently altered during generation. Their study focuses on the clinical domain and evaluates preservation through level matching; ours complements it by spanning three domains and 15 models, decomposing errors into six types, and examining prompting, scale, domain adaptation, and token-level generation preferences. We do not evaluate directly against their annotations because the two pipelines operationalize uncertainty differently, in-

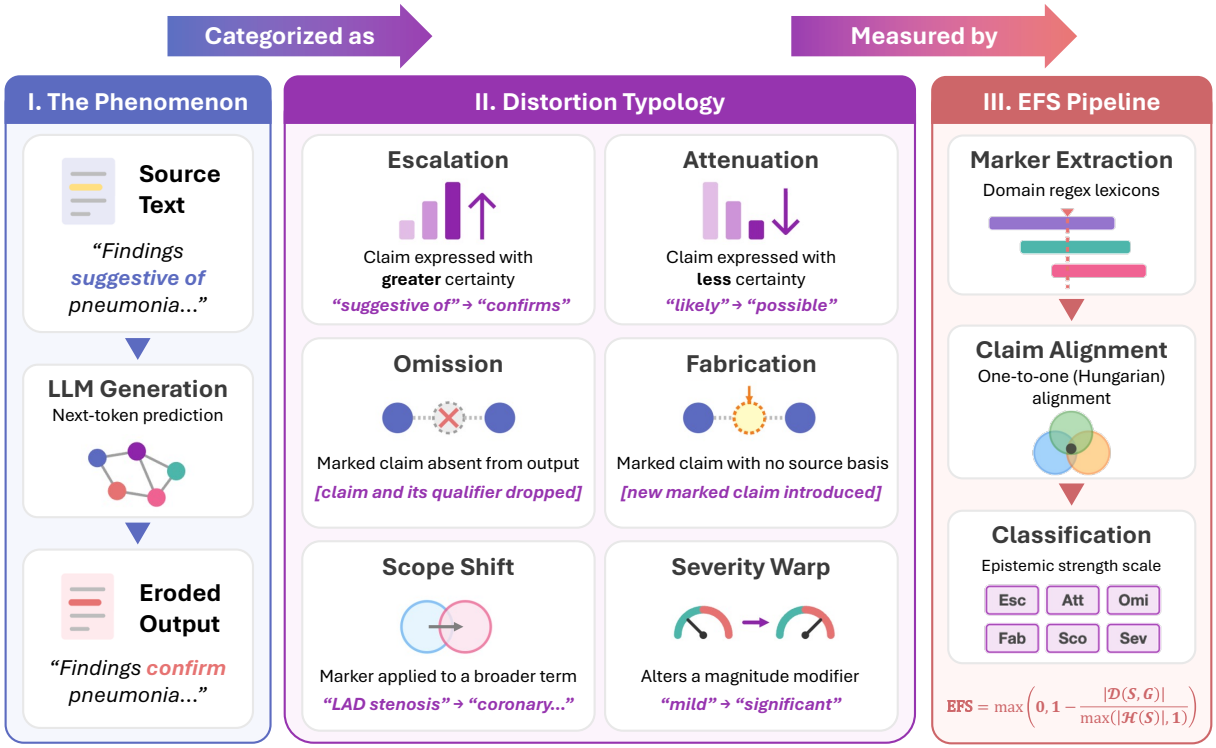


Figure 1: Overview of epistemic erosion. **(I)** An LLM transforms source text containing epistemic markers into an output whose expressed certainty differs from the source. **(II)** The change falls into one or more of six distortion types (escalation, attenuation, omission, fabrication, scope shift, severity warp). **(III)** Our EFS pipeline detects these distortions in three stages: regex-based epistemic-marker extraction, exact-anchored one-to-one claim alignment, and per-type classification.

cluding different treatments of negation and hedging.

3 Problem Formulation

3.1 Definition

Let S denote a source text containing epistemic markers $\mathcal{H}(S) = \{h_1, \ldots, h_n\}$, expressions that qualify the certainty of claims (e.g., "possible," "suggestive of," "the evidence suggests"). Given a generated text $G = \mathrm{LLM}(S)$, we define **epistemic erosion** as a divergence between the certainty expressed in G and that expressed in S. Although the term suggests loss, we use it as an umbrella for departures from the source's epistemic stance in either direction. The divergence arises in two ways: (i) the certainty attached to a source claim changes, or the claim is omitted together with its qualifier; or (ii) G introduces an epistemically marked claim absent from S. Case (ii) can therefore overlap with hallucination.

3.2 Types of Epistemic Erosion

We define six types of epistemic distortion, each paired with an operational detection criterion. An *epistemically marked claim* is a sentence containing both (i) an epistemic marker from the domain-specific lexicon (Appendix A) and (ii) a domain-relevant claim term.

1. **Escalation.** A claim is expressed with greater certainty than in the source (*"Findings suggestive of pneumonia"* → *"Findings confirm pneumonia."*). *Test:* an epistemically marked source claim aligned to an output sentence that either has strictly higher epistemic strength on a four-point scale (from speculative to definitive) or carries no epistemic marker at all.

2. **Attenuation.** A claim is expressed with less certainty than in the source (*"Likely fracture"* → *"possible fracture."*). *Test:* an epistemically marked source claim aligned to an output sentence that retains a marker and whose epistemic strength is strictly lower than the source's.

3. **Omission.** An epistemically marked source claim is absent from the output. *Test:* no claim-aligned output sentence exists for it.

4. **Fabrication.** The output introduces an epistemically marked claim with no basis in the source. *Test:* an unmatched epistemically marked output claim that shares no domain term with any epistemically marked source claim.

5. **Scope shift.** An epistemic marker applies to a broader or narrower scope than in the source (*"Mild stenosis in the left anterior descending artery"* → *"mild coronary artery stenosis."*). *Test:* we operationalize broadening shifts as cases where an output epistemic marker applies to terminology absent from the corresponding source claim. Narrowing shifts are not scored.

6. **Severity warp.** A severity modifier is altered (*"Mild stenosis"* → *"significant stenosis."*). *Test:* an epistemically marked source claim contains a severity modifier from a closed lexicon {mild, moderate, severe, ...} and the aligned output sentence either drops the modifier or uses one at a different level on a three-point ordered scale.

The categories are not mutually exclusive; a generation may exhibit multiple distortion types (examples of all six are in Appendix O). Scope shift and severity warp alter the scope or degree of a qualifier and are therefore treated as boundary categories.

3.3 EFS Measures Fidelity to the Source, Not Epistemic Accuracy

EFS measures fidelity to the certainty expressed in the source, not whether that certainty is itself well calibrated. We therefore treat the source author's epistemic stance as the reference: a model that faithfully preserves poorly calibrated uncertainty scores high, while one that alters it is penalized even if the altered statement would be more accurate under an external ground truth. This source-based definition reflects the transformation tasks we study, which do not provide the model with independent ground truth for revising the author's stated certainty.

4 Method

4.1 Epistemic Faithfulness Score

We quantify epistemic erosion with EFS:

$$\mathrm{EFS} = \max\left(0,\ 1 - \frac{|\mathcal{D}(S,G)|}{\max(|\mathcal{H}(S)|,\ 1)}\right) \quad (1)$$

Here, $\mathcal{D}(S,G)$ is the collection of detected distortion events and $\mathcal{H}(S)$ is the collection of epistemic marker occurrences extracted from the source. EFS is a score derived from a normalized distortion rate, not the fraction of source markers preserved. A score of 1 means that the pipeline detects no distortions; the score reaches 0 when the number of detected distortions is at least as large as the number of source markers. Every source document in the scored generation sets contains at least one marker, so the denominator guard never binds. Because fabrications add distortions with no source marker and an aligned claim can carry two labels, $|\mathcal{D}(S,G)|$ can exceed $|\mathcal{H}(S)|$; hence the lower bound at zero. An unclamped-score analysis and controlled tests of identity, monotonicity, and scale invariance are reported in Appendix J.3.

4.2 Rule-Based Detection Pipeline

Our detection pipeline operates in three stages (Figure 1):

Stage 1: Epistemic-marker extraction. We use domain-specific regex lexicons to identify candidate epistemic markers. Clinical patterns span six categories: diagnostic hedges ("possible," "suggestive of," "cannot rule out"), evidential hedges ("reportedly," "equivocal"), severity modifiers ("mild," "moderate," "severe"), temporal hedges ("acute," "evolving"), recommendation hedges ("consider," "if clinically indicated"), and negation hedges ("unlikely," "no evidence of"). Legal and scientific domains have analogous category structures (Appendix A). We compute $|\mathcal{H}(S)|$ by counting the raw lexicon matches. We evaluate a stricter contextual filter only as an ablation (Appendix D). Claim-level filtering enters in Stage 2.

Stage 2: Claim alignment. Source and output texts are segmented into sentences. Sentences containing both an epistemic marker and domain-relevant terminology are treated as epistemically marked claims; on the output side, every sentence with domain terminology is an alignment candidate, with or without a marker, so that dehedged paraphrases can align. Exact matches receive the maximum similarity; other pairs are scored by a weighted combination of token-level F1 and term containment. We obtain a one-to-one alignment using the Hungarian algorithm (Kuhn, 1955) and retain matches whose similarity reaches a fixed threshold τ (weights and threshold in Appendices C and J.5). Unmatched source claims are candidates for omission, while unmatched output

claims are candidates for fabrication.

Stage 3: Distortion classification. Matched and unmatched claims are evaluated using the six operational criteria defined in §3.2. An aligned claim receives at most two labels: escalation, attenuation, and scope shift are mutually exclusive, and severity warp may co-occur with any of them, so a dropped severity modifier can be charged as both escalation and severity warp. Sentence-level strength is the maximum over the sentence's scaled markers (no marker-to-subclaim scope resolution is performed); markers without a scale entry (severity, temporal, recommendation, evidential) carry no strength, and a sentence with no scaled marker defaults to the definitive level. For fabrication, the clinical pipeline applies the strict term test of §3.2 alone, while the legal and scientific pipelines additionally apply a loose test that flags marked output sentences whose content-bigram overlap with the source falls below a fixed threshold, de-duplicating spans flagged by both tests (Appendix A).

4.3 Learned Distortion Detector

We train four sentence-pair encoders on the pipeline's own labels, which measures learnability rather than independent validity. Each distortion event detected on the clinical generations of one depth-study seed becomes one training instance (labels come from a greedy-alignment variant of the pipeline). Term-sharing source–output claim pairs with equal epistemic strength supply a seventh, "equal strength," class (at most three per generation). Label provenance, dataset sizes, and the label distribution are given in Appendix N. We use a random instance-level split for training and validation, without grouping by document or sentence pair. The encoders are RoBERTa-base and RoBERTa-large (Liu et al., 2019), PubMedBERT (Gu et al., 2021), and SciBERT (Beltagy et al., 2019). Each model receives a concatenated sentence pair (`[CLS] source [SEP] output`) and assigns it to one of seven categories (the six distortion types plus "equal strength"). Omission and fabrication have no counterpart on one side, so omission examples pair the source sentence with an empty output segment and fabrication examples pair an empty source segment with the output sentence. Because the rule-based taxonomy is multi-label, an aligned pair carrying two distortion labels contributes one training instance per label; the resulting identical inputs with different labels introduce ambiguity into the single-label classification task. Training uses AdamW (Loshchilov and Hutter, 2019) with class-weighted cross-entropy loss to handle the natural imbalance across distortion types (hyperparameters in Appendix C). Results are reported in §6.9.

5 Experimental Setup

Data. We evaluate across three professional domains:

(1) **Clinical**: discharge summaries from Medical Information Mart for Intensive Care IV (MIMIC-IV) (Johnson et al., 2023a,b, 2024), stratified across four specialties (cardiology, general medicine, neurology, surgery), case complexity, and diagnostic-hedge density (low/medium/high), restricted to notes of 3,000–20,000 characters. For the depth study, three independently seeded stratified draws of 192 episodes yield 560 unique episodes in total (184–192 retained per seed); generating three task outputs per episode gives 1,680 generations per model. A 1,000-episode breadth study on three models (Llama-3.1-8B, Med42-v2-8B, Mistral-7B-Instruct) complements the depth study.

(2) **Legal**: a stratified 192-case sample from Multi-LexSum (Shen et al., 2022), a corpus of civil rights litigation with multi-granularity summaries. The sample is drawn from a screened pool of 1,881 cases with at least three source markers and is used in all open-weight legal experiments ($n = 1{,}152$: 192 cases $\times$ 3 models $\times$ 2 tasks).

(3) **Scientific**: a stratified 192-document sample from SciTLDR (Cachola et al., 2020), a dataset of computer-science papers with author-written and expert-derived single-sentence summaries. The sample is drawn from a screened pool of 2,615 papers with at least three source markers ($n = 1{,}152$: 192 documents $\times$ 3 models $\times$ 2 tasks). The legal and scientific experiments use the three breadth-study models.

Models. We evaluate 11 instruction-tuned models spanning three model lineages (Llama, Mistral, and Gemma across two generations) and a range of scales: **Llama** (Llama-3.1-8B, Grattafiori et al., 2024; Llama-3.3-70B;[1] Med42-v2-8B and Med42-v2-70B, Christophe et al., 2024; OpenBioLLM-70B, Pal and Sankarasubbu, 2024), **Mistral** (Mistral-7B-Instruct, Jiang et al., 2023; BioMistral-7B, Labrak et al., 2024),

[1] Meta, `Llama-3.3-70B-Instruct`; model card: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Gemma 3 (MedGemma-27B, Sellergren et al., 2025), and **Gemma 4** (Gemma-4-E4B, Gemma-4-31B, Gemma-4-26B-MoE; Gemma Team, 2026; E4B: effective 4B parameters, MoE: mixture-of-experts). Five models (Med42-v2-8B/70B, OpenBioLLM-70B, BioMistral-7B, MedGemma-27B) are medical/domain-adapted variants, enabling comparison between domain-adapted and general-purpose models within the same lineage for Llama and Mistral. Gemma-3-4B (Gemma Team, 2025) is used only in the token-distribution probe (§6.8).

We separately evaluate four hosted models: GPT-4o (OpenAI, 2024), GPT-5.4,[2] Phi-4 (Abdin et al., 2024), and Mistral-Large-3[3] (3 seeds $\times$ 100 documents per domain $\times$ 2 tasks). Legal and scientific evaluations use the same domain datasets described above. In keeping with the MIMIC-IV data-use agreement, we use PubMedQA (Jin et al., 2019) as a proxy free of protected health information for the hosted-model biomedical evaluation rather than sending MIMIC-IV notes to the hosted generation APIs. The hosted runs use shorter task prompts and system prompts that ask the model to preserve the source's qualifiers and the limits on its conclusions (both released with the code). We report this comparison separately because it is not directly equivalent to the open-weight evaluation.

Tasks. For clinical text: hospitalization summary, referral letter, and discharge instructions. For legal text: case summary and legal memorandum. For scientific text: paper summary and plain-language summary. Open-weight generations use greedy decoding (temperature T=0), except in the temperature ablation, with a 2,048-token output cap; the hosted-model runs and the temperature ablation cap output at 1,024 tokens. Source text is truncated to 12,000 characters (one exception is described in Appendix C). The open-weight experiments comprise 43,908 scored generations; the 6,336-generation seed-123 clinical slice used for the cross-domain comparison is part of the depth study, not additional data. The full prompts and a worked input-output example are in Appendices S and R, and all hyperparameters are consolidated in Appendix C.

[2]OpenAI, `gpt-5.4`; system card: `https://deploymentsafety.openai.com/gpt-5-4-thinking`.

[3]Mistral AI, `Mistral-Large-3`; release announcement: `https://mistral.ai/news/mistral-3/`.

Mitigation strategies. We test seven prompting-based approaches to reducing erosion. Three strategies are evaluated on two models that share a base architecture but differ in training: Llama-3.1-8B and Med42-v2-8B, a medical fine-tune of the closely related Llama-3-8B (n = 1,152 generations per strategy, balanced across both models). We replicate these experiments on Mistral-7B-Instruct and Gemma-4-E4B (n = 576 per cell, Appendix H). The strategies are: (1) *Instruction*: an explicit system prompt directing the model to preserve all uncertainty markers from the source; (2) *Few-shot*: four examples pairing an input with both a correct (hedge-preserving) and incorrect (hedge-eroding) output; (3) *Post-hoc verification*: a self-verification step in which the model is instructed to review its own output for hedge preservation and emit only the final text. Four additional strategies (a strict verbatim-copy instruction, reason-first self-review, a few-shot+post-hoc combination, and a hard-constraint framing) are evaluated on three models (Llama-3.1-8B, Med42-v2-8B, Mistral-7B; n = 576 per model-strategy cell). We also test sampling at $T \in \{0.3, 0.7\}$ (50 episodes $\times$ 3 tasks, Llama-3.1-8B), compared against greedy decoding.

Validation. We use four complementary validation checks. A controlled test removes a known fraction of source hedges and checks that EFS falls accordingly (Appendix J.7). Llama-3.3-70B evaluates a balanced reference set as a separate LLM judge that receives no pipeline labels (§6.6, Appendix I). A 20-pair multi-domain pilot, independently annotated by three raters, provides a human agreement check on the six-type taxonomy (Appendix K). Finally, a lexicon-blind evaluation against radiologist uncertainty annotations tests preservation of human-marked uncertainty (§6.7).

6 Results

6.1 Epistemic Distortion Appears Across All Tested Models

Table 1 reports EFS for all 11 open-weight models on the multi-seed depth study. Mean EFS ranges from 0.415 to 0.493 (overall mean 0.450), with significant differences across models (Kruskal–Wallis $H = 409.8$, $p < 10^{-81}$). Across these 11 models, model-level EFS is not positively associated with parameter count over the E4B–70B range (Pearson $r = -0.281$, $p = 0.402$). Medical/domain-adapted models score slightly higher than general-

Table 1: Complete results across all 11 open-weight models (depth study, 3 random seeds, 560 unique episodes, $n = 1{,}680$ per model). **Bold** = medical/domain-adapted. EFS reported as mean ± standard deviation (std) across the 3 seeds. Dist/Gen = mean detected distortions per generation. Distortion columns show percentage of each type out of total distortions for that model, pooled across seeds: escalation (Esc.), attenuation (Att.), omission (Omi.), fabrication (Fab.), scope shift (Scope), and severity warp (Sev.W.).

Model	Params	EFS	Dist/Gen	Esc.%	Att.%	Omi.%	Fab.%	Scope%	Sev.W.%
BioMistral-7B	7B	.493±.002	10.423	7.6	0.3	82.9	3.2	3.3	2.7
Llama-3.1-8B	8B	.481±.005	10.611	10.6	1.5	72.3	4.9	6.4	4.2
OpenBioLLM-70B	70B	.478±.001	10.660	10.4	1.2	75.3	4.1	5.0	3.9
Mistral-7B-Inst.	7B	.478±.002	10.619	11.0	1.3	73.4	4.5	5.6	4.2
Med42-v2-8B	8B	.477±.006	10.521	10.5	1.6	69.6	7.1	7.2	4.0
Llama-3.3-70B	70B	.447±.008	11.113	11.3	1.9	67.5	7.7	7.3	4.4
Gemma-4-E4B	E4B	.432±.011	11.278	13.4	1.8	58.4	12.4	8.7	5.2
Med42-v2-70B	70B	.420±.010	11.524	12.6	2.7	60.9	10.3	8.0	5.5
Gemma-4-26B-MoE	26B	.418±.020	11.573	13.3	2.1	58.2	12.3	8.7	5.4
Gemma-4-31B	31B	.416±.013	11.608	13.4	1.8	59.4	12.0	8.0	5.4
MedGemma-27B	27B	.415±.015	11.629	14.3	3.1	55.0	12.7	8.8	6.1

purpose models (0.457 vs. 0.445); the difference is statistically detectable at the generation level (Mann–Whitney $p = 3 \times 10^{-4}$, Appendix B) but small in magnitude. These patterns are robust to the pipeline's two main settings. Restricting $|\mathcal{H}(S)|$ to certainty-bearing markers lowers the three-seed mean from 0.450 to 0.366, while the model ranking stays strongly correlated with the original (Spearman 0.873). Varying the alignment threshold over $[0.20, 0.50]$ on a seed-123 subsample moves the mean by less than 0.01 (Appendices J.4 and J.5).

Cross-seed standard deviations per model range from 0.001 to 0.020 (mean 0.009). In the breadth study (3 models, 1,000 episodes), EFS also differs significantly across clinical specialties (Kruskal–Wallis, Appendix B). Distortion *profiles* vary substantially: BioMistral-7B has a relatively low fabrication share (3.2%) and a high omission share (82.9%), while MedGemma-27B has the largest fabrication share (12.7%) and a lower omission share (55.0%).

6.2 Omission Is the Largest Detected Distortion Category

In the multi-seed depth study, omission accounts for 66.3% of detected distortions at the operating alignment threshold (mean 7.327 per generation), a share that is sensitive to the threshold (Appendix J.5). Escalation follows at 11.7%, fabrication at 8.4%, scope shift at 7.0%, severity warp at 4.7%, and attenuation at 1.8% (Table A18, Appendix P). Because a dropped epistemically marked claim can reflect either epistemic loss or legitimate compression (omission shows the widest inter-annotator disagreement in the pilot study, Appendix K), its absolute rate should not be read as pure de-hedging. However, omission remains common in the longest output quintile (0.290 per source marker), and substantial distortion remains when omission events are excluded altogether (seed-123 clinical EFS 0.747 against 0.443 with omission included; Appendix E). Restricting $|\mathcal{H}(S)|$ to certainty-bearing markers increases the relative share of escalation to 21.3% (Appendix J.4).

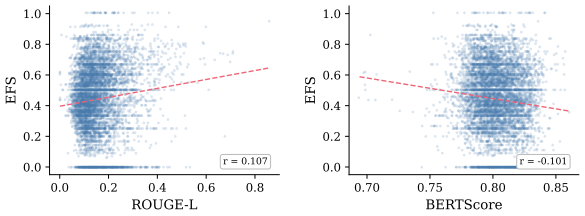


Figure 2: ROUGE-L (left) and BERTScore (right) against EFS for $N = 6{,}336$ clinical source–output pairs. Dashed lines are least-squares fits.

6.3 Standard Metrics Show Little Sensitivity to Epistemic Distortion

Figure 2 plots ROUGE-L against EFS for $N = 6{,}336$ clinical source–output pairs (all 11 open-weight models, seed 123), with the source document as the reference. ROUGE-L is only weakly correlated with EFS ($r = 0.107$, $p = 1.2 \times 10^{-17}$; $R^2 = 0.011$). BERTScore (Zhang et al., 2020) shows only a weak negative association ($r = -0.101$, $p = 6.5 \times 10^{-16}$). A controlled perturbation confirms the distinction: mechanically removing hedges lowers EFS from 1.000 to 0.268 while ROUGE-L remains 0.994, whereas removing unrelated content lowers both (Appendix F). In the

Table 2: Multi-seed EFS (mean ± cross-seed std) for four Azure-hosted models on legal, biomedical (PubMedQA), and scientific text. 3 seeds × 100 documents per domain × 2 tasks per document; 7,005 of 7,200 requests are scored; the remainder failed at generation or scoring (Appendix B). **Bold** indicates the highest mean EFS per column. Legal values are partly affected by the 1,024-token output cap (Appendix E).

Model	Legal	Biomedical	Scientific
Phi-4 (14B)	**.416±.045**	**.406±.057**	**.483±.006**
GPT-4o	.342±.037	.373±.045	.468±.016
Mistral-Large-3	.302±.034	.335±.056	.443±.006
GPT-5.4	.201±.022	.343±.060	.422±.006

controlled test of §5, EFS decreases monotonically with the removed fraction (Spearman $\rho = -0.891$; Appendix J.7). One likely reason is that many epistemic distinctions are expressed through short lexical cues such as "may," "possible," and "suggestive of," while most of the surrounding propositional content remains unchanged. Such sentence pairs can therefore remain highly similar under both overlap-based and embedding-based metrics despite differing in expressed certainty.

6.4 Prompting Produces Limited and Inconsistent Improvements

Prompting effects are architecture-dependent rather than uniformly absent: the three initial prompting strategies produce little change on the two Llama-based models, with similarly small effects on Mistral-7B but a larger gain on Gemma-4-E4B (best: few-shot, Cohen's $d = 0.379$). Four additional strategies yield no improvement larger than 0.01 over the three-model baseline (Table A6; Appendices G and H).

6.5 Cross-Domain Patterns and Hosted-Model Results

Among the three models evaluated in all domains, mean EFS is 0.478 for clinical, 0.427 for scientific, and 0.359 for legal text (Kruskal–Wallis $H = 135.2$, $p = 4.4 \times 10^{-30}$). Because the domains use different lexicons and claim distributions, absolute cross-domain values should be interpreted cautiously. Restricting the analysis to the four core distortion categories preserves the same qualitative pattern (Appendix Q).

Table 2 shows that mean EFS is at most 0.483 for all four Azure-hosted models in every domain. Phi-4 has the highest mean EFS in each domain (seven of the nine two-sided pairwise comparisons between Phi-4 and the other hosted models are significant; Table A2). The legal setting yields the lowest mean EFS for three of the four models under these task and pipeline configurations (for Phi-4 the biomedical value is marginally lower); many legal outputs reach the $1{,}024$-token cap, so part of this gap reflects truncation (Appendix E).

6.6 LLM-as-Judge Validation

Using the Llama-3.3-70B judge of §5, we first assess false-positive behavior on a byte-identical identity control (output = source): the pipeline yields $\mathrm{EFS}(s, s) = 1$, while the judge, using the reference-set prompt of Appendix I, flags $1/242$ cases. In a separate directional test on 750 source–output pairs (clinical: three models; scientific: two), the judge sees only the raw source text and the output. Among the 541 responses accepted by the JSON parser, the judge rates the output as *more certain* than the source in 96.3% of the cases (clinical 98.3%, scientific 94.6%; 2.8% less certain). Recovering the 209 otherwise valid responses wrapped in code fences gives an overall rate of 97.1% across all 750 responses (Appendix J.8).

6.7 Independent Evidence from Radiologist-Annotated Uncertainty

We treat observations marked uncertain by radiologists in RadGraph (Jain et al., 2021a,b) (680 per model over 200 reports) as the reference set and have a lexicon-blind judge (Llama-3.3-70B, shown neither our lexicon, EFS, nor any pipeline label) decide whether each is preserved, hardened (stated with more certainty), or omitted in a summary. In a paired design, the same 389 observations are judged against the source report itself and against a model summary: uncertainty is preserved for 83.8% in the source arm but 54.5% in the model arm, a 29.3-point drop (exact McNemar $p = 1.2 \times 10^{-25}$; report-clustered bootstrap 95% confidence interval (CI) $[23.3, 35.3]$). The judge labels 12.8% of observations mentioned in the source arm as hardened, compared with 36.1% of those mentioned in the model arm ($2.816\times$). For the 329 observations mentioned in both arms, the corresponding rates are 14.3% and 36.5%, respectively. Across all $2{,}720$ judged pairs, the RadGraph-anchored uncertainty-survival rate is 0.510 (clustered CI $[0.465, 0.555]$). This measure is conceptually distinct from EFS and should not be interpreted as a direct numerical validation of the score.

The clinical extractor, restricted to markers with non-zero epistemic strength, agrees with the radiologists' uncertainty labels at F1 0.588 over 3,622 observation-bearing sentences (precision 0.570, recall 0.609), compared with an inter-radiologist F1 of 0.884 on the doubly annotated test split. It transfers zero-shot to BioScope abstracts at F1 0.395 (0.559 for the scientific extractor). Using a separate Spanish lexicon, we obtain F1 0.612 on Spanish clinical text (Appendix J.1). Additional subset analyses, caveats, and per-model results are reported in Appendix J.2.

6.8 Generation-Time Preference for Assertive Tokens

To test whether erosion has a correlate in the next-token distribution, we place 500 hedge-bearing MIMIC sentences and 500 length-matched non-hedge sentences in a fixed summarization prompt and read the probability mass that Llama-3.1-8B assigns to a prespecified set of assertive tokens versus hedge tokens at the position after "The CT". Without a hedge in the source, assertive tokens receive $24.818\times$ the mass of hedge tokens; with a hedge in the source the ratio falls to $3.449\times$, an 89.7% reduction of the excess over parity, yet 90.0% of hedged prompts still favor assertive tokens. The same design on Mistral-7B-Instruct, Med42-v2-8B and Gemma-3-4B gives baseline ratios of 28.696–$83.080\times$, reductions of 90.3–96.9%, and residuals of 3.574–$7.605\times$ (Appendix M). This is correlational evidence from one local generation context and does not establish a causal mechanism.

6.9 Learned Detector Results

Sentence-pair encoders trained on pipeline labels achieve similar validation accuracy across architectures, with no evidence of an advantage for domain-specific encoders; because this is the label-reproduction probe of §4.3, absolute accuracy is optimistic (multi-label ambiguity and repeated inputs across the random split). Full results and construction diagnostics are reported in Appendix N, and cross-domain transfer in Appendix L.

7 Conclusion

We formalize and measure *epistemic erosion*, the distortion of uncertainty expressed in source text during LLM-mediated transformation. Across the models, domains, and generation settings studied here, detected epistemic distortion is pervasive, arising mostly from dropped epistemically marked claims and, for retained claims, from shifts toward greater certainty. Model scale and domain adaptation are not associated with consistently higher faithfulness, and prompting produces modest, architecture-dependent improvements. Token-distribution probes across four models show a residual preference for assertive over hedged continuations after conditioning on hedged source text, providing correlational rather than causal evidence about the observed erosion. Complementary LLM-judge and radiologist-annotated analyses indicate that the shift toward greater expressed certainty is not an artifact of EFS alone. Together, these results motivate epistemic faithfulness as a distinct dimension of source fidelity that should be evaluated alongside factual consistency in professional text generation.

Limitations

EFS is a pipeline-relative measure of source fidelity, not of epistemic accuracy, perceived severity (Appendix K), or harm. Its hand-built lexicons are incomplete, and the marker inventory includes severity and temporal modifiers that are not purely epistemic (Appendix J.4).

Omission is hard to separate from benign compression (Appendix E), and scope shift captures broadening only. Fabrication targets new marked claims, so invented certain assertions remain the domain of factuality evaluation. Type shares are more threshold-sensitive than EFS, and detection rules differ across domains, so absolute values are not directly calibrated across domains and should be compared cautiously (Appendices A and J.5).

The human pilot is small. The RadGraph analysis relies on an LLM judge from the same family as two of the four generators. Statistical tests treat shared-episode generations as independent, although episode-level re-analysis changes no conclusion (Appendix B). The hosted biomedical evaluation uses PubMedQA rather than MIMIC-IV, and hosted legal outputs are often truncated. The lexicons used in the main EFS analysis are English (Appendix F). The learned detector only reproduces pipeline labels (Appendix N), so its accuracy measures agreement with the pipeline rather than detection of real distortions; the rule-based results are unaffected.

Ethical Considerations

MIMIC-IV discharge notes and RadGraph reports were used under PhysioNet (Pollard et al., 2026; Goldberger et al., 2000) credentialed data-use agreements; no patient-identifying information is reproduced in any example, figure, or appendix. The remaining corpora (Multi-LexSum, SciTLDR, PubMedQA, BioScope, NUBes/IULA+) are public. Human annotation was performed by the authors and expert volunteers on de-identified text; no new personal data were collected. Claude Sonnet 4.6, used as a judge and annotator, received MIMIC-derived excerpts through the Anthropic API, which PhysioNet's guidance on credentialed data lists as acceptable because prompts are not used for training by default and are not routinely reviewed.[4] Our results identify a form of source distortion that current evaluation practice can miss; they do not establish the downstream safety of any deployment, nor argue against the use of LLMs in these domains.

Code Availability

We release the pipeline. The breadth-study sampler, one Claude judge control script, and the human annotation files (which identify the raters) are withheld; no data, model outputs, or weights are redistributed, and MIMIC-derived inputs require PhysioNet access.

Use of AI Assistants

For manuscript preparation, LLMs were used as general-purpose assistive tools to polish the manuscript's language and to refine instructions within our prompt templates. Specifically, LLMs aided in improving grammar, clarity, and style, and in suggesting alternative phrasings for prompt templates. Experimental uses of LLMs as generators, judges, and annotators are described in §5 and the appendices.

Acknowledgments

The authors acknowledge the AI Makerspace of the College of Engineering (RRID:SCR_028058), provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA (RRID:SCR_027619), for computational resources used in this work. We also gratefully acknowledge funding and fellowships that contributed to this work, including a Wallace H. Coulter Distinguished Faculty Fellowship, a Petit Institute Faculty Fellowship, and research funding from Amazon and Microsoft Research awarded to Professor May D. Wang.

[4] https://physionet.org/news/post/gpt-responsible-use

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, et al. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Arnav Attri and Anuj Attri. 2026. Bridging cognition and affect: Emotion-aware opinion summarization using LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 4506–4565. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777. Association for Computational Linguistics.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical LLMs. *Preprint*, arXiv:2408.06142.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jiaxi Cui, Munan Ning, Zongjian Li, Hao Li, Yang Ya, Bohua Chen, Bin Ling, Yonghong Tian, and Li Yuan. 2026. Chatlaw: A multi-agent legal assistant based on a role-aligned mixture-of-experts architecture. *Fundamental Research*.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984. Association for Computational Linguistics.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345. Association for Computational Linguistics.

Hongbo Du, Zixin Lu, and Jiaming Qu. 2026. Possible or definite? A benchmark for evaluating diagnostic uncertainty preservation in clinical text. *Preprint*, arXiv:2606.18471.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601. Association for Computational Linguistics.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12. Association for Computational Linguistics.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

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

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1).

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2).

Ken Hyland. 1998. *Hedging in scientific research articles*. John Benjamins Publishing Company.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021a. RadGraph: Extracting clinical entities and relations from radiology reports (version 1.0.0). PhysioNet.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis P Langlotz, and Pranav Rajpurkar. 2021b. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, et al. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577. Association for Computational Linguistics.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV (version 3.1). PhysioNet.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*. Version 2.2.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023b. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Zahra Kolagar and Alessandra Zarcone. 2024. Aligning uncertainty: Leveraging LLMs to analyze uncertainty transfer in text summarization. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 41–61, St Julians, Malta. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346. Association for Computational Linguistics.

Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Salvador Lima Lopez, Naiara Perez, Montse Cuadros, and German Rigau. 2020. NUBes: A corpus of negation and uncertainty in Spanish clinical texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5772–5781. European Language Resources Association.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, KDD '25, page 6107–6117, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Montserrat Marimon, Jorge Vivaldi, and Núria Bel. 2017. Annotation of negation in the IULA Spanish clinical record corpus. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 43–52, Valencia, Spain. Association for 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. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *Preprint*, arXiv:2303.13375.

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. 2026. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, et al. 2025. MedGemma technical report. *Preprint*, arXiv:2507.05201.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-world summaries of civil rights lawsuits at multiple granularities. In *Advances in Neural Information Processing Systems*, volume 35, pages 13158–13173.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Sumithra Velupillai, Maria Skeppstedt, Maria Kvist, Danielle Mowery, Brian E Chapman, Hercules Dalianis, and Wendy W Chapman. 2014. Cue-based assertion classification for Swedish clinical text—Developing a lexicon for pyConTextSwe. *Artificial Intelligence in Medicine*, 61(3):137–144.

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(Suppl 11):S9.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Hedge Detection Patterns

Our clinical lexicon covers six categories of epistemic markers: **diagnostic hedges** ("possible," "likely," "cannot rule out," "suggestive of," "concerning for"), **evidential hedges** ("reportedly," "per report," "equivocal"), **severity modifiers** ("mild," "moderate," "severe," "significant"), **temporal hedges** ("acute," "chronic," "evolving," "new"), **recommendation hedges** ("consider," "if clinically indicated," "recommend evaluation"), and **negation hedges** ("unlikely," "no evidence of," "low probability").

Legal lexicons cover five categories: **legal hedging** ("allegedly," "may constitute," "arguable"), **evidential qualifiers** ("the record suggests," "the evidence indicates," "insufficient evidence," "fails to establish"), **evidential standards** ("preponderance," "substantial evidence," "clear and convincing"), **conditional hedging** ("assuming arguendo," "to the extent that," "absent a showing"), and **scope qualifiers** ("in some circumstances," "generally," "not necessarily").

Scientific lexicons cover five categories: **results hedging** ("suggests," "indicating," "may contribute," "appears to"), **evidential hedging** ("our results suggest," "the findings indicate," "based on our data"), **scope qualifiers** ("in our study," "in this study," "under these conditions"), **limitation markers** ("further research is needed," "preliminary," "beyond the scope of"), and **quantitative hedges** ("approximately," "modest," "substantial").

Epistemic strength scale (4-point). The scale draws on the hedging literature (Hyland, 1998; Vincze et al., 2008) adapted to clinical hedge categories. Escalation and attenuation detection use an ordered scale; the assignment of each clinical hedge term to a strength level is:

- **1 (speculative)**: possible, possibly, may, might, could, cannot rule out, cannot exclude, rule out, equivocal, indeterminate, questionable; and the negation hedges unlikely, less likely, low probability, low suspicion, without definite, no definite evidence (listed explicitly so that they do not resolve by substring to "likely" or "definite").
- **2 (suggestive)**: suspected, suggestive of, suspicious, concerning for, consistent with, compatible with, appears to, seems to.
- **3 (probable)**: likely, probable, most likely, presumed.
- **4 (definitive)**: confirmed, definite, diagnostic of, proven.

Level-4 terms are not lexicon markers and are therefore not matched on their own; a sentence reaches level 4 only through the default for sentences with no scaled marker (§4), so a sentence whose only cue is a level-4 term does not form an epistemically marked claim (the one exception is the negated lexicon pattern "no definite sign," which resolves to level 4 by substring). The same scale is applied in the legal and scientific pipelines; markers containing none of these terms (e.g., "allegedly,"

"suggests," "indicates") carry no strength, so for them escalation is detected only through loss of the marker.

Severity scale (3-point). Severity warp detection uses a closed lexicon of magnitude modifiers: **1 (mild)**: mild, minimal, slight, subtle, trace, borderline, small, low-grade; **2 (moderate)**: moderate, intermediate; **3 (severe)**: severe, significant, marked, extensive, profound, large. Severity terms are matched as substrings in all three domains (so "statistically significant" registers as level 3).

Contextual filter (ablation only). Under this filter, a hedge candidate is retained only if its sentence also contains at least one domain claim signal (clinical: diagnoses, findings, treatments, procedures; legal: holdings, rulings, motions, statutes; scientific: results, methods, datasets, hypotheses) and does not match a domain false-positive pattern (e.g., "may we help" or "stable relationship" in clinical text, "arguable case" framings without a substantive claim in legal text). Full pattern lists, claim-signal regexes, and false-positive filters are in the code release.

Fabrication detector (§3.2, §4). Fabrication detection combines two stages. The *strict detector* applies the test of §3.2; terms that occur only in unmarked source sentences are not consulted. The *loose detector* flags an output sentence that contains a modal or evidential hedge from a separate closed list (severity terms excluded) and a domain claim term, and whose content-bigram overlap with the source is below 0.30, catching cases where the domain term appears in the source but is attached to a different claim. A manual spot check of 16 spans flagged on 44 source-output pairs judged 11 to be fabrications (precision $11/16 = 0.688$). Because the per-span adjudication record was not retained, we treat this as a descriptive check rather than a formal validation. The clinical pipeline applies the strict detector only; the legal and scientific pipelines apply both; the loose detector operates on output sentences independently of the alignment, and spans flagged by both detectors are de-duplicated on their raw text. Because the two detectors segment sentences differently, this de-duplication misses 34 spans in 26 legal outputs (none in scientific), which are counted twice; removing them would raise the legal mean EFS from 0.359 to 0.361. We report the scores as computed by the released code. Applying the loose detector to the clinical depth study would add fewer than 3% to the 204,220 detected distortions (2.4–2.7% depending on how spans already flagged by the strict detector are de-duplicated), touching roughly one generation in five.

B Statistical Tests

Table A1 reports the tests.

Table A1: Statistical tests on EFS across experimental factors. Every effect has the same sign and significance under the greedy baseline aligner of Appendix J.5.

Factor	Test	Stat.	p
Model (11, depth)	Kruskal–Wallis	$H = 409.8$	$< 10^{-81}$
Task (3, depth)	Kruskal–Wallis	$H = 113.5$	$< 10^{-24}$
Specialty (4, breadth)	Kruskal–Wallis	$H = 966.9$	$< 10^{-208}$
Scale (E4B–70B, per-generation)	Pearson r	$r = -.044$	$< 10^{-8}$
Scale (E4B–70B, per-model)	Pearson r	$r = -.281$	.402
Medical vs. general	Mann–Whitney	$U = 4.4 \times 10^7$	3×10^{-4}
Domain (3-way, matched models)	Kruskal–Wallis	$H = 135.2$	$< 10^{-29}$
Temperature (T=0.3)	Mann–Whitney	-	.041
Temperature (T=0.7)	Mann–Whitney	-	.057

Cluster-level re-analysis. Generations that share a source episode are not independent, so we repeat the main tests at the cluster level (EFS rounded to nine decimals first). Across the 11 models, a Friedman test with episode×task×seed blocks ($n = 1{,}680$) gives $\chi^2 = 1863.6$ and with episode-level means ($n = 560$) $\chi^2 = 1437.1$, both $p < 10^{-300}$. The medical/domain-adapted versus general-purpose difference, computed per episode as the mean over the five adapted models minus the mean over the six general-purpose models, is $+0.011$ (episode-level bootstrap 95% CI $[0.008, 0.014]$; Wilcoxon signed-rank $p = 1.1 \times 10^{-16}$, $n = 560$), so the difference is real but small. On document-level means (192 documents per domain; the three matched models, clinical seed 123) the cross-domain Kruskal–Wallis test gives $H = 28.9$, $p = 5.3 \times 10^{-7}$, and all three pairwise Mann–Whitney comparisons are significant (clinical vs. scientific $p = 0.008$; the other two $p < 0.002$). For mitigation, episode-paired Wilcoxon tests against the seed-123 baseline agree with the unpaired tests of Tables A5 and A7: none of the three initial strategies changes EFS on Llama-3.1-8B+Med42-v2-8B ($p \geq 0.11$) or Mistral-7B ($p \geq 0.07$), while all three improve Gemma-4-E4B ($p < 10^{-11}$). Among the four additional strategies (Table A6), the paired tests detect the few-shot+post-hoc gain ($+0.009$, $p = 0.001$) and the hard-constraint loss (-0.010, $p = 0.02$); both are about 0.01 in absolute size, so the conclusion that prompting yields at most modest gains is

unchanged.

Hosted-Model Pairwise Comparisons

Of the 7,200 hosted requests (Table 2), 7,005 are scored; the 195 excluded are content-filter rejections (36 legal, 5 biomedical), 2 legal requests that returned no output, and 152 seed-0 biomedical generations whose 19 source abstracts were unavailable at scoring time. All scientific requests are scored. Table A2 reports all $\binom{4}{2} = 6$ pairwise Mann–Whitney comparisons across the four hosted models (two-sided). As for the open-weight runs, the domain is passed to the pipeline explicitly, with PubMedQA abstracts scored under the clinical lexicon. The two non-significant Phi-4 comparisons are against GPT-4o on biomedical and scientific text.

Table A2: Pairwise two-sided Mann–Whitney p-values across the four hosted models. $^{**}p < 0.01$, $^{*}p < 0.05$, ns = not significant; Mistral-L-3 = Mistral-Large-3.

Pair	Legal	Biomedical	Scientific
Phi-4 vs. GPT-4o	$2.5\times10^{-4\,**}$	.146 ns	.465 ns
Phi-4 vs. Mistral-L-3	$4.0\times10^{-9\,**}$	$5.8\times10^{-4\,**}$	$9.7\times10^{-3\,**}$
Phi-4 vs. GPT-5.4	$2.6\times10^{-30\,**}$	$.003^{**}$	$4.8\times10^{-5\,**}$
GPT-4o vs. Mistral-L-3	$.014^{*}$	$.044^{*}$	.068 ns
GPT-4o vs. GPT-5.4	$4.2\times10^{-17\,**}$	.112 ns	$.001^{**}$
Mistral-L-3 vs. GPT-5.4	$5.8\times10^{-9\,**}$	.701 ns	.151 ns

C Hyperparameters

Table A3 consolidates every hyperparameter used across generation, the claim aligner, the learned detector, and the LLM-as-judge. The 43,908 scored open-weight generations comprise 18,480 depth, 9,000 breadth, 13,824 mitigation, 1,152 legal, 1,152 scientific, and 300 temperature-ablation generations. Decoding is greedy (T=0) for every open-weight run except the temperature-ablation runs and for three of the four hosted models; the GPT-5.4 API exposes no temperature parameter, so GPT-5.4 is evaluated using the API's default decoding settings. The open-weight output cap is the generator default of 2,048 new tokens (the 1,024 cap applies only to the hosted-model and temperature runs); 1,004 of the 18,480 depth-study outputs exceed 1,024 tokens (135 reach the cap), as do 73 of the 1,152 scored legal and 5 of the 1,152 scored scientific outputs. For 192 seed-123 generations of Llama-3.1-8B and Med42-v2-8B (the 32 notes longer than 12,000 characters, three tasks each), an earlier generator passed the full note or its first 15,000 characters; scoring uses the first 12,000 characters for every generation. Excluding these rows would give 0.484 instead of 0.481 for Llama-3.1-8B and 0.478 instead of 0.477 for Med42-v2-8B. The aligner's only hand-set values are its match threshold τ and its $0.65/0.35$ score weights (threshold sensitivity in Appendix J.5).

Table A3: Consolidated hyperparameters for every component.

Component	Setting
Generation (open-weight)	
Decoding	greedy, $T = 0$
Max new tokens	2,048; 1,024 for the temperature ablation
Source truncation	12,000 characters
Backend	Hugging Face Transformers
Temperature ablation	$T \in \{0.3, 0.7\}$
Generation (hosted)	
Access / decoding	Azure API; T=0, except GPT-5.4 (API default decoding)
Max output tokens	1,024
Claim aligner	
Assignment	Hungarian (one-to-one)
Pair score	$0.65\cdot$F1 $+ 0.35\cdot$containment
Exact-match anchor	score $= 1.0$
Match threshold	$\tau = 0.30$
Learned detector	
Encoders	RoBERTa-base, RoBERTa-large, PubMedBERT, SciBERT
Optimizer	AdamW, learning rate 3×10^{-5}
Epochs	5
Max input length	256 tokens
Loss	class-weighted cross-entropy
LLM-as-judge (reference set)	
Model / decoding	Llama-3.3-70B, greedy
Max new tokens	400; 128 for the identity stratum
Batching	none; left-padded batches of 8 for the identity stratum
Severity scale	integer 0–5
LLM-as-judge (directional)	
Model / decoding	Llama-3.3-70B, greedy, unbatched
Max new tokens	300

D Contextual Filter Ablation

The headline pipeline uses the raw lexicon match as $\mathcal{H}(S)$: on the 192 seed-123 clinical notes this yields 15.8 markers per document, the same figure reported in §J.4. Applying the contextual filter of Appendix A instead (retaining only candidates that co-occur with a domain claim signal) leaves 6.0 per document, a 62% reduction; the removed candidates are predominantly severity and temporal terms ("chronic," "new," "acute") matched outside a clinical finding. Because the filter shrinks the denominator, it moves EFS in the same direction as the coherent certainty-bearing restriction of Ap-

pendix J.4, which removes the same two categories. The filter also has far lower recall on BioScope than the unfiltered extractor (Appendix J.1), so we do not adopt it.

E Length-Control Robustness

Omission persists at controlled output length. We extract per-generation output lengths (in word counts) from the raw clinical generations (n = 6,336 no-mitigation generations, after filtering to non-zero source-hedge counts), bin into output-length quintiles, and report mean *omission per source hedge* within each (Table A4). Output length is significantly but weakly associated with omission rate (ordinary least squares, OLS, on log output words, controlling for model identity: coefficient $= -0.060$, 95% CI $[-0.074, -0.045]$, $p = 9 \times 10^{-16}$, $R^2 = 0.035$). At the *longest* quintile (Q5; mean 581 output words) every model still omits between 0.245 and 0.332 source hedges per hedge present.

Table A4: Omission per source hedge by output-length quintile, pooled across 11 open-weight clinical models, no-mitigation runs ($n = 6{,}336$). Generations with zero omissions are included in the mean.

Quintile	n	Mean words	Omission / hedge	EFS
Q1 (shortest)	1,272	224	0.377	0.473
Q2	1,274	339	0.320	0.441
Q3	1,256	390	0.313	0.418
Q4	1,273	441	0.292	0.425
Q5 (longest)	1,261	581	0.290	0.455

Erosion without omission. Excluding every omission event, EFS is 0.747 (clinical, all 11 models, seed 123), 0.574 (legal) and 0.620 (scientific), against 0.443, 0.359 and 0.427 with omission included, and at least one escalation occurs in 57%, 38% and 70% of outputs respectively. Erosion therefore does not reduce to dropped claims.

Hosted-model legal outputs and the 1,024-token cap. Legal outputs often reach the 1,024-token cap (414 of 592 for Mistral-Large-3, 196 of 590 for GPT-5.4, 73 of 588 for GPT-4o, none for Phi-4), and capped GPT-5.4 and GPT-4o outputs score lower than uncapped ones (0.100 vs. 0.251 and 0.267 vs. 0.353), so part of the legal gap in Table 2 reflects truncation.

Phi-4's higher mean EFS is not a length artifact. Among the four hosted models, Phi-4 actually produces *shorter* mean outputs (348 completion tokens) than two of the other three, Mistral-Large-3 (456) and GPT-5.4 (399), so longer outputs do not explain its higher EFS. Restricting to a controlled length band of 200–600 completion tokens (n = 2,937) preserves the ranking: Phi-4 mean EFS $= 0.487$ vs. 0.396 for the other three combined (Mann–Whitney U = 979,855, $p = 1.6 \times 10^{-11}$, two-sided). An OLS on the 6,971 hosted-model generations with non-empty completions, with log completion tokens, model, and domain as predictors, yields a small length coefficient (-0.076, 95% CI $[-0.103, -0.049]$; full-model $R^2 = 0.056$); log length alone explains 2.6% of EFS variance.

F Robustness and Sensitivity

F.1 Severity-Weighted EFS

Using fixed severity weights (fabrication 3.0, escalation 2.5, severity warp 2.0, attenuation 1.5, scope shift 1.0, omission 1.0, normalized to mean 1), EFS_w correlates with EFS at $r = 0.874$–0.893 per generation within each domain (pooled $r = 0.875$ over the 8,640 cross-domain generations), and at $r = 0.992$ across the 11 clinical per-model means; the within-domain model ranking is essentially unchanged (Spearman 0.982 clinical, 1.000 legal and scientific). Domain means are 0.531 (clinical, all 11 models), 0.378 (legal), 0.411 (scientific).

F.2 Metric Sensitivity to De-Hedging

A controlled manipulation shows EFS responds to certainty changes while n-gram overlap does not: mechanically stripping hedges drops EFS from 1.000 to 0.268 while ROUGE-L stays at 0.994; removing unrelated content instead drops both (EFS $= 0.757$, ROUGE-L $= 0.676$).

F.3 Diagnostic Post-Hoc Hedge Restoration

A post-hoc restoration step that re-inserts dropped source hedges raises EFS from 0.493 to 0.832 and reduces detected distortions by 66% (n = 200; 0.458 to 0.732 and -50% under the greedy baseline aligner of Appendix J.5). Because re-inserting markers directly targets what EFS measures, this experiment shows only that a deterministic post-processing step can recover a substantial portion of the measured loss; it is a diagnostic intervention, not evidence about the causal origin of erosion.

F.4 Compression-Controlled Human-vs-Model Comparison

In a regression of hedge hardening (escalation plus severity warp per source marker) on compression ratio (summary words divided by source words) and a model/human indicator (identical specification in both domains), the legal model coefficient is positive but not significant at the 0.05 level (0.025, $p = 0.057$, 95% CI $[-0.001, 0.050]$). Legal is the domain in which the control is identified, because human and model compression distributions overlap substantially: 180 human and 1,152 model summaries fall in the common band, where hardening is 0.082 for humans versus 0.107 for models (one-sided Mann–Whitney $p = 0.024$). The scientific coefficient is larger (0.117, $p = 1.6 \times 10^{-11}$), but we do not treat it as a compression-controlled estimate: human scientific summaries are far shorter relative to their source (mean ratio 0.021) than model summaries (mean 0.317), so the two distributions are nearly disjoint and only four human summaries fall in the common band.

F.5 Downstream Decision Study: What It Does and Does Not Show

Reading source versus eroded output, a Claude Sonnet 4.6 clinician-proxy (API-default sampling) changes its decision in 82.7% of cases ($n = 150$; urgency 65%, recommended action 46%), against a 27.3% identity-pair floor obtained by querying the same proxy twice on the same text (McNemar exact, paired, $p = 1.1 \times 10^{-20}$).

Stratifying a fresh sample across the full EFS range and querying a locally run Llama-3.3-70B proxy (greedy decoding), we find that decision change does *not* track EFS: point-biserial $r = 0.010$ ($p = 0.908$, $n = 150$, EFS range 0 to 1), and the change rate is essentially flat across five equal-width EFS bins (70.8%, 75.0%, 82.1%, 73.1%, 75.0%). The 82.7% therefore measures the consequence of *reading a model summary instead of the source document*, which conflates summarization with erosion. We do not present it as evidence that erosion is decision-relevant.

F.6 De-Hedging with the Remaining Text Held Fixed

The intervention removes only spans identified by the extractor while leaving the remaining text unchanged, thereby isolating the contribution of detected qualifier spans from broader summarization changes. Because the extractor also includes some severity, temporal, and negation markers, this intervention should not be interpreted as isolating pure epistemic certainty in every case. On 185 hedge-rich notes, the de-hedged arm changes at least one clinical decision in 15.7% of notes, against a 1.1% identity floor measured by scoring the same notes twice under a reversed batch ordering. The excess is 14.6 points (bootstrap 95% CI $[9.2, 20.0]$, 5,000 resamples; McNemar exact, discordant counts $b = 28$, $c = 1$, $p = 1.1 \times 10^{-7}$). This study uses the Llama-3.3-70B proxy with greedy decoding and fixed-length padding, because with left padding its logits depend on batch position and the same text scored twice disagrees 12.5% of the time.

F.7 Preliminary Multilingual Probe

When a model summarizes an English Cochrane review abstract excerpt (Devaraj et al., 2021) into English, Spanish, or French, we measure hedge-density retention, the hedge density of the summary divided by the hedge density of the English source, using small parallel epistemic lexicons of comparable size (English 17, Spanish 19, French 16 patterns). Within each model, retention is lower outside English: Gemma-4-E4B scores 0.955 (English), 0.531 (Spanish), 0.411 (French), and Llama-3.1-8B scores 1.541 (English), 0.562 (Spanish), 0.434 (French). Llama's English retention exceeds 1 because it inserts additional hedges relative to the source rather than dropping them; the informative quantity is the within-model drop, which holds the generator fixed. Because lexical coverage is not controlled across languages, we treat this as a preliminary multilingual signal rather than a full replication of the main EFS analysis.

G Mitigation Results

Tables A5 and A6 give the per-strategy results referenced in §6.

H Cross-Architecture Mitigation

To test whether the null prompting effect on Llama (§6, Appendix G) generalizes across architectures, we re-run the three prompting strategies (instruction, few-shot, post-hoc verification) on **Mistral-7B-Instruct** and **Gemma-4-E4B**, two architecturally distinct general-purpose models. Episodes, tasks, and prompt content match the main mitigation protocol ($n = 576$ generations per cell).

Table A5: Mitigation results ($n = 1{,}152$ per prompting strategy; $n = 150$ per temperature setting). EFS is mean $\pm$ std over generations; Cohen's d and two-sided Mann–Whitney p are against the same 1,152-generation baseline. The temperature runs use a 50-episode subset and a 1,024-token cap, whereas the baseline pools both models over all 192 episodes with a 2,048-token cap, so that comparison is unpaired and the small positive Δ should not be read as a temperature effect.

Strategy	EFS	ΔEFS	d	p
Baseline (reference)	$.477 \pm .217$	-	-	-
Instruction	$.474 \pm .216$	$-.002$	$-.011$	.959
Few-shot	$.471 \pm .216$	$-.006$	$-.026$	.719
Post-hoc	$.480 \pm .214$	$+.004$	$+.017$	.653
T=0.3	$.517 \pm .170$	$+.040$	$+.189$	.041
T=0.7	$.508 \pm .201$	$+.032$	$+.147$	.057

Table A6: Four additional prompting strategies, pooled over Llama-3.1-8B, Med42-v2-8B and Mistral-7B-Instruct, against the same-model no-mitigation baseline (EFS $= .478$, $n = 1{,}728$). p from two-sided Mann–Whitney U; none is significant under this unpaired test; the episode-paired re-analysis is in Appendix B.

Strategy	n	EFS	ΔEFS	p
Strict verbatim-copy	1,728	.469	$-.008$	.332
Reason-first self-review	1,728	.478	$+.001$	.631
Few-shot + post-hoc	1,728	.486	$+.009$	.213
Hard-constraint framing	1,728	.467	$-.010$	.231

Results are reported in Table A7.

Mistral shows similarly small prompting effects. All three strategies leave Mistral-7B's mean EFS essentially unchanged ($|\Delta\text{EFS}| \leq 0.009$, all $p > 0.490$).

Gemma shows small-to-moderate positive effects. On Gemma-4-E4B all three strategies produce statistically significant, small-to-moderate improvements (instruction $d = 0.278$, few-shot $d = 0.379$, post-hoc $d = 0.190$; all $p < 0.01$). The best strategy (few-shot, EFS $= 0.501$) lifts Gemma from a below-median baseline up to the level of the highest-EFS open-weight baseline (BioMistral, seed-123 baseline EFS $= 0.492$), but EFS remains far below the identity value of 1.

I Pipeline Validation

Balanced reference set. We construct a reference set of 484 pairs in two strata: (i) 242 **positives**: outputs flagged by the pipeline (greedy baseline aligner) with ≥ 3 distortions and EFS ≤ 0.4, stratified across the 11 open-weight models; and (ii) 242 **source-self pairs** built from the same number of notes. Positives are pipeline-defined and therefore not independent of the system under test; the comparison against the judge on positives measures *convergence on flagrant cases* and *ranking under shared severity criteria*, not unbiased accuracy. Llama-3.3-70B-Instruct serves as the judge, receiving no pipeline labels; 22 of the 242 positives are Llama-3.3-70B generations, so the judge scores its own model's outputs on that subset. Each pair is scored on an integer 0–5 severity scale; we treat score ≥ 1 as a positive binary call to match the pipeline's distortion-count ≥ 1 criterion.

Table A7: Cross-architecture mitigation results ($n = 576$ per cell). ΔEFS vs. same-model no-mitigation baseline; effect size is Cohen's d; p from two-sided Mann–Whitney U. $^{**}p < 0.01$, $^{*}p < 0.05$, ns = not significant.

Model	Strategy	EFS	ΔEFS	d	p
Mistral-7B-Inst.	Baseline	0.480	–	–	–
Mistral-7B-Inst.	Instruction	0.475	-0.005	-0.023	0.780 ns
Mistral-7B-Inst.	Few-shot	0.488	$+0.008$	$+0.038$	0.491 ns
Mistral-7B-Inst.	Post-hoc	0.484	$+0.004$	$+0.019$	0.862 ns
Gemma-4-E4B	Baseline	0.420	–	–	–
Gemma-4-E4B	Instruction	0.480	$+0.061$	$+0.278$	$<10^{-5}$ **
Gemma-4-E4B	Few-shot	0.501	$+0.082$	$+0.379$	$<10^{-8}$ **
Gemma-4-E4B	Post-hoc	0.461	$+0.042$	$+0.190$	0.002**

Truncation stress test and identity control. The judge also flags all 242 pipeline-selected positives. For the second stratum we report two constructions. In the *truncation stress test*, the source is the full note and the "output" is only the note's first 3,000 characters. Because the notes average 10,210 characters, that "output" retains just 24.7% of the source's hedges, and 236 of the 242 pairs drop at least one hedge relative to the 6,000-character source excerpt shown to the judge. These pairs are *ambiguous* rather than gold-negative: the surviving text is a verbatim prefix, so the certainty of what remains is untouched, yet most of the note's epistemically marked claims are absent. Detectors consequently disagree about them by 88 points (the pipeline with the greedy baseline aligner flags 89.7% on the same excerpt, the Llama-3.3-70B judge 57.4%, and a Claude Sonnet 4.6 judge[5] 1.7% of a 60-pair subsample), so no false-positive rate computed on this stratum is interpretable. Specificity is therefore measured on a true *identity control* on the same 242 notes, in which source and output are byte-identical and the pipeline's labels are

[5] Anthropic, `claude-sonnet-4-6`; system card: https://www.anthropic.com/claude-sonnet-4-6-system-card.

computed. Under the greedy baseline aligner the identity false-positive rate is substantial and grows with source length (Table A8, Appendix J.5); every such flag is a claim-alignment failure, since on identical text a distortion can arise only from a mispairing (Appendix J.5). The one-to-one Hungarian aligner used throughout this paper removes these false positives (Table A8). Both judges are specific on the identity stratum. Llama-3.3-70B flags $1/242 = 0.4\%$ of true identity pairs (Wilson 95% CI $[0.1, 2.3]$; zero parse failures, and all 242 pairs rated "same"), and the single flagged pair reports no distortion in its own free-text field. A Claude judge flags $0/242$ (Wilson 95% CI $[0.0, 1.6]$), and 0.5% when matched at the 6,000-character source length the prompt itself uses, where the greedy baseline aligner flags 50.5%. These near-zero rates are not a constant response: under a positive control with the identical prompt the Claude judge flags 98.3% of real eroded generations (mean severity 2.833, with 60 distinct written explanations) and 60.0% of mechanically de-hedged outputs. Among the 242 positives, pipeline distortion count and judge 0–5 severity correlate at $r = 0.186$ ($p = 0.004$), indicating partial but not strong agreement on ranking among real generations.

We recommend identity-pair controls, with the caveat that a control is only as good as its construction: truncating the output converts intended gold-negatives into genuine omissions and can invert the conclusion drawn from them. The identity control is informative about rule-based detection: a greedy claim aligner has a substantial identity false-positive rate that grows with source length, and any EFS computed with such an aligner must be read against that measured ceiling.

The prompt used for this balanced-set validation judge is shown below.

Balanced-Set Validation Judge Prompt

You are an expert evaluator assessing whether an LLM output faithfully preserves the uncertainty expressed in a source clinical note.

Source text (excerpt): `{source_excerpt}`
LLM output: `{output_text}`

Rate on the scales provided and respond ONLY with a JSON object with fields: `uncertainty_distortion_score` (integer 0 to 5, where 0 is perfect preservation and 5 is severe distortion), `hedges_lost_count` (number of uncertainty markers dropped or changed), `certainty_shift` (one of more_certain, same, less_certain, or mixed), and `worst_example` (one sentence describing the worst distortion, or none if faithful).

J Independent Validation

This appendix collects validations that do not use the pipeline's own labels, against human annotations, controlled ground truth, and LLM judges, together with aligner and marker-set ablations.

J.1 Human-Gold Validation of the Hedge Extractor

We evaluate the extractor against three independently annotated corpora.

BioScope, all extractors. On the 11,871-sentence, human-annotated abstracts sub-corpus of BioScope (Vincze et al., 2008; Farkas et al., 2010), the *scientific* extractor reaches F1 $= 0.559$ (precision $= 0.628$). Evaluated on the same sentences, the clinical extractor that computes clinical EFS reaches F1 $= 0.395$ (precision $= 0.533$, recall $= 0.313$), and the contextual clinical variant reaches F1 $= 0.082$ (precision $= 0.405$, recall $= 0.046$). All three are zero-shot transfers: the extractors are narrow epistemic lexicons applied to BioScope without tuning, whereas the best supervised systems in the CoNLL-2010 shared task score substantially higher on biomedical articles (Farkas et al., 2010). BioScope's speculation inventory is also broader than our epistemic definition, annotating experimental-purpose speculation ("to determine whether") that carries no claim-level certainty. We make no directional claim about how the transfer gap propagates into EFS: false-positive cues enlarge $|\mathcal{H}(S)|$ and raise EFS, missed cues shrink it and lower EFS, so the net effect is not identified. BioScope's own clinical subcorpus cannot help here: it is distributed fully anonymized (every token replaced by "*"), and the source corpus needed to restore the text is no longer online.

NUBes, human gold in clinical text. We therefore validate against NUBes and IULA+ (Lima Lopez et al., 2020), the latter a re-annotation of the IULA Spanish Clinical Record Corpus (Marimon et al., 2017): publicly released anonymized Spanish health records annotated by humans with uncertainty cues. We use all 615 released annotation files, of which 355 contain at least one uncertainty cue (2,795 cues, 321 distinct surface forms). We use five random $50/50$ document splits; any data-driven lexicon is derived on the training half and all scoring is done on the held-out half. Under this protocol our hand-built 19-pattern Spanish lexicon recovers 0.448 ± 0.019 of human cues, with sentence-level precision 0.824, recall 0.487, F1 0.612 (mean $\pm$ std over splits; worst split 0.420 cue recall). Expanding the lexicon does not help: a data-driven 100-form lexicon raises cue recall to 0.909 ± 0.008 but collapses precision to 0.250

Table A8: Stratified judge-validation results on the $n = 484$ set. FPR = false-positive rate; 3k and 12k = source truncated to 3,000 and 12,000 characters; 70B = the Llama-3.3-70B judge; ceiling = mean EFS on identical text; detector = the learned detector; Claude = the Claude Sonnet 4.6 judge.

Stratum	n	Pipeline	Judge	Notes
Truncation stress test (not a specificity test)				
Truncation pairs	242	flags 89.7%	139/242	97.5% genuinely lose a hedge
Identity control (specificity)				
True identity, greedy aligner	242	FPR 27.7% @ 3k	0.4% (70B), 0.0% (Claude)	18.8% @ 3k, 66.7% @ 12k on the 192-doc set; ceiling 0.864
True identity, Hungarian	242	FPR 0.0%	0.4% (70B), 0.0% (Claude)	ceiling 1.000; detector 0.0%
Pearson r: pipeline count vs judge severity				
Positives only	242	$+0.186$ ($p = 0.004$)		real generations

and F1 to 0.394, because the frequent gold forms include generic words.

RadGraph, human gold in English clinical text. We further validate against RadGraph (Jain et al., 2021b), using only the 600 chest-radiograph reports annotated by board-certified radiologists and excluding the 221,263 reports labeled automatically by a model. A sentence is gold-uncertain when a radiologist marked an observation as uncertain. Over 3,622 observation-bearing sentences, the clinical extractor, restricted to spans with nonzero epistemic strength, reaches precision 0.570, recall 0.609, F1 0.588. Two radiologists independently annotate the test split and agree on 98.9% of its 1,011 sentences (Cohen's $\kappa = 0.879$; one labeler's uncertain sentences are a strict subset of the other's, giving an inter-radiologist F1 of 0.884). The extractor scores F1 0.606 and 0.691 against them respectively, so its error is substantially *larger* than the disagreement between two experts. This matters for what EFS counts. Scored against *all* spans the extractor returns, precision on those same 3,622 sentences collapses to 0.170, because the lexicon also emits severity and temporal modifiers ("mild", "small", "stable", "acute", "new") that carry epistemic strength zero and that radiologists do not treat as uncertainty. Spans with epistemic strength zero (the severity and temporal categories plus recommendation, evidential and negation markers outside the 4-point scale) are 84.3% of all lexicon spans in the 192 seed-123 MIMIC notes; the two named categories alone are 74.7% (§J.4). All of them enter the EFS denominator $|\mathcal{H}(S)|$. EFS should therefore be read as a score derived from a distortion rate normalized by the source's full *marker set*, of which the epistemic subset is a minority; restricting to that subset raises precision to 0.570.

J.2 Human-Grounded Erosion on RadGraph (Lexicon-Blind)

The extractor validation above asks whether our lexicon finds what radiologists marked. This subsection asks the paper's question with the radiologists' marks as gold. From the 600 radiologist-annotated RadGraph reports we take 200 chest-radiograph reports and every observation a radiologist labeled uncertain, 680 per model, and summarize each report with four open-weight models (BioMistral-7B, Llama-3.1-8B, Med42-v2-8B, Mistral-7B). A judge (Llama-3.3-70B) that never sees our lexicon, EFS, or any pipeline label is shown the report and a summary and labels each gold observation PRESERVED, HARDENED (stated with more certainty) or OMITTED; 2,720 (model, observation) judgments result. As a control, every seventh judgment item (389 observations from 161 reports) is additionally scored with the source report itself in place of the summary, giving a paired source-arm versus model-arm comparison on identical observations (Table A9).

Over all 2,720 model-arm judgments the human-grounded survival rate (share of gold-uncertain observations preserved) is 0.510, report-clustered 95% CI $[0.465, 0.555]$; per model, BioMistral-7B 0.512, Llama-3.1-8B 0.540, Med42-v2-8B 0.513, Mistral-7B 0.476 ($n = 680$ each). Every model's paired drop is significant (McNemar $p \leq 2 \times 10^{-4}$). Three caveats. (i) Observations cluster within 200 reports $\times$ 4 models, so per-observation Wilson intervals are anticonservative; we therefore use report-clustered bootstrap intervals throughout. (ii) The control arm shows the judge the whole report (mean 795 characters) whereas the model arm shows a summary (mean 367), so part of the preservation drop is compression; the mention-conditional hardening rate removes omissions from the numerator and denominator and therefore reduces the direct influence of compression (Ta-

Table A9: Human-grounded survival of radiologist-marked uncertainty, paired design (same observations judged against the source report and against a model summary). The primary rows use the full paired set; the secondary rows exclude observations that the source report also states *definitely* (present or absent) elsewhere ("def-dup"). The last row is restricted to observations mentioned in each arm, so n gives the source-arm and model-arm counts and the final cell gives the rate ratio with its CI. p: exact McNemar. CI: report-clustered bootstrap (10,000 resamples) on the difference in percentage points.

Quantity	**Set**	n	**Source arm**	**Model arm**	Δ (pts)	p; clustered 95% CI
Preserved	primary (all)	389	83.8%	54.5%	−29.3	1.2×10^{-25}; [−35.3, −23.3]
Preserved	secondary (def-dup excl.)	327	89.0%	57.5%	−31.5	6.4×10^{-25}; [−38.0, −25.0]
Hardened	primary	389	12.3%	30.8%	+18.5	1.3×10^{-15}; [13.4, 23.5]
Hardened	secondary	327	6.4%	25.4%	+19.0	6.4×10^{-15}; [13.8, 24.2]
Hardened | mentioned	primary	374 / 332	12.8%	36.1%	+23.3	ratio 2.816, CI [2.130, 3.954]; Δ CI [17.5, 29.2]

ble A9). Restricting further to the 329 observations that both arms mention, so that the comparison is fully paired, hardening is 14.3% versus 36.5% (exact McNemar $p = 2.5 \times 10^{-16}$). (iii) The judge is Llama-3.3-70B and two of the four judged models (Llama-3.1-8B, Med42-v2-8B) are Llama-family; the paired design cancels any constant per-model judge offset within a pair, but no cross-family judge replication exists for this experiment. The control arm also measures the judge's own error: it labels 3.9% of gold-uncertain observations omitted when shown the source itself.

J.3 Measure Properties and Discriminance

On 60 clinical notes (16.9 lexicon markers per note) we test EFS against pre-specified properties. *Identity*: $\text{EFS}(s, s) = 1$ on all 60 notes (the greedy baseline aligner described below scores 0.823 with an 80.0% document false-positive rate on the same notes). *Local monotonicity*: deleting source hedges one at a time never raises EFS (0 violations in 919 deletion steps; mean EFS after all deletions 0.340). *Scale invariance*: duplicating both source and output leaves EFS unchanged (maximum $|\Delta| = 0$). *Sensitivity to hedged fabrication*: grafting an epistemically marked claim from another note lowers EFS in 59/60 notes (1.000 → 0.907). One scope property is negative and we state it: an invented *certain* assertion is charged in only 1/60 notes, because fabrication requires the unmatched output claim to carry a hedge; EFS scores fidelity to expressed uncertainty and is blind to invented certainty. Table A10 compares EFS with a raw hedge counter and ROUGE-L under controlled corruptions of a second 60-note sample drawn by the same procedure (≥ 5 markers, 19.1 per note; 53 notes are shared with the sample above). EFS is the only measure that responds to relocating a hedge to a different claim and that penalizes added or fabricated hedges, which the counter rewards as "better than perfect"; it is nearly invariant to sentence reordering (0.998), on which ROUGE-L collapses. On pure deletion the counter tracks the ground-truth fraction more linearly than EFS (Spearman −0.988 vs. −0.897; ROUGE-L −0.864).

Floor, clamp and unclamped score. Distortions are charged per claim, more than one may fall on the same claim, and 39% of markers occur in sentences that never form a claim and can never be charged; at $\text{EFS} = 0$ a mean of 68% of markers are implicated, not all of them. The floor is reached on 9.0% (clinical), 12.8% (scientific), and 34.7% (legal) of generations; the clamp is strictly binding ($|\mathcal{D}| > |\mathcal{H}(S)|$) on 4.8%, 5.2% and 20.7% respectively, the last driven by heavily eroded complaints that accumulate many distortions per source marker. Reporting the *unclamped* score lowers the domain means (clinical $0.443 \to 0.422$, scientific $0.427 \to 0.411$, legal $0.359 \to 0.209$) but leaves the within-domain model ordering essentially unchanged (rank Spearman 0.964 clinical, 1.000 legal and scientific); we clamp to $[0, 1]$ so that EFS remains bounded and comparable across documents.

Table A10: Mean value of each measure under controlled corruptions of 60 clinical notes. Hedge count is the output/source marker-count ratio. A faithful measure should read 1 on identity and reorder and fall below 1 elsewhere.

Corruption	EFS	Hedge count	ROUGE-L
Identity	1.000	1.000	1.000
Reorder sentences	0.998	1.000	0.334
Delete 50% of hedges	0.670	0.498	0.996
Relocate hedge to another claim	0.806	1.000	0.997
Add a hedge	0.935	1.218	0.998
Fabricate a marked claim	0.931	1.070	0.993

J.4 Robustness: What EFS Counts as a Marker

The clinical lexicon's severity and temporal categories are not treated as uncertainty by radiologists

(RadGraph validation, §J.1). Their inclusion is by design: the lexicon is organized into named categories, two of which are `severity_modifier` (mild, moderate, severe) and `temporal_hedge` (acute, subacute, chronic). Together they are 74.7% of detected clinical spans on the same 192 notes, and they enter both $|\mathcal{H}(S)|$ and the definition of an epistemically marked claim. The scientific lexicon has an analogous `quantitative_hedge` category (approximately, marginal, substantial), 44% of its spans; the legal lexicon has no degree or time category at all.

We therefore recompute the entire study with those categories removed and every certainty-bearing category retained (diagnostic, evidential, recommendation, negation for clinical; results, scope, limitation, evidential for scientific; all for legal). The restriction is applied coherently to the denominator, the claim extractor and each domain's hedge finder. The mean clinical marker count falls from 20.5 to 6.2 per document (three-seed means).

Clinical EFS falls from 0.450 to 0.366 over the three-seed depth study. The model ranking is correlated with the unrestricted metric (Spearman 0.873, $p < 10^{-3}$), with BioMistral-7B still the highest-EFS open-weight model and MedGemma-27B the lowest. The between-domain ordering is preserved (for the three matched models, clinical 0.456, scientific 0.415, legal 0.359; Kruskal–Wallis $H = 44.8$, $p = 1.9\times10^{-10}$), and omission accounts for 60.3% of distortions. Escalation increases from 11.7% to 21.3% of detected distortions, because severity-driven distortions are no longer counted. Cross-seed variability rises (maximum per-model standard deviation 0.020 to 0.067), as expected when the denominator shrinks and episode-level hedge density varies: the seed-0 and seed-1 episode samples carry 23.5 and 22.4 detected spans per document against 15.8 for seed 123.

We keep EFS over the full marker set as the headline metric, because it is what the released pipeline computes, and report its composition above.

J.5 Aligner Design: Greedy versus One-to-One Assignment

The identity-pair control above motivates the aligner design, which we document as an ablation against a greedy baseline. The greedy baseline aligner scores each candidate output claim by the *raw count* of shared domain terms and takes the first argmax, allowing many source claims to map onto one output claim:

```
overlap = len(s_terms & o_terms)
if overlap > best_overlap: best =
o_claim
```

An epistemically marked source claim (for instance, a history sentence stating that an event was *likely* due to a named condition) therefore ties with, and loses to, a Past Medical History list repeating the same diagnoses. Feeding this aligner identical text produces a 66.7% document-level false-positive rate at the 12,000-character truncation used throughout, growing with source length (18.8% at 3,000 characters, 50.5% at 6,000; the floor by distortion type is in Appendix J.6); all 452 spurious distortions have a source span distinct from their output span.

The aligner of §4 (Table A3) instead anchors identical normalized text at similarity 1.0, so it cannot be mis-paired, and treats a source claim below the threshold as a genuine omission. Extraction, hedging, epistemic strength, severity levels and the distortion-typing rules are identical in the two variants, with one exception: the greedy clinical scorer also applies the loose fabrication detector of Appendix A, which the Hungarian clinical pipeline does not, so the greedy clinical figures below include loose fabrications.

The Hungarian aligner satisfies $\text{EFS}(s, s) = 1$ on all 1,728 document, length and domain combinations tested, as does the learned detector on the same pairs. Sensitivity is preserved (mechanically de-hedged text scores 0.268, against 0.270 under the greedy baseline) and the controlled ground-truth correlation strengthens from $\rho = -0.721$ (greedy baseline) to $\rho = -0.891$ (three seeds each; std 0.009 and 0.006), measured under the identical 40-document protocol. The main patterns persist under both aligners. On the three-seed depth study (n = 18,480, the basis used for all model-level claims) clinical EFS moves from 0.416 (greedy) to 0.450 (Hungarian). On the single-run cross-domain study, where legal and scientific were generated with one seed and clinical is therefore matched to seed 123, EFS moves from the greedy values in the Observed column of Table A11 to 0.443 (clinical), 0.359 (legal) and 0.427 (scientific); between the two aligners, open-weight model ranking Spearman is 0.936 (three-seed means, n = 18,480) and hosted-model ranking Spearman 1.000; the highest-EFS open-weight model (BioMistral-7B) and the lowest (MedGemma-27B) are the same under both, as are Phi-4 and GPT-5.4 among hosted models. Omission rises from 43.5% to 66.3% of detected

distortions. On 1,000 generations drawn uniformly at random from the 6,336 seed-123 clinical generations (all 11 models), mean EFS ranges from 0.441 to 0.449 for thresholds in $[0.20, 0.50]$ (0.442 at the operating value $\tau = 0.30$, against 0.443 on the full set), so the score is insensitive to it; the omission share is not (56.3% at $\tau = 0.20$, 61.2% at 0.30, 74.3% at 0.50), so we do not lean on the exact type distribution.

J.6 Type-Dependence of the Greedy Aligner's Identity Floor

With the Hungarian aligner the identity floor is zero for every type, so nothing below affects the reported results; the analysis shows which conclusions are sensitive to alignment quality, given the greedy baseline's 66.7% identity false-positive rate. Under the greedy baseline the floor is strongly type-dependent, and the distinction is structural rather than empirical. Omission requires a source hedge to have no counterpart in the output, and fabrication requires a hedge absent from the source; neither can fire when output equals source. The remaining four types are established by comparing an *aligned pair* of claims, so they inherit the aligner's mispairings. Expressed as a share of each type's mean count among the clinical generations that exhibit that type, the identity floor at operational length is 0.0% for omission, 0.0% for fabrication, 6.7% for attenuation, 12.9% for scope shift, 31.6% for severity warp, and 51.6% for escalation.

Two consequences follow. First, omission and fabrication together account for 74.7% of all distortion instances in the multi-seed depth study and have a zero identity floor, so the largest detected category is not produced by identity-level alignment errors. Second, escalation counts under the greedy baseline are materially inflated, so under that aligner we would not treat escalation as a co-equal driver alongside omission. The floor is not an estimate of inflation in real generations: it is measured with a full-length output, which gives the aligner far more candidate sentences than a real summary does, so it likely overstates the inflation present in actual generations.

Under the greedy baseline, normalizing each domain by its own ceiling gives ceiling-normalized shortfalls ($1 - \text{EFS}/\text{ceiling}$) of 52.0% (clinical), 58.6% (legal) and 49.9% (scientific), and the raw between-domain ordering is not robust: scientific has a lower ceiling than the other two, so clinical and scientific exchange places once that is accounted for. The Hungarian aligner gives every domain an identity ceiling of 1, which removes the confound and makes the raw ordering interpretable (§6.5).

Table A11: Identity ceiling per domain under the greedy baseline aligner (source and output byte-identical, 12,000-character truncation, $n = 192$ documents each; observed EFS also under the greedy baseline). FPR = false-positive rate.

Domain	Identity EFS	Identity FPR	Observed EFS
Clinical	0.864	66.7%	0.414
Legal	0.872	43.2%	0.361
Scientific	0.788	69.8%	0.395

J.7 Controlled Ground-Truth Validation (Primary Validation Anchor)

Here the ground truth is known by construction rather than estimated by a rater, a judge, or the pipeline itself: we programmatically remove a known fraction $f \in \{0, .25, .5, .75, 1\}$ of each document's hedges and test whether EFS tracks f. It does, monotonically (Table A12).

Table A12: Mean EFS by ground-truth fraction of hedges removed (40 clinical documents; means over three random-removal seeds). Spearman $\rho = -0.891$, computed per seed over the 40 documents $\times$ 5 fractions and averaged over the three seeds.

Fraction removed	0.00	0.25	0.50	0.75	1.00
Mean EFS	1.000	0.830	0.644	0.443	0.249

An independent Claude judge's 0 to 5 severity rating also tracks the ground-truth fraction ($\rho = 0.650$).

J.8 Independent Directional Judge

For the directional judgment of §6, 541 of the 750 submitted items returned JSON that the parser accepted; the remaining 209 returned complete JSON wrapped in a Markdown code fence, which the parser rejected, and are excluded from the reported figures. Including them ($n = 750$) gives 97.1% more certain (clinical 98.7%, scientific 94.7%; 2.0% less certain). The prompt is shown below with its output-format instructions condensed.

Relying on an LLM judge raises the question of when it may legitimately be trusted, and our own controls answer it. Asked an *absolute* question about a single pair, an LLM reader is noisy: queried twice on byte-identical text at the API's

default sampling temperature, the Claude clinician-proxy of §F changed at least one clinical decision in 27.3% of cases. Asked a *directional* question, which of two texts expresses greater certainty, its answers agree with the RadGraph and controlled-removal evidence. We therefore use the judge for directional or span-anchored questions, not as a substitute for EFS in estimating the rate or magnitude of distortion, and we anchor quantitative claims about EFS's sensitivity on the controlled ground truth ($\rho = -0.891$). The judge's identity false-positive rate is measured separately on byte-identical pairs (§I).

Independent Directional Judge Prompt

You are an expert evaluator of how faithfully a summary preserves the CERTAINTY of its source. You are given a SOURCE passage and an OUTPUT derived from it.

Judge ONLY epistemic certainty (how confidently claims are asserted), not factual correctness, fluency, or completeness. Source authors deliberately hedge (e.g., "possible", "likely", "suggests", "may", "cannot rule out") or assert firmly; a faithful output keeps each claim at the SAME confidence level as the source.

Source: `{source}`
Output: `{output}`

Compared to the SOURCE, how does the OUTPUT present the certainty of its claims overall? Respond with ONLY a JSON object with fields `certainty_shift` (one of more_certain, less_certain, same, or mixed), `n_hedges_dropped`, `n_hedges_added`, `confidence`, and `worst_example`.

J.9 Multi-Annotator Agreement (Human and AI)

Three human annotators plus four independent LLMs (Claude Sonnet 4.6, Llama-3.3-70B, gpt-oss-20B (OpenAI, 2025), and Gemma-4-31B) label the same six-type schema. The vs-pipeline column uses the 164 pairs the LLM annotators labeled: 64 legal and scientific pairs, including the 20 of the human pilot, and 100 clinical pairs. On the decision-relevant question of whether any distortion is present, agreement is high for three of the four LLM annotators under a prevalence-robust coefficient (Table A13; Gwet's AC1, Gwet, 2008, a chance-corrected agreement coefficient on the same scale as κ that remains stable when one label is highly prevalent; Cohen's κ, Cohen, 1960, is degenerate under the skewed prevalence here).

The two strongest independent models (Llama-3.3-70B and Claude) agree with each other at AC1 $= 0.948$. Fine-grained six-way typing remains hard for all annotators: agreement with the human majority is lowest on omission for three of the four models, as it was for the human raters, and on fabrication for Llama-3.3-70B.

Table A13: Agreement on whether any distortion is present (Gwet's AC1); n is the number of source–output pairs compared. gpt-oss-20B produced no parseable labels for some pairs, so its values are computed on 18 and 139 pairs.

Annotator	vs human (n=20)	vs pipeline (n=164)
Claude Sonnet 4.6	0.947	0.907
Llama-3.3-70B	0.942	0.921
gpt-oss-20B	0.934	0.792
Gemma-4-31B	0.559	0.607

K Pilot Human Annotation Study

We piloted human validation on a stratified sample of 20 source-output pairs across legal and scientific domains, annotated independently by three raters using the six-category typology and an overall ordinal severity rating ({none, low, moderate, high}); the same pairs are labeled by four LLM annotators in Appendix J.9.

Inter-annotator agreement. On the ordinal severity rating, mean pairwise quadratic-weighted Cohen's κ is 0.430 (the three rater pairs: 0.554, 0.543, 0.194; all three raters within ± 1 severity level on 70% of pairs). Per-type Fleiss' κ (Fleiss, 1971) across the three raters is moderate to substantial for fabrication ($\kappa = 0.603$; positive rates 13/14/15 of 20 across raters) and lower for the other five types. Two things drive the lower values. Heavily skewed label prevalences suppress κ through the prevalence paradox (Feinstein and Cicchetti, 1990), and the raters disagree sharply on how often *omission* occurs (19/20 vs. 6/20). The latter points to under-specification in the typology, which §3.2 operationalizes but which does not yet match every rater's implicit criteria. We therefore report raw agreement alongside κ and read the per-type κ of the skewed types as a lower bound on agreement rather than a direct measure of it. On whether *any* certainty distortion is present, agreement is high: raw pairwise agreement is 93% and Gwet's AC1 $= 0.926$.

EFS vs. annotator-rated severity. On the same 20-pair pilot, the Pearson correlation between EFS and the 3-annotator-averaged severity is $r = -0.034$ (bootstrap 95% CI $[-0.523, +0.457]$); Spearman $\rho = -0.070$ ($[-0.554, +0.440]$). The EFS formula, which counts each distortion event with equal weight, may differ structurally from gestalt severity judgments, where annotators implicitly weight fabrication and scope shift more

heavily than omission. EFS is a text-level property (its fabrication labels show moderate-to-substantial inter-annotator agreement, $\kappa = 0.603$ above) that *underdetermines* annotator-perceived severity. A severity-weighted variant with fixed weights (EFS_w, Appendix F) preserves the model rankings but, like EFS, does not track annotator-rated severity at $n = 20$; deriving weights from an independent severity-elicitation panel is future work.

L Cross-Domain Transfer

Table A14 reports cross-domain transfer performance for clinical-trained distortion detectors. Fine-grained 7-class accuracy drops 14–38 points when transferring from clinical to legal or scientific text, indicating that the detector's fine-grained discriminations are substantially domain-specific. Binary separation between typed-distortion instances and the equal-strength class transfers better than fine-grained seven-class classification (legal 55–84%, scientific 75–93%), although the equal-strength class is not a clean no-distortion control (Appendix N). **RoBERTa-large** transfers best on both legal (69.9%) and scientific (73.5%), with the smallest transfer gap (14–18 points).

Table A14: Cross-domain transfer accuracy (%). Models trained on clinical data ($n = 52{,}748$); tested on a random 2,000-example subsample of the clinical validation split, legal ($n = 482$), and scientific ($n = 279$) splits. **Bold**: encoder with the best legal and scientific transfer.

Model	Clinical	Legal	Scientific
7-class accuracy			
PubMedBERT	87.6	49.2	58.8
RoBERTa-base	87.9	64.5	71.0
RoBERTa-large	**87.8**	**69.9**	**73.5**
SciBERT	88.0	63.1	64.5
Binary accuracy (distortion vs. equal strength)			
PubMedBERT	96.6	55.2	74.6
RoBERTa-base	97.0	74.9	87.5
RoBERTa-large	**97.0**	**84.0**	**93.2**
SciBERT	96.9	72.8	80.6

M Token-Distribution Probe of Assertive Preference

The erosion observed across the tested models motivates the hypothesis that generation retains a preference for assertive continuations even when the source is hedged. We probe this hypothesis at the token-distribution level on the same Llama-3.1-8B used in our main results, with a length-matched non-hedge control.

Lexicons (a priori). **Assertive verbs:** {is, has, shows, confirms, reveals, demonstrates, displays, indicates}, expanded to 20 single-token IDs across lowercased / capitalized / leading-space variants. **Hedge markers:** {may, could, might, possibly, appears, seems, suggests, likely}, expanded to 22 single-token IDs.

Prompt and probe position. Each source sentence is placed in the template "*Summarize the following clinical sentence in a single sentence. Sentence: {sentence}. One-sentence summary: The CT*", and we capture the next-token probability distribution at the position immediately after "The CT". The source sentence (with its hedge, in the hedge condition) is in the prompt.

Length-matched conditions ($n = 500$ each). We sample 500 MIMIC sentences containing a hedge marker; for each, we draw at random, without replacement, a non-hedge MIMIC sentence whose length is within $\pm 15\%$ (at least ± 8 characters). Lengths are matched (hedge 108.8 vs. non-hedge 106.7 characters; Mann–Whitney $p = 0.501$).

Result. On length-matched pairs:

- **Non-hedge baseline:** $P(\text{assertive}) = 0.020$, $P(\text{hedge}) = 0.001$; sum ratio $= 24.818\times$, per-token ratio $= 27.299\times$; $499/500$ prompts (99.8%, Wilson 95% CI $[98.9, 100.0]$) place more mass on assertive than hedge tokens.
- **Hedge context:** $P(\text{assertive}) = 0.020$, $P(\text{hedge}) = 0.006$; sum ratio $= 3.449\times$, per-token ratio $= 3.794\times$; $450/500$ prompts (90.0%, Wilson 95% CI $[87.1, 92.3]$).

Treating $1\times$ as no preference, the source hedge reduces the excess assertive advantage from 23.818 to 2.449, a reduction of approximately 89.7%. A residual $3.449\times$ class-level advantage persists when the source hedges.

Interpretation. Under greedy decoding the emitted token is the single most probable token, not a sample from the aggregate class, so a hedge marker can still win the position when its individual probability exceeds that of the top assertive token. This is correlational evidence about the next-token distribution; a stronger causal probe (internal-representation interventions or hedge-balanced fine-tuning with downstream EFS measurement) is future work.

Cross-architecture replication. To check whether the pattern is specific to Llama-3.1-8B, we re-ran the identical length-matched probe ($n = 500$ per condition) on Mistral-7B-Instruct, Med42-v2-8B, and Gemma-3-4B (Gemma Team, 2025) (Table A15). The baseline assertive-class advantage varies widely across architectures (24.818× to 83.080×); the hedge-context reduction is consistently large (89.7–96.9%); and the post-conditioning residual is non-trivial in every model (3.449 to 7.605×, largest for Gemma-3-4B).

Table A15: Cross-architecture replication of the token-distribution probe ($n = 500$ length-matched hedge / non-hedge MIMIC sentences per condition, per model). Baseline = $P(\text{assertive})/P(\text{hedge})$ in length-matched non-hedge contexts; Hedge ctx. = same ratio when the source contains a hedge marker; Reduction = $(\text{baseline}-\text{hedge})/(\text{baseline}-1)$. Gemma uses bfloat16 inference (fp16 underflows on this architecture).

Model	Baseline	Hedge ctx.	Reduction
Llama-3.1-8B-Instruct	24.818×	3.449×	89.7%
Mistral-7B-Instruct-v0.3	83.080×	3.574×	96.9%
Med42-v2-8B (medical fine-tune)	28.696×	3.673×	90.3%
Gemma-3-4B	82.334×	7.605×	91.9%

N Learned Detector: Full Results

Training examples are the distortion events found by the greedy baseline pipeline (Appendix J.5; its clinical scorer includes the loose fabrication detector) on the 6,336 seed-123 clinical generations of the 11 open-weight models, so the detector's numbers are not directly comparable to the EFS values reported elsewhere. There are 62,057 labeled examples (label distribution: omission 40.6%, escalation 15.5%, fabrication 12.6%, scope shift 11.7%, severity warp 8.3%, attenuation 3.4%, equal strength 7.9%). The four encoders are trained with identical hyperparameters (Appendix C) on the same 52,748 training / 9,309 validation split. Per-type metrics in Table A16 are computed on the validation split, a random instance-level 85/15 split that is not document-disjoint: 57.8% of validation inputs also occur in the training set (42.1% with the same label; 99.6% for omission, where the same source sentence is dropped by several models), so validation accuracy overstates generalization to unseen sentences. In addition, 24.5% of distorted aligned pairs carry two labels (always severity warp plus a second type) and appear once under each, and 27% of equal-strength pairs also occur as scope-shift or severity-warp instances, so the equal-strength class is not a clean no-distortion control.

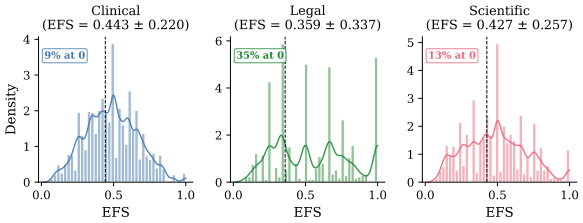


Figure A1: EFS distributions across three professional domains (stratified samples of 192 sources per domain, clinical seed 123; the clinical panel pools all 11 open-weight models, $n = 6{,}336$, while the legal and scientific panels use Llama-3.1-8B, Mistral-7B and Med42-v2-8B, $n = 1{,}152$ each). Histograms and density curves show generations with EFS above 0, and the inset gives the share at the floor of 0; panel titles give mean ± std over all generations, and dashed lines mark the mean; the clinical mean is 0.443 for the 11-model pool (0.478 for the same three models as the other panels).

O Distortion Type Examples

The examples in Table A17 are condensed and lightly paraphrased from patterns observed in Llama-3.1-8B clinical generations rather than quoted verbatim.

P Distortion Type Counts

Table A18 gives the raw counts and per-generation means underlying the percentages in §6 ("Omission Is the Largest Detected Distortion Category").

Q Cross-Domain EFS Distributions

Figure A1 pairs with the cross-domain Kruskal–Wallis test in §6.5 by showing the per-domain EFS distributions. For the same three models, restricting EFS to the four core categories (escalation, attenuation, omission, fabrication) yields 0.392–0.541, compared with 0.359–0.478 under the full taxonomy (0.510 against 0.443 for the full 11-model clinical pool), so the cross-domain pattern is not driven by the two boundary categories. One possible contributor to the wide legal dispersion is that complaint documents mix formulaic procedural language with substantive epistemically marked claims.

R Prompt-Input-Output Example

Figure A2 shows the prompts, a source excerpt, an output in the style of a model hospitalization summary, and the pipeline's result on that pair, so the

Table A16: Learned distortion detector performance (7-class classification, 62,057 total examples; 52,748 training / 9,309 validation); mean ± std across 3 training seeds ($s \in \{42, 123, 2024\}$). Acc. = validation accuracy at the best epoch; macro-F1 and per-type F1 are from the final epoch, on the validation split. F1 = 1.000 on omission and fabrication is expected: pipeline labels for these two types are derived from a structural cue (*source claim absent from output* / *output claim absent from source*) that leaves one side of the input pair empty. The discriminating types are attenuation, escalation, scope shift (Sco.), and severity warp. **Bold** rows are the domain-specific encoders.

Encoder	Acc.	Macro-F1	Omi.	Fab.	Esc.	Att.	Sco.	Sev.W.
RoBERTa-base	88.7 ± 0.2	0.836 ± 0.003	1.000	1.000	0.814	0.853	0.736	0.620
PubMedBERT	$\mathbf{88.8 \pm 0.1}$	$\mathbf{0.835 \pm 0.002}$	**1.000**	**1.000**	**0.817**	**0.863**	**0.735**	**0.612**
SciBERT	$\mathbf{88.8 \pm 0.1}$	$\mathbf{0.838 \pm 0.002}$	**1.000**	**1.000**	**0.817**	**0.864**	**0.738**	**0.615**
RoBERTa-large	88.5 ± 0.1	0.833 ± 0.001	1.000	1.000	0.812	0.848	0.727	0.621

Table A17: Illustrative examples of each distortion type in clinical text. "S:" = source, "O:" = LLM output.

Type	Source → Output
Escalation	S: "CT chest shows findings *suggestive of* pneumonia in the right lower lobe." O: "CT chest *confirms* pneumonia in the right lower lobe."
Attenuation	S: "*Likely* left hip fracture on X-ray imaging." O: "X-ray shows *possible* left hip fracture."
Omission	S: "Worsening renal function is *likely* due to obstruction of the left ureter." O: [claim absent from output]
Severity warp	S: "Echocardiogram shows *mild* aortic stenosis." O: "Echocardiogram shows *significant* aortic stenosis."
Fabrication	S: [not in source] O: "Patient was *possibly* diagnosed with renal insufficiency."
Scope shift	S: "*Mild* stenosis in the left anterior descending artery." O: "*Mild* coronary artery stenosis."

Table A18: Distribution of distortion types across the multi-seed depth study (all 11 models over three seeds and three tasks; $n = 18{,}480$ generations). Mean/Gen = mean count per generation.

Type	Count	%	Mean/Gen
Omission	135,412	66.3	7.327
Escalation	23,979	11.7	1.298
Fabrication	17,205	8.4	0.931
Scope Shift	14,397	7.0	0.779
Severity Warp	9,548	4.7	0.517
Attenuation	3,679	1.8	0.199
Total	**204,220**	**100**	**11.051**

reader can see how an individual generation maps to the per-type distortion counts aggregated in Table A18. Both texts are synthetic, because MIMIC-IV note text may not be reproduced in publications;

System Prompt

"You are a clinical documentation assistant. Generate accurate, professional clinical text based on the provided source notes. Ensure all clinical details are faithfully represented."

Task Prompt (abridged)

"Based on the following clinical notes from a hospitalization episode, generate a structured summary covering: (1) Reason for Admission, (2) Key Findings, (3) Hospital Course, (4) Diagnoses, (5) Discharge Plan."

Source Excerpt (synthetic; 7 markers in bold)

Chest CT showed an ***acute*** right lower lobe consolidation, ***likely*** secondary to aspiration. Another possibility is organizing pneumonia given the ***subacute*** course, although ***less likely*** given the imaging findings. It is ***possible*** that this represents a post-obstructive pneumonia from an endobronchial lesion. It is ***likely*** that the pleural effusion was a consequence of the pneumonia. Pulmonology was consulted for evaluation of ***possible*** underlying interstitial lung disease.

LLM Output (synthetic; style of a model hospitalization summary)

+ Chest CT: showed an ***acute*** right lower lobe consolidation, ***likely*** secondary to aspiration. * Pulmonology was consulted for evaluation of ***possible*** underlying interstitial lung disease (ILD). * Working Diagnoses: + Aspiration pneumonia + Post-obstructive pneumonia from an endobronchial lesion.

Pipeline: $|\mathcal{H}(S)| = 7$, $|\mathcal{D}| = 3$ (2 omissions, 1 escalation), EFS $= 1 - 3/7 = 0.571$

Figure A2: Worked prompt-input-output example (hospitalization-summary prompt). The pipeline charges an *omission* for the organizing-pneumonia sentence ("less likely"), which has no aligned output sentence; an *escalation* for "*possible* . . . post-obstructive pneumonia", which the output lists as a working diagnosis with no marker; and an *omission* for the "*likely*" pleural-effusion sentence. The CT and consultation sentences keep their markers and are not charged.

they mirror patterns observed in the breadth study, and the counts shown are the released pipeline's output on exactly this pair.

S Prompts

We reproduce the generation and mitigation prompts used for the open-weight models below; the hosted-model prompts (§5) are in the code re-

lease. Each domain uses a system prompt followed by task-specific user prompts. `{source_text}` is replaced with the actual input document at generation time. For models without a system role (the Gemma-4 models, MedGemma-27B, BioMistral-7B) the system prompt is prepended to the user turn.

System Prompt (Baseline)

"You are a clinical documentation assistant. Generate accurate, professional clinical text based on the provided source notes. Ensure all clinical details are faithfully represented."

Task A: Hospitalization Summary

Based on the following clinical notes from a hospitalization episode, generate a structured summary covering:
1. **Reason for Admission**: Chief complaint and presenting symptoms
2. **Key Findings**: Relevant diagnostic results (labs, imaging, procedures)
3. **Hospital Course**: Major events, consultations, and treatment decisions
4. **Diagnoses**: Final and working diagnoses at discharge
5. **Discharge Plan**: Follow-up appointments, pending results, and instructions

---

Source Clinical Notes:
`{source_text}`

Generate the hospitalization summary now.

Task B: Referral Letter

Based on the following clinical notes, generate a specialist referral letter to `{referral_specialty}`.
The letter should include:
1. **Patient Summary**: Brief demographic and clinical context
2. **Reason for Referral**: Why this specialist consultation is being requested
3. **Relevant Clinical History**: Key findings, test results, and current treatment
4. **Specific Questions**: What the referring team would like the specialist to address

---

Source Clinical Notes:
`{source_text}`

Generate the referral letter now.

Task C: Patient-Facing Discharge Instructions

Based on the following discharge summary, generate clear, plain-language discharge instructions for the patient.
The instructions should include:
1. **What happened during your hospital stay**: Brief, accessible explanation
2. **Your diagnoses**: What conditions were identified or being monitored
3. **Medications**: Any new, changed, or stopped medications and why
4. **Follow-up**: When and where to follow up, any pending test results
5. **Warning signs**: When to seek immediate medical attention

Use clear, simple language appropriate for a general adult audience. Avoid medical jargon where possible.

Source Discharge Summary:
`{source_text}`

Generate the patient discharge instructions now.

Figure A3: Clinical-domain prompts (used in the depth, breadth, and mitigation experiments).

Clinical Domain

Clinical generations use the baseline system prompt below followed by one of three task prompts: hospitalization summary, referral letter, or patient-facing discharge instructions (Figure A3).

Legal Domain

Two task prompts are used: case summary and legal memorandum (Figure A4).

System Prompt (Legal)

"You are a legal documentation assistant. Generate accurate, professional legal text based on the provided source documents. Ensure all legal details are faithfully represented."

Task A: Case Summary

Based on the following court opinion/order, generate a concise case summary covering:
1. **Case Background**: Parties, claims, and procedural posture
2. **Key Legal Issues**: The central legal questions addressed
3. **Court's Analysis**: How the court reasoned through the issues
4. **Holding**: What the court decided
5. **Implications**: Any noted limitations, conditions, or open questions

---

Source Document:
`{source_text}`

Generate the case summary now.

Task B: Legal Memorandum

Based on the following case documents, draft a legal memorandum for a supervising attorney covering:
1. **Issue Statement**: The legal question(s) to be addressed
2. **Brief Answer**: Short preliminary conclusion
3. **Facts**: Relevant facts from the record
4. **Analysis**: Application of law to facts, including strengths and weaknesses
5. **Conclusion**: Recommended course of action

Source Documents:
`{source_text}`

Generate the legal memorandum now.

Figure A4: Legal-domain prompts.

Scientific Domain

Two task prompts are used: a structured paper summary and a plain-language summary aimed at a non-expert audience (Figure A5).

System Prompt (Scientific)

"You are a scientific writing assistant. Generate accurate, professional scientific text based on the provided source material. Ensure all technical details are faithfully represented."

Task A: Paper Summary

Based on the following research paper content, generate a structured summary covering:
1. **Motivation**: What problem does this work address and why is it important?
2. **Approach**: What method or framework is proposed?
3. **Key Results**: What are the main findings or contributions?
4. **Limitations**: What are the noted limitations or caveats?
5. **Impact**: What are the implications for the field?

Source Paper:
`{source_text}`

Generate the paper summary now.

Task B: Plain-Language Summary

Based on the following research paper content, generate a plain-language summary suitable for a non-expert audience (e.g., science journalists, policymakers, or the general public).

The summary should:
- Explain the research question in accessible terms
- Describe the key findings without jargon
- Note any important caveats or limitations
- Be 150-250 words

Source Paper:
`{source_text}`

Generate the plain-language summary now.

Figure A5: Scientific-domain prompts.

Mitigation Prompts

The three mitigation strategies evaluated in §6 (Table A5, Figure A6) modify the clinical baseline as described in §5; the four additional strategies use the prompts in Figure A7.

Mitigation 1: Instruction (System Prompt Override)

"You are a clinical documentation assistant. Generate accurate, professional clinical text based on the provided source notes. Ensure all clinical details are faithfully represented. ***IMPORTANT:*** *Maintain the exact level of certainty expressed in the source notes. If a finding is described as possible, likely, suspected, or uncertain, you MUST preserve that qualification in your output. Do not upgrade uncertain findings to definitive claims, and do not add uncertainty where none exists in the source."*

Mitigation 2: Few-Shot (System Prompt with Examples)

"... Ensure all clinical details are faithfully represented.

CRITICAL: Preserve the exact level of certainty from the source. Study these examples:

CORRECT: Source says 'possible pneumonia' → Output says 'possible pneumonia'
WRONG: Source says 'possible pneumonia' → Output says 'pneumonia'
CORRECT: Source says 'cannot rule out malignancy' → Output says 'malignancy cannot be ruled out'
WRONG: Source says 'cannot rule out malignancy' → Output says 'malignancy'
CORRECT: Source says 'likely demand ischemia' → Output says 'likely demand ischemia'
WRONG: Source says 'likely demand ischemia' → Output says 'demand ischemia'
CORRECT: Source says 'mild aortic stenosis' → Output says 'mild aortic stenosis'
WRONG: Source says 'mild aortic stenosis' → Output says 'aortic stenosis'

Every hedge word (possible, likely, suspected, may, cannot rule out, suggestive of, mild, moderate, severe) must appear in your output if it appears in the source."

Mitigation 3: Post-Hoc Verification (Appended to Task Prompt)

IMPORTANT: After generating your output, review it against the source for uncertainty preservation.

Step 1: Generate the requested clinical text.
Step 2: List every uncertainty/hedge marker in the source (e.g., "possible", "likely", "cannot rule out", "suggestive of", "mild", "suspected").
Step 3: For each marker, verify it is preserved in your output. If any marker was dropped or changed, revise your output to restore it.
Step 4: Output ONLY the final revised clinical text (no commentary).

Figure A6: Mitigation prompts for the three strategies evaluated against the clinical baseline (Figure A3); the strategies are described in §5.

Mitigation 4: Strict Verbatim-Copy (System Prompt Override)

"You are a clinical documentation assistant with ONE overriding rule: epistemic fidelity. Generate the requested clinical text, but you MUST reproduce every uncertainty marker from the source VERBATIM. Hedge words and phrases — possible, possibly, likely, probable, presumed, suspected, may, might, could, cannot rule out, suggestive of, concerning for, consistent with, appears to, questionable, mild, moderate, borderline, equivocal, indeterminate — carry clinical meaning and MUST NOT be dropped, paraphrased into certainty, or strengthened. If the source says 'possible pneumonia', your output says 'possible pneumonia' — never 'pneumonia'. Copy the qualifier exactly. Do not add certainty the source did not assert."

Mitigation 5: Reason-First Self-Review (System Prompt + Task Suffix)

System prompt: "You are a clinical documentation assistant. Before writing, you reason explicitly about the source's uncertainty structure, then write faithfully."

Appended to task prompt: Work in two stages before producing the final text:

STAGE 1 (planning): Read the source and list EVERY uncertainty/hedge marker it contains (e.g., "possible", "likely", "suspected", "cannot rule out", "suggestive of", "mild"). Note which claim each qualifies.

STAGE 2 (writing): Write the requested clinical text so that every marker from your Stage-1 list appears, attached to the same claim, at the same strength. Output ONLY the final clinical text (do not print the Stage-1 list).

Mitigation 6: Few-Shot + Post-Hoc Combination

The Mitigation 2 few-shot system prompt and the Mitigation 3 post-hoc verification suffix (Figure A6), applied together.

Mitigation 7: Hard-Constraint Framing (System Prompt Override)

"You are a clinical documentation assistant operating under a hard constraint: your output is REJECTED if it removes or strengthens any uncertainty expressed in the source. Treat each source hedge as a non-negotiable token that must survive into your output unchanged. It is better to keep an awkward hedge than to lose it. Faithfully represent all clinical content, but never trade uncertainty for fluency or brevity. Preserve possible/likely/suspected/may/cannot-rule-out/suggestive-of/mild/moderate exactly as written in the source."

Figure A7: Additional mitigation prompts (§5).