Possible or Definite? A Benchmark for Evaluating Diagnostic Uncertainty Preservation in Clinical Text

Hongbo Du[*] HDU24@MY.TRINE.EDU
Trine University

Zixin Lu[*] LZIXIN@UMICH.EDU
University of Michigan

Jiaming Qu[†] QJIAMING@AMAZON.COM
Amazon

Abstract

Large language models (LLMs) are increasingly used for clinical text tasks such as summarization and revision. While most studies evaluate the fluency and coherence of LLM-generated text, whether LLMs correctly *preserve diagnostic uncertainty* remains underexplored. In clinical practice, phrases such as "possible pneumonia" communicate the strength of available evidence and directly guide decisions about follow-up testing and treatment. Altering these uncertainty expressions can change the clinical meaning entirely. In this paper, we systematically evaluated this problem in two steps. First, we constructed a benchmark of 1,200 clinical documents with 9,184 uncertainty annotations across five levels. Second, we evaluated three LLMs on this benchmark. Our results show that (1) LLMs preserve the original uncertainty cues poorly, often less than half the time; (2) LLMs struggle with nuanced distinctions between adjacent levels. This work reveals a failure mode not captured by standard evaluation metrics and provides implications for the safe deployment of LLMs in clinical workflows.

1. Introduction

Large language models (LLMs) have been widely used in biomedical natural language processing (NLP) tasks such as clinical summarization, report generation, and question answering (Tian et al., 2024). Studies have shown that LLMs fine-tuned on specialized corpora can achieve human-level performance on complex clinical tasks (Van Veen et al., 2024). To systematically evaluate LLM performance on these tasks, prior work has introduced a variety of benchmarks. For example, ProbSum targets problem-list summarization (Gao et al., 2023a), and (Xu et al., 2024) proposed a shared task for medical discharge note generation. However, existing evaluation metrics remain limited in scope, and many assessments do not directly measure real clinical impact (Bednarczyk et al., 2025).

Despite this progress, whether LLMs correctly interpret and preserve diagnostic uncertainty in clinical text remains underexplored. In clinical practice, phrases such as "*possible* pneumonia" and "*cannot exclude* pulmonary embolism" are not simply cautious wording.

[*] Equal contribution.
[†] Corresponding author. This research was conducted independently in a personal capacity and does not reflect the author's position at Amazon.

They communicate the strength of available evidence, indicate alternative diagnoses under consideration, and signal whether further testing is needed. These expressions directly influence clinical decision-making: a "possible" diagnosis may require additional imaging, while a definite assertion may lead to immediate treatment. Prior work has shown that uncertainty is a central and multidimensional component of clinical reasoning, and that clinicians use a wide range of uncertainty expressions in daily practice (Han et al., 2011; McGowan et al., 2025; Panicek and Hricak, 2016). However, LLMs can generate fluent and coherent text while altering the degree of uncertainty—for example, rewriting "*possible* pneumonia" as "pneumonia." If such changes propagate in clinical settings, they risk distorting the original clinical meaning. This issue has received little attention in prior work.

In this work, we study the following question: **Can LLMs correctly preserve diagnostic uncertainty in clinical NLP tasks?** Prior research on uncertainty in LLMs mainly follows two directions. The first studies uncertainty quantification, which uses logits, entropy, or related signals to detect unreliable outputs or hallucinations (Farquhar et al., 2024). The second studies uncertainty expressed *within* LLM-generated text, evaluating whether models express uncertainty appropriately (Kolagar and Zarcone, 2024; Yang et al., 2025a). Our work follows the second direction. We do not measure the model's internal confidence during generation. Instead, we evaluate whether LLMs preserve the uncertainty that is already present in clinical source text on specific tasks.

To address this question, we constructed a benchmark for diagnostic uncertainty in clinical text. Our benchmark is based on two widely used datasets: MIMIC-IV-Note (Johnson et al., 2023), which contains discharge summaries and radiology reports, and TCGA-Reports (Kefeli and Tatonetti, 2024), a collection of pathology reports. We defined five levels of uncertainty, ranging from certain absent to non-asserted. To support fine-grained analysis, we used proposition-level annotations rather than assigning a single uncertainty label to each document. Each document is represented as a set of cue-target pairs, where each pair includes a target concept (e.g., "pneumonia") and its associated uncertainty cue (e.g., "possible"). In total, our benchmark contains **1,200 documents across six clinical text types and 9,184 annotated cue-target pairs across five uncertainty levels**. We evaluated three LLMs on this benchmark using two complementary assessments: an indirect assessment that measures whether LLMs preserve uncertainty during text transformation tasks, and a direct assessment that tests whether an LLM can classify and rank uncertainty cues when explicitly prompted.

In summary, this paper systematically evaluates how LLMs handle diagnostic uncertainty in clinical NLP tasks. Our contributions are as follows. First, we constructed a benchmark covering multiple document types and uncertainty levels that can be reused in future work. Second, we provide an empirical evaluation of three LLMs across tasks and prompt conditions. Our results show that **LLMs frequently distort uncertainty: without appropriate prompting, they preserve the original uncertainty level less than half the time, and roughly two in five cue-target pairs are rewritten as definite assertions**. Additionally, LLMs classify individual uncertainty levels with moderate accuracy but struggle with fine-grained distinctions between adjacent levels. Together, these findings identify a systematic failure mode in LLM-based clinical text processing and highlight the urgent need for responsible LLMs usage in clinical workflows.

We release our codebase and the benchmark at https://github.com/HongboD/Clinical-Uncertainty.

Generalizable Insights about Machine Learning in the Context of Healthcare

- Clinical meaning depends not only on *which* condition is mentioned, but also on *how certainly* it is expressed. Our results show that LLMs can produce fluent, factually grounded outputs that still distort meaning by altering uncertainty, suggesting that uncertainty preservation should be evaluated as a separate dimension.
- Evaluation results on our proposition-level benchmark reveals that the dominant failure mode is not random noise but a systematic bias toward certainty assertion—a pattern invisible to document-level evaluation.
- Even explicit instructions to preserve uncertainty in the prompt are insufficient to fully prevent distortion, suggesting that more sophisticated interventions may be needed for safe and responsible LLM usage in clinical text processing tasks.

2. Related Work

2.1. Uncertainty in Natural Language

From a linguistic perspective, uncertainty is expressed through a wide range of forms rather than a single cue word. These include hedge words, modal verbs, differential diagnoses, incomplete evidence statements, and recommendations for follow-up testing (Kilicoglu and Bergler, 2008). These expressions are particularly important in biomedical settings, where they communicate limits of evidence, possible alternatives, and planned next steps in patient care. Prior work has shown that uncertainty in healthcare is multidimensional rather than incidental, spanning a wide range of linguistic forms and clinical functions (Han et al., 2011). Clinical NLP studies have developed resources for detecting related phenomena such as negation, hedging, and assertion status (Uzuner et al., 2011; Vincze et al., 2008; Peng et al., 2018). However, these datasets are limited in the number of uncertainty levels, the types of clinical text, and the overall scale. We extend this line of work by constructing a benchmark with broader coverage of uncertainty levels, document types, and dataset size.

2.2. Evaluating LLMs' Awareness of Uncertainty

Prior work on studying uncertainty in LLMs can be grouped into two main directions. The first direction studies LLMs' internal uncertainty, including methods based on logits, entropy, calibration, and related signals that estimate the likelihood of errors or hallucinations (Farquhar et al., 2024; Kadavath et al., 2022; Kapoor et al., 2024). This line of work evaluates how reliably LLMs generate responses based on their internal state. The second direction studies uncertainty expressed in LLM-generated language, examining whether uncertainty is expressed appropriately in the output. Examples include uncertainty transfer in summarization and long-form text generation (Kolagar and Zarcone, 2024; Yang et al., 2025a,b), as well as uncertainty-aware diagnosis and explanation (Zhou et al., 2025). Our work follows the second direction. Specifically, we focus on patient-specific clinical notes and evaluate whether diagnostic uncertainty is preserved from source text to output.

2.3. Benchmarks in Clinical LLMs

Prior studies have curated various benchmarks to evaluate LLMs across different tasks, including clinical note summarization, report generation, and question answering (Gao et al., 2023b; Van Veen et al., 2024; Xu et al., 2024; Kweon et al., 2024; Liu et al., 2024). These benchmarks mainly focus on evaluating the fluency and coherence of LLM-generated text and whether LLMs can produce accurate responses. However, a plausible and coherent output does not guarantee clinically meaningful assessment (Bednarczyk et al., 2025; Bedi et al., 2025; Gong et al., 2025). To address this gap, we constructed a benchmark to evaluate LLMs' understanding of diagnostic uncertainty. Motivated by prior work, our benchmark construction combines two approaches: (1) deriving supervision from existing clinical document collections (Gao et al., 2023a) and (2) combining automated data extraction with human review (Kweon et al., 2024; Grazhdanski et al., 2025).

3. Method

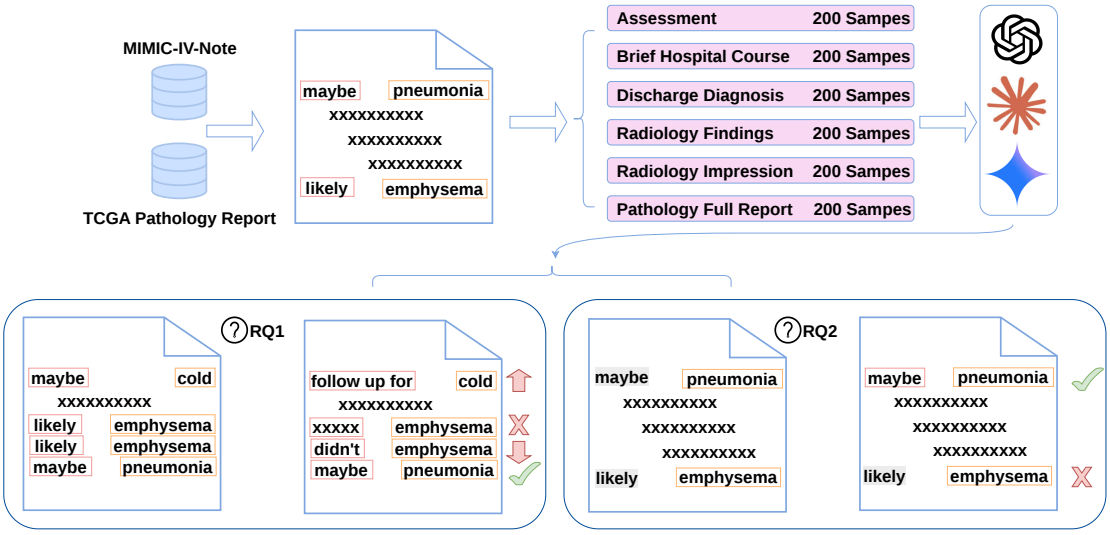

Figure 1: Overview of the study workflow.

The main goal of this study is to evaluate **whether LLMs can correctly interpret uncertainty cues in clinical text**. We do not evaluate LLMs on traditional measures such as generation quality, coherence, or factual accuracy. Instead, we study two complementary research questions (RQs):

- **RQ1 (Indirect assessment):** Do LLMs preserve diagnostic uncertainty when transforming clinical text in downstream tasks such as summarization and patient-friendly rewriting?
- **RQ2 (Direct assessment):** Can LLMs correctly identify and rank uncertainty cues when explicitly asked to interpret them?

To study these RQs, we took two steps (Figure 1). First, because no suitable dataset was readily available, we constructed a benchmark of 1,200 clinical documents spanning six clinical text types. We defined five uncertainty levels, ranging from certain absent to non-asserted. Second, we evaluated three LLMs on this benchmark. We describe our benchmark construction (Section 3.1) and evaluation approaches (Section 3.2) below.

3.1. Benchmark Construction

Data Pre-processing. We constructed the benchmark from two datasets that are widely used in biomedical NLP research: (1) MIMIC-IV-Note (Johnson et al., 2023), which contains 2.6M clinical notes, and (2) TCGA-Reports (Kefeli and Tatonetti, 2024), a dataset of 9,523 pathology reports. For MIMIC-IV-Note, we focused on two document types: discharge summaries and radiology reports, as they are relatively long and provide rich material for studying uncertainty. Each MIMIC note is organized into sections, and we selected the most clinically relevant ones: assessment, brief hospital course, and discharge diagnosis for discharge summaries; impressions and findings for radiology reports. Each section was extracted as a standalone document. For TCGA pathology reports, we used the full document because structured sections were not available. In total, our initial document pool included six document types, all of which are interpretive clinical text in which clinicians record conclusions, exclusions, contingencies, and unresolved issues about the patient.

Defining Uncertainty Cues. A key step in our benchmark construction was extracting uncertainty cue-target pairs from each document. We define an *uncertainty cue* as a text span (a word or phrase) that expresses the writer's degree of commitment toward a clinical proposition. The *target* is the clinical proposition modified by that cue. For example, in the sentence *"Limited evaluation due to suboptimal contrast timing; no large central filling defect is seen, but pulmonary embolism cannot be excluded,"* "cannot be excluded" is the uncertainty cue and "pulmonary embolism" is the target.

Informed by prior work on clinical uncertainty, negation, and assertion status in biomedical text (Han et al., 2011; Bhise et al., 2018; McGowan et al., 2025; Vincze et al., 2008; Uzuner et al., 2011; Peng et al., 2018; Lima Lopez et al., 2020; Thompson et al., 2011; Callen et al., 2020), we defined a five-level uncertainty schema. The five levels range from certain absent to non-asserted mentions:

- **Certain absent:** The clinical proposition is not present for the patient, expressed with clear and confident language that leaves no doubt about its absence.
- **Probable present:** The clinical proposition is more likely present than absent, but the language indicates that the writer is not fully certain.
- **Possible present:** The clinical proposition may be present, but the language indicates that this is only one reasonable possibility rather than a confident conclusion.
- **Indeterminate or unresolved:** The clinical proposition cannot yet be determined because the available information is incomplete, mixed, or unclear.
- **Non-asserted evaluation target:** The clinical proposition is mentioned as something to check, test for, monitor, or rule out. The text does not state that it is present or absent; it is only the target of evaluation.

Table 1: We defined five uncertainty levels (ULs), with UL5 representing the most uncertainty and UL1 representing the least. We developed a set of cue words for each level. N denotes the unique number of cues for each level.

Level	**Type**	**Uncertainty cues**
UL1	Certain absent	($N = 9$): no evidence of; without evidence of; negative for; absence of; free of
UL2	Probable present	($N = 21$): likely; probable; favored; most consistent with; favored to represent
UL3	Possible present	($N = 45$): possible; may represent; might represent; could represent; concern for; concerning for; suspicious for; question of
UL4	Indeterminate	($N = 19$): cannot determine; cannot be determined; indeterminate; unclear; unknown
UL5	Non-asserted	($N = 16$): ?; rule out; follow-up to exclude; workup for; query; evaluate for

We began with an uncertainty cue inventory drawn from the studies above, identifying approximately five common cues for each level. To increase coverage, we took two approaches. First, each author independently reviewed a small set of MIMIC notes and conducted qualitative coding of sentences expressing uncertainty. All authors then met twice to discuss and select the most frequent cues. Second, we invited a linguist and a clinician to a discussion panel in which all authors participated for final refinement. This step helped capture conventions specific to clinical notes, such as the use of the question mark as a common symbol for non-asserted information. Table 1 summarizes definitions and representative cues for each level. UL3 (*Possible present*) has the most unique cues ($N = 45$), while UL1 (*Certain absent*) has the fewest ($N = 9$). A full list of cues for each level is provided in Appendix A.

It is notable that this schema captures uncertainty at the proposition level rather than a fully symmetric ontology of present and absent probability: certain present propositions are treated as definite assertions during output-side scoring, while UL1 captures explicitly asserted absence. UL2 and UL3 distinguish stronger versus weaker possible presence, UL4 captures indeterminacy due to insufficient or mixed evidence, and UL5 captures non-asserted evaluation targets such as rule-out, follow-up, workup, or monitoring mentions. Because some surface forms can appear in different clinical roles, final labels are assigned using the cue, target, syntactic direction, and clinical function together. For example, "possible pneumonia" is labeled UL3 when it asserts possible current presence, whereas "follow up to exclude pneumonia" or "? pneumonia" is labeled UL5 when pneumonia is only the target of evaluation.

Cue-Target Pair Extraction. To extract cue-target pairs from each document, we built a rule-based pipeline that operates in two stages: extraction and refinement. We deliberately chose a rule-based approach rather than using LLMs to avoid circular evaluation, where the same type of model would both generate the benchmark labels and be evaluated

against them. We also prioritized precision over recall, as the benchmark requires reliable labels rather than exhaustive coverage. Our pipeline proceeds as follows:

- **Step 1: Extraction.** We first segment each document into sentences and scan for uncertainty cues using regular expressions (regex) over the cue inventory defined in Table 1. Each cue pattern encodes a syntactic direction: some cues modify the phrase that follows (e.g., "*no evidence of* pneumonia"), while others modify the phrase that precedes them (e.g., "the adrenal nodule *is indeterminate*"). When a cue is detected, we extract a candidate target span from the appropriate direction within the same sentence. If the candidate span contains coordination structures (e.g., "PE or pneumonia" after "negative for"), we split it into separate targets and create one cue-target pair per target.

- **Step 2: Refinement.** The extraction stage produces candidate pairs whose target spans are raw text fragments. These fragments may include extraneous tokens (e.g., trailing punctuation, leading articles, overly broad clauses) or may miss the precise medical concept boundary. The refinement stage addresses this by independently extracting medical concepts from the same text and linking them back to the pairs from Step 1. This stage consists of three consecutive sub-steps.

- **Step 2a: Medical Concept Extraction.** We run medSpaCy (Eyre et al., 2022) for medical named entity recognition on each sentence. We then use QuickUMLS (Soldaini and Goharian, 2016) to match each entity to a Concept Unique Identifier (CUI) in the Unified Medical Language System (Bodenreider, 2004). This step produces a set of medical concept candidates for each sentence.

- **Step 2b: Candidate Scoring and Linking.** For each cue-target pair from Step 1, we score every medical concept candidate from Step 2a based on span overlap with the original regex capture, character proximity to the cue, and directional consistency with the cue's syntactic pattern. If all candidates score below a minimum threshold, the pair is flagged for manual review.

- **Step 2c: Target Replacement.** When a high-confidence medical concept is found, we replace the raw regex-captured target span with the matched medical concept. If no suitable match exists for a given target—which is common for clinical abbreviations (e.g., "FNH," "PNA," "AMS")—we retain the original target from Step 1.

After applying the pipeline to each sentence within a document, we perform deduplication on the target text to obtain a set of unique targets. For target texts with exact string matches, we prioritize the pair that fills a gap in the document's uncertainty level coverage. For target texts that differ lexically but share the same CUI, we prioritize the more specific surface form. Finally, each document has a set of cue-target pairs with unique targets.

Sampling. After extracting cue-target pairs from each document, we conducted a pre-screening step by excluding documents that covered fewer than two unique uncertainty levels. These documents provided insufficient uncertainty-related content for the benchmark. We then sampled from the pre-screened documents, giving higher priority to documents with broader coverage of uncertainty levels. For example, documents containing pairs from

all five levels were sampled first. We sampled 200 documents from each of the six document types, resulting in a final benchmark of 1,200 documents. We adopt this sampling strategy to create a balanced stress-test set: sampling an equal number of documents from each document type prevents longer or more frequent note types from dominating the benchmark. We prioritized documents with broader within-document uncertainty coverage because RQ1 transforms whole documents and RQ2 evaluates co-occurring uncertainty signals.

We performed human validation for quality control. We randomly sampled 15 documents from each document type, for a total of 90 documents (7.5% of the benchmark). Two authors[1] independently reviewed all extracted target texts and their associated cues, covering 891 pairs in total. For each pair, they annotated whether *both* the target text and the cue were extracted correctly. The results showed an average accuracy of 82%, with an inter-reviewer agreement of Cohen's $\kappa = 0.61$, indicating moderate agreement (Landis and Koch, 1977). These results suggest that the rule-based pipeline achieved reasonable quality for benchmark construction. Here, the validation task assessed whether uncertainty cue-target pairs were correctly extracted; it did not adjudicate clinical diagnosis or treatment correctness. The 82% accuracy is therefore a strict pair-level measure: a pair was counted as correct only when both the cue and the target were correct. Under this criterion, 82% accuracy and Cohen's $\kappa = 0.61$ indicate reasonable quality for a nuanced proposition-level extraction task. Comparable clinical NLP annotation and assertion-status evaluations have reported similar ranges (Chaturvedi et al., 2023; Savova et al., 2010).

Benchmark Statistics. The final benchmark contains 1,200 documents across six clinical document types, with 200 documents per type: assessment, brief hospital course, discharge diagnosis, radiology findings, radiology impression, and full pathology reports. Document length ranges from 17 to 4,013 words (mean = 572.41; S.D. = 482.4). Because we prioritized documents with broader uncertainty level coverage during sampling, each document covers 3.54 unique uncertainty levels on average. There is slight variance across document types due to the nature of the data: brief hospital course documents tend to cover all five levels because they are longer and describe key events, treatments, and outcomes during a patient's stay; in contrast, assessment, discharge diagnosis, and TCGA reports are dominated by documents covering only three levels, as they tend to be conclusion-oriented rather than focused on clinical findings. In total, the benchmark contains 9,184 cue-target pairs,[2] with an average of 7.65 pairs per document.

3.2. Evaluations

Our RQ1 evaluates whether LLMs preserve uncertainty when performing downstream text transformation tasks. Our RQ2 evaluates whether LLMs can directly identify and rank uncertainty cues when explicitly prompted. We describe the experimental setup and metrics for each RQ below.

1. Both were biomedical/clinical NLP researchers with 3 and 5 years of relevant experience, respectively.
2. The detailed distribution across uncertainty levels is: 3,233 pairs in UL1, 2,147 in UL2, 2,631 in UL3, 467 in UL4, and 706 in UL5.

3.2.1. RQ1: Indirect Assessment

Experimental design. We evaluate LLMs on two free-text clinical transformation tasks. The first task is *clinician handoff summarization*, in which the model produces a concise clinician-facing summary for continuity of care. This task places strong pressure on compression and reorganization, making it a natural setting for observing uncertainty preservation or erosion. The second task is *patient-friendly revision*, in which the model rewrites the source text in plain language while preserving its meaning.

Each task is evaluated under two prompt conditions. In the Baseline condition, the model receives only the core task instruction and the source text. In the Guarded condition, the instruction additionally requires the model to rely only on source information, avoid adding new diagnoses or interpretations, and preserve uncertain or hedged statements rather than converting them into definite claims. This yields a 2 (task) $\times$ 2 (prompt condition) design. All prompts are provided in Appendix B.

Models and infrastructure. We evaluated three LLMs: `gpt-oss-120b` (OpenAI, 2025), `gemini-2.5-flash` (Google, 2025), and `claude-haiku-4.5` (Anthropic, 2025). To comply with data retention policies for the MIMIC-IV-Note dataset, we ran the three models on separate platforms: Groq[3] for `gpt-oss-120b`, Google Vertex AI[4] for `gemini-2.5-flash`, and AWS Bedrock[5] for `claude-haiku-4.5`. The decoding temperature was set to 0 for all models, as we focus on faithful transformation rather than response diversity.

Metrics. To evaluate RQ1, we need to determine whether each source cue-target pair is preserved in the LLM-generated text. Let $\mathcal{S}$ denote the set of all cue-target pairs in the source document. For each pair, we search the LLM output for the target using exact string match, normalized lexical forms, phrases linked to the same CUI, concept-key matches, or broader/narrower concept matches. After a target is matched in the generated output, we assign its output uncertainty level using linked uncertainty events, local cue patterns around the matched target, and medSpaCy assertion status. If a source target is retained but no output-side uncertainty cue or assertion evidence is found, we count the output as a definite assertion, contributing to CAR. This design allows paraphrased targets to be credited while still measuring whether the uncertainty attached to the proposition was preserved. This produces two subsets: $\mathcal{S}_+$, the pairs whose targets are retained in the output, and $\mathcal{S}_-$, the pairs whose targets are absent. We define five metrics:

- *Target Retained Rate (TRR):* The proportion of source pairs whose targets appear in the model output (i.e., $|\mathcal{S}_+|/|\mathcal{S}|$). TRR measures whether the model keeps the clinical proposition at all, regardless of its uncertainty level.
- *Uncertainty Retained Rate (URR):* Among $\mathcal{S}_+$, the proportion of pairs for which the model preserves the same uncertainty level as the source. URR is the primary metric, as it directly measures whether the model faithfully reproduces the original degree of uncertainty.
- *Certainty Assertion Rate (CAR):* Among $\mathcal{S}_+$, the proportion of pairs for which the model removes the uncertainty cue entirely, rewriting the target as a definite assertion.

3. https://console.groq.com/docs
4. https://cloud.google.com/vertex-ai
5. https://aws.amazon.com/bedrock/

CAR captures the most clinically concerning form of distortion: collapsing hedged language into certain claims.

- *Partial Certainty Rate (PCR):* Among $\mathcal{S}_+$, the proportion of pairs for which the model shifts the uncertainty level toward certainty without reaching full assertion. PCR complements CAR by capturing partial movement toward certainty.
- *Over-hedging Rate (OHR):* Among $\mathcal{S}_+$, the proportion of pairs for which the model shifts the uncertainty level away from certainty. OHR captures cases where the model introduces more hedging than the original text.

3.2.2. RQ2: Direct Assessment

Experimental design. We evaluate whether LLMs can directly interpret clinical uncertainty when explicitly prompted. We sampled 500 cue-target pairs from the benchmark, with 100 pairs from each uncertainty level. Each sample contains a short sentence snippet, a target concept, and a ground truth uncertainty level. The model is prompted to (1) identify the shortest quote in the snippet that signals the uncertainty status of the target, and (2) classify the corresponding uncertainty level. We provide the definitions of UL1–UL5 in the prompt. We use `gemini-2.5-flash` for this experiment because it was the strongest-performing model in RQ1.

We evaluate two settings. In the single-signal setting (N=1), the model receives one cue-target pair at a time and classifies its uncertainty level. In the multi-signal setting (N=2, 3, 4, 5), the model receives a bundle of N cue-target pairs simultaneously and must classify each pair and rank them from strongest to weakest certainty. The multi-signal setting tests whether the model can distinguish between uncertainty levels when several signals must be compared jointly. For the multi-signal setting, we construct 100 bundles for each value of N, yielding 400 bundles in total across N=2–5. The full prompts for both settings are provided in Appendix C.

Metrics. We define four metrics for RQ2:

- *Uncertainty Level Accuracy (ULA):* The proportion of pairs for which the predicted uncertainty level exactly matches the gold label. ULA is the primary classification metric and is reported for all settings.
- *Macro-F1:* The unweighted mean of per-level F1 scores across UL1–UL5. Each level contributes equally regardless of classification difficulty. Macro-F1 complements ULA by accounting for potential imbalances in per-level performance.
- *Per-level F1:* The F1 score computed separately for each uncertainty level. Per-level F1 reveals which levels the model handles well and which it struggles with.
- *Exact Rank Accuracy:* The proportion of bundles for which the model predicts the entire strongest-to-weakest ordering correctly. This is only for the N=2–5 setting.

4. Results

4.1. RQ1: Indirect Assessment

Table 2 reports all five metrics for the three LLMs across both tasks and prompt conditions. We identify five trends.

Table 2: RQ1 results. All values are percentages. Higher TRR and URR indicate better preservation; higher CAR, PCR, and OHR indicate more distortion.

Model	Task / Condition	TRR	URR	CAR	PCR	OHR
`claude-haiku-4.5`	Revision / Baseline	32.82	33.64	44.96	3.88	1.26
	Revision / Guarded	36.57	43.35	40.07	3.57	1.37
	Summarization / Baseline	55.25	43.32	41.90	5.30	1.73
	Summarization / Guarded	67.79	62.51	24.59	4.22	1.78
`gemini-2.5-flash`	Revision / Baseline	46.54	38.98	42.65	4.16	1.54
	Revision / Guarded	48.41	46.06	37.04	3.89	1.50
	Summarization / Baseline	49.99	45.55	38.12	4.98	2.24
	Summarization / Guarded	62.45	62.53	23.64	3.54	1.95
`gpt-oss-120b`	Revision / Baseline	48.08	37.77	42.80	3.79	2.01
	Revision / Guarded	50.59	47.72	34.42	4.10	1.63
	Summarization / Baseline	53.44	44.36	37.88	6.12	1.83
	Summarization / Guarded	59.21	56.18	29.28	5.36	1.69

Models frequently omit clinical targets. Under the Baseline condition, TRR ranges from 32.82% to 55.25%, indicating that LLMs often drop clinical targets entirely. Revision yields substantially lower TRR than summarization across all models. For example, `claude-haiku-4.5` retains only 32.82% of targets during revision compared to 55.25% during summarization. This suggests that patient-friendly rewriting leads to more aggressive content removal than clinician handoff summarization.

Uncertainty preservation is poor under baseline. Under the Baseline condition, URR ranges from 33.64% to 45.55% across all model-task combinations, indicating that even when a model retains a clinical target, it preserves the original uncertainty level correctly less than half the time. The pattern is consistent across all three models. Summarization yields slightly higher URR than revision (e.g., 45.55% vs. 38.98% for `gemini-2.5-flash`), suggesting that the more extensive rewriting required by revision makes uncertainty more vulnerable to distortion. The document-clustered bootstrap 95% CIs (Table 8, Appendix D) support this trend: baseline URR ranges from 33.6 [31.2, 36.3] to 45.5 [43.5, 47.5].

Distortion is overwhelmingly toward certainty. Among the three distortion metrics (CAR, PCR, OHR), certainty assertion is by far the dominant failure mode. Under the Baseline condition, CAR ranges from 37.88% to 44.96%, meaning that roughly two in five retained targets are rewritten as definite assertions with the uncertainty cue removed entirely. In contrast, PCR ranges from 3.79% to 6.12% and OHR ranges from 1.26% to 2.24%. Together, partial certainty shifts and over-hedging account for less than 8% of retained pairs. This indicates that when LLMs distort uncertainty, they almost always

collapse hedged language into definite assertions rather than make subtle shifts along the uncertainty scale.

Guarded prompting helps but does not eliminate distortion. Comparing the Guarded condition to the Baseline, URR increases for all six model-task combinations, with gains ranging from +7.08 to +19.19 percentage points. The improvements are consistently larger for summarization than for revision. CAR also decreases under the Guarded condition by 4.89–17.31 percentage points. However, even the best guarded configuration achieves only 62.53% URR, leaving substantial room for improvement. All guarded-prompt URR gains have document-clustered bootstrap 95% CIs (Table 8, Appendix D) above zero, ranging from +7.1 percentage points [5.0, 9.2] to +19.2 percentage points [17.2, 21.1], indicating that the improvements are robust.

Variation across tasks. Across all models and prompt conditions, summarization consistently outperforms revision on both TRR and URR. This pattern likely reflects the nature of the two tasks: summarization compresses content for a clinical audience that expects hedged language, while revision rewrites for patients and tends to simplify or remove clinical nuance. The gap is most pronounced for `claude-haiku-4.5`, where baseline summarization URR exceeds baseline revision URR by nearly 10 percentage points.

4.2. RQ2: Direct Assessment

Table 3 reports the overall classification and ranking metrics across the single-signal and multi-signal settings. Table 4 reports per-level F1 scores. We only experimented with `gemini-2.5-flash` based on the overall best performance on RQ1. We identify three trends.

Classification is moderately accurate; ranking degrades with signal load. In the single-signal setting (N=1), it achieves 78.4% ULA and 78.6% Macro-F1 (Table 3). When the model receives multiple signals simultaneously (N=2–5), classification accuracy remains stable, indicating that additional signals do not impair per-item labeling. However, exact rank accuracy decreases steadily from 86.0% at N=2 to 48.0% at N=5, as the number of possible orderings grows factorially. Kendall's τ remains above 0.72 across all values of N, confirming that predicted rankings stay positively correlated with the gold rankings even when the exact order is incorrect.

Table 3: RQ2 results by number of simultaneously presented signals (N).

N	ULA	Macro-F1	Exact Rank Accuracy
1	78.4	78.6	—
2	80.5	79.8	86.0
3	80.3	80.0	78.0
4	82.0	81.8	68.0
5	81.6	81.3	48.0
2–5	81.3	81.0	70.0

Per-level performance and the effect of contrastive signals. In the single-signal setting, the model performs best on UL1 (Certain absent), which involves clear negation cues, and worst on UL3 (Possible present), a level that requires fine-grained distinction from

adjacent levels (Table 4). This pattern suggests that the model handles categorical distinctions (present vs. absent) more reliably than graded ones along the uncertainty spectrum. In the multi-signal setting, the F1 scores for UL2 and UL3 increases by 10.5 and 9.2 points respectively compared to N=1, suggesting that contrastive signals from other levels help the model calibrate the boundary between probable and possible. In contrast, F1 for UL4 and UL5 decreases by 3.2 and 4.8 points, indicating that indeterminate and non-asserted levels become harder to distinguish when multiple signals are presented simultaneously.

Table 4: RQ2 per-level F1 scores under the single-signal (N=1) and multi-signal (N=2–5, averaged) settings.

Level	N=1	N=2–5
UL1 (Certain absent)	94.6	94.8
UL2 (Probable present)	72.2	82.7
UL3 (Possible present)	67.0	76.2
UL4 (Indeterminate)	83.1	79.9
UL5 (Non-asserted)	76.1	71.3

The aggregate confusion matrix (Table 9, Appendix D) further shows that errors concentrate between neighboring or conceptually close levels, especially UL2/UL3 and UL4/UL5, while UL1 remains the easiest level to identify.

5. Discussion

5.1. Summary of Findings

Our study evaluated whether LLMs correctly preserve diagnostic uncertainty in clinical text through two complementary assessments on a benchmark of 1,200 documents with 9,184 proposition-level uncertainty annotations across five levels.

In the indirect assessment (RQ1), we found that LLMs frequently distort uncertainty during both summarization and revision. Under baseline prompting, all three LLMs preserve the original uncertainty level less than half the time. The dominant failure mode is certainty assertion: roughly two in five retained propositions are rewritten as definite claims with the uncertainty cue removed entirely. LLMs also frequently omit clinical targets altogether, particularly during revision, where the more extensive rewriting may cause models to drop uncertainty-bearing content. The Guarded condition slightly improves uncertainty preservation across all models and tasks.

In the direct assessment (RQ2), `gemini-2.5-flash` classifies uncertainty levels with moderate accuracy when presented with a single signal. Classification accuracy remains stable as the number of simultaneously presented signals increases, but exact ranking accuracy degrades from 86% at N=2 to 48% at N=5. Per-level analysis reveals that the model handles categorical distinctions reliably but struggles with fine-grained distinctions along the uncertainty spectrum, particularly between probable and possible.

5.2. Follow-Up Analysis on Additional Models

To increase the coverage of LLMs, we additionally evaluated `Sonnet 4.6` (Anthropic, 2026) and `GPT-5.5` (OpenAI, 2026) as frontier models and `MedGemma-27B` (Sellergren et al., 2025) as a clinical-domain model. We report these as follow-up analyses rather than as part of the main benchmark comparison.

Table 5: Follow-up RQ1 results for additional models. All values are percentages.

Model	Task (Baseline)	TRR	URR	CAR	PCR	OHR
`Sonnet 4.6`	Revision	42.05	37.18	38.94	4.84	1.94
	Summarization	57.72	46.27	38.67	5.47	1.43
`GPT-5.5`	Revision	47.20	48.24	32.11	5.01	2.31
	Summarization	54.18	53.16	30.10	6.23	2.41
`MedGemma-27B`	Revision	40.82	37.74	44.78	3.98	2.57
	Summarization	43.97	41.41	39.75	5.13	3.76

Table 6: Follow-up RQ2 results for additional models. ULA and Macro-F1 are shown for N=1/2/3/4/5; Exact Rank Accuracy is shown for N=2/3/4/5.

Model	ULA	Macro-F1	Exact Rank Accuracy
Sonnet 4.6	74.0/73.5/76.7/76.8/77.2	73.8/72.9/76.8/77.1/77.5	84/66/50/34
GPT-5.5	79.8/77.5/78.3/76.5/77.6	79.7/77.5/78.5/76.6/77.7	84/77/52/34
MedGemma-27B	72.6/73.5/76.3/74.5/72.4	72.3/73.3/76.3/74.3/72.2	74/70/48/27

The follow-up results support the same overall conclusion. `GPT-5.5` performs best among the added models, but still preserves uncertainty for only 48.24% of retained targets in revision and 53.16% in summarization. `MedGemma-27B`, despite being a clinical-domain model, shows similar vulnerability, with URR of 37.74% in revision and 41.41% in summarization and CAR near 40–45%. In RQ2, all three added models achieve moderate classification performance but continue to struggle with exact ordering as the number of simultaneously presented signals increases. These results suggest that uncertainty erosion is not limited to the original three general-purpose models.

5.3. Implications for Using LLMs in Clinical Tasks

Our study offers three implications for using LLMs in clinical tasks.

First, our findings suggest that uncertainty preservation should be treated as a distinct evaluation dimension for clinical LLM systems. Current evaluation of clinical text generation focuses on content coverage, factual consistency, and fluency, yet recent reviews have noted that safety-relevant evaluation remains underdeveloped (Bednarczyk et al., 2025; Bedi et al., 2025), mirroring how other hidden biases in clinical ML pipelines can corrupt downstream inference (Liang et al., 2025). BERTScore for RQ1 outputs (Table 7, Appendix D) was around 82% across all conditions. However, despite appearing fluent and factually plausible, the LLM-generated outputs still changed the clinical meaning by altering the degree

of certainty attached to a diagnosis. This can lead to harmful consequences in healthcare, where uncertainty communicates the strength of evidence, the openness of the differential diagnosis, and the need for further testing (Han et al., 2011; Bhise et al., 2018). Our study highlights the importance of developing benchmarks and metrics that evaluate uncertainty preservation to support safe and responsible use of LLMs in clinical tasks.

Second, the certainty distortion is systematic rather than random. Across models and tasks, the dominant error mode is movement toward certainty by removing uncertainty cues entirely, e.g., *"~~possible~~ pneumonia"*. This is especially concerning because the most dangerous distortion is not adding false information, but converting a tentative proposition into a definite one. One plausible explanation is that LLMs inherit a bias from training data in which hedges are often removed for simplicity (Kolagar and Zarcone, 2024). Our RQ2 results reinforce this: the model classifies uncertainty accurately when asked, yet fails to preserve same cues during downstream tasks. This gap suggests that the bottleneck is not whether the model can recognize uncertainty, but whether it prioritizes preserving uncertainty while optimizing for compression and fluency (Gong et al., 2025).

Finally, our results show that guarded prompting helps but is not sufficient, and that vulnerability varies by task. The Guarded condition improved URR by 7–19 percentage points, but even the best configuration reached only about 63% URR. Moreover, patient-friendly revision is consistently more vulnerable than clinician-facing summarization across all models and prompt conditions. This matters because patient-facing text is where uncertainty distortion has the most direct impact: patients and caregivers typically cannot independently assess the underlying evidence and may take a definite-sounding statement at face value. Together, these findings suggest that guarded prompting should be interpreted as a lightweight, architecture-free baseline mitigation rather than a complete solution. Future work can explore stronger interventions such as fine-tuning on our benchmark, preference optimization with uncertainty-aware reward signals, or post-hoc verification (Yang et al., 2025a,b; Zhou et al., 2025).

5.4. Limitations and Future Work

Our study has several limitations. First, the benchmark was constructed using a rule-based pipeline rather than full manual annotation. Although human validation showed reasonable quality (82% accuracy, Cohen's $\kappa = 0.61$), some extraction errors may remain. Second, we evaluated three general-purpose LLMs as off-the-shelf tools. While the Guarded condition suggests improvement over the baseline, we did not fine-tune on clinical data or incorporate uncertainty-aware reward signals, which may further improve performance. Finally, perceptions of uncertainty are inherently subjective and may vary across clinicians. We plan to investigate whether distorted uncertainty leads to different clinical decisions through clinician-facing user studies that measure real downstream impact. We also plan to explore a multilingual uncertainty cue inventory in future work.

6. Conclusion

In this paper, we presented a systematic evaluation of whether LLMs preserve diagnostic uncertainty in clinical text. We constructed a benchmark of 1,200 clinical documents with 9,184 proposition-level uncertainty annotations across five levels and evaluated three LLMs

using two complementary assessments. Our results show that LLMs frequently distort uncertainty during text transformation, with the dominant failure mode being the collapse of hedged language into definite assertions. Guarded prompting reduces but does not eliminate this distortion. When directly prompted to classify and rank uncertainty levels, the model classifies individual levels with moderate accuracy but struggles with fine-grained distinctions between adjacent levels. These findings identify a clinically important failure mode not captured by standard evaluation metrics. We will release our benchmark to support future work and encourage joint effort between computer scientists and clinicians for safe and responsible usage of LLMs in clinical NLP tasks.

References

Anthropic. Introducing claude haiku 4.5. Anthropic announcement, oct 2025. URL https://www.anthropic.com/news/claude-haiku-4-5.

Anthropic. Introducing claude sonnet 4.6. Anthropic product announcement, February 2026. URL https://www.anthropic.com/news/claude-sonnet-4-6.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*, 333(4):319–328, 2025.

Lydie Bednarczyk, Daniel Reichenpfader, Christophe Gaudet-Blavignac, Amon Kenna Ette, Jamil Zaghir, Yuanyuan Zheng, Adel Bensahla, Mina Bjelogrlic, and Christian Lovis. Scientific evidence for clinical text summarization using large language models: Scoping review. *Journal of Medical Internet Research*, 27:e68998, 2025.

Viraj Bhise, Suja S. Rajan, Dean F. Sittig, Robert O. Morgan, Parashar Chaudhary, and Hardeep Singh. Defining and measuring diagnostic uncertainty in medicine: A systematic review. *Journal of General Internal Medicine*, 33(1):103–115, 2018.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

Andrew L. Callen, Sara M. Dupont, Adi Price, Ben Laguna, David McCoy, Bao Do, Jason Talbott, Marc Kohli, and Jared Narvid. Between always and never: Evaluating uncertainty in radiology reports using natural language processing. *Journal of Digital Imaging*, 33(5):1194–1201, 2020.

Jaya Chaturvedi, Sumithra Velupillai, Robert Stewart, and Angus Roberts. Identifying mentions of pain in mental health records text: A natural language processing approach, 2023. URL https://arxiv.org/abs/2304.01240.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. Launching into clinical

space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings*, volume 2021, pages 438–447, 2022.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, and Majid Afshar. Overview of the problem list summarization (probsum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 461–467. Association for Computational Linguistics, 2023a.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. DR.BENCH: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of Biomedical Informatics*, 138:104286, 2023b.

Eun Jeong Gong, Chang Seok Bang, Jae Jun Lee, and Gwang Ho Baik. Knowledge-practice performance gap in clinical large language models: Systematic review of 39 benchmarks. *Journal of Medical Internet Research*, 27:e84120, 2025.

Google. Gemini 2.5 flash. Google AI for Developers documentation, 2025. URL https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash.

Georgi Grazhdanski, Vasil Vasilev, Sylvia Vassileva, Dimitar Taskov, Izabel Antova, Ivan Koychev, and Svetla Boytcheva. Synthmedic: Utilizing large language models for synthetic discharge summary generation, correction and validation. *Journal of Biomedical Informatics*, 170:104906, 2025.

Paul K. J. Han, William M. P. Klein, and Neeraj K. Arora. Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, 31(6):828–838, 2011.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14. Association for Computational Linguistics, 2024.

Jenna Kefeli and Nicholas Tatonetti. Tcga-reports: A machine-readable pathology report resource for benchmarking text-based ai models. *Patterns*, 5(3), 2024.

Halil Kilicoglu and Sabine Bergler. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(11):S10, 2008.

Zahra Kolagar and Alessandra Zarcone. 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. Association for Computational Linguistics, 2024.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. EHRNoteQA: An LLM benchmark for real-world clinical practice using discharge summaries. *arXiv preprint arXiv:2402.16040*, 2024.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

Zhongyuan Liang, Arvind Suresh, and Irene Y. Chen. Treatment non-adherence bias in clinical machine learning: A real-world study on hypertension medication. In *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pages 430–442. PMLR, 2025.

Salvador Lima Lopez, Naiara Perez, Montse Cuadros, and German Rigau. 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, 2020.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. Large language models in the clinic: A comprehensive benchmark. *arXiv preprint arXiv:2405.00716*, 2024.

Samuel K. McGowan, Maria-Jose Corrales-Martinez, Teva Brender, Alexander K. Smith, Shannen Kim, Krista L. Harrison, Hunter Mills, Albert Lee, David Bamman, Julien Cobert, et al. Unclear trajectory and uncertain benefit: Creating a lexicon for clinical uncertainty in patients with critical or advanced illness using a delphi consensus process. *Medical Decision Making*, 45(1):34–44, 2025.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. OpenAI model card, aug 2025. URL https://openai.com/index/gpt-oss-model-card/.

OpenAI. Introducing gpt-5.5. OpenAI product announcement, April 2026. URL https://openai.com/index/introducing-gpt-5-5/.

David M. Panicek and Hedvig Hricak. How sure are you, doctor? a standardized lexicon to describe the radiologist's level of certainty. *AJR. American Journal of Roentgenology*, 207(1):2–3, 2016.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Joint Summits on Translational Science Proceedings*, 2017:188–196, 2018.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010. doi: 10.1136/jamia.2009.001560.

Andrew Sellergren et al. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4, 2016.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12:393, 2011.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493, 2024.

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

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142, 2024.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9, 2008.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis P. Langlotz, and Jean-Benoit Delbrouck. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98. Association for Computational Linguistics, 2024.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Sen Yang, Nigel Collier, Dong Yu, and Deqing Yang. Logu: Long-form generation with uncertainty expressions. In

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18947–18968. Association for Computational Linguistics, 2025a.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Dong Yu, Nigel Collier, and Deqing Yang. UNCLE: Benchmarking uncertainty expressions in long-form generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30340–30356. Association for Computational Linguistics, 2025b.

Shuang Zhou, Jiashuo Wang, Zidu Xu, Song Wang, David Brauer, Lindsay Welton, Jacob Cogan, Yuen-Hei Chung, Lei Tian, Zaifu Zhan, Yu Hou, Mingquan Lin, Genevieve B. Melton, and Rui Zhang. Uncertainty-aware large language models for explainable disease diagnosis. *npj Digital Medicine*, 8(1):690, 2025.

Appendix A. Full List of Uncertainty Cues

All uncertainty cues used in this study are summarized below:

- UL1 ($N = 9$): no evidence of; without evidence of; negative for; absence of; free of; no; denies; denied; absent
- UL2 ($N = 21$): most consistent with; favored to represent; favoured to represent; most likely represent; most likely represents; likely secondary to; likely due to; likely from; likely represent; likely represents; most likely; likely; probable; probably; suggestive of; compatible with; consistent with; favors; favours; favored; favoured
- UL3 ($N = 45$): may represent; might represent; could represent; may reflect; might reflect; could reflect; may be; might be; could be; possible; possibly; concern for; concerning for; suspicious for; question of; appears to be; seems to be; differential includes; differential diagnosis includes; suggests the possibility of; questionable; suspect; suspected; difficult to exclude; hard to exclude; cannot be excluded; cannot be completely excluded; cannot be fully excluded; cannot be entirely excluded; cannot exclude; cannot completely exclude; cannot fully exclude; cannot entirely exclude; can't be excluded; can't be completely excluded; can't be fully excluded; can't exclude; can't completely exclude; can't fully exclude; cannot rule out; can't rule out; is not excluded; are not excluded; was not excluded; were not excluded
- UL4 ($N = 19$): cannot determine; cannot be determined; indeterminate; unclear; unknown; equivocal; nondiagnostic; not clear; too small to characterize; too small to fully characterize; too small to adequately characterize; limited evaluation for; limited exam for; suboptimal for; not clearly; not entirely clear; not well visualized; motion degraded; degraded by motion
- UL5 ($N = 16$): ?; follow-up to exclude; follow up to exclude; to exclude; evaluate for; workup for; rule out; r/o; query; monitor for; watch for; repeat imaging; repeat ct; could indicate; may indicate; might indicate

  Contextual note: The phrases "could indicate," "may indicate," and "might indicate" are assigned to UL5 only in conditional-contingency constructions concerning a future observation or event. Unconditional uses describing a target's possible current presence are assigned to UL3.

Appendix B. Prompts in RQ1 Experiments

Baseline Prompt for the Summarization task

You are a clinical handoff assistant helping prepare a concise clinician-to-clinician summary for the next provider taking over care.
Your task is to summarize the source text into a brief clinician-facing handoff summary.
The output will be read by another clinician during a care transition. Focus on the most clinically important information needed for continuity of care, including major problems, key findings, current status, and important near-term considerations. Keep the summary concise, accurate, and professionally written.
{{input text}}

Baseline Prompt for the Revision task

You are a patient education assistant helping rewrite clinical text into plain-language information that a patient or family member can understand.
Your task is to rewrite the source text in clear, patient-friendly language without changing its meaning.
The output will be read by a patient or caregiver. Use plain, direct, non-technical language where possible. Preserve the clinical meaning, but make the wording easier to understand. Keep the tone respectful, calm, and supportive.
{{input text}}

Guarded Prompt for the Summarization task

You are a clinical handoff assistant helping prepare a concise clinician-to-clinician summary for the next provider taking over care.
Your task is to summarize the source text into a brief clinician-facing handoff summary.
The output will be read by another clinician during a care transition. Focus on the most clinically important information needed for continuity of care, including major problems, key findings, current status, and important near-term considerations. Keep the summary concise, accurate, and professionally written.
Use only information that appears in the source text. Do not add new facts, interpretations, diagnoses, actions, recommendations, timelines, or causes. Do not convert uncertain or hedged statements into definite statements. Preserve the original meaning and level of detail as closely as possible. Improve clarity, grammar, fluency, and organization. Keep the tone professional and neutral. Output only the rewritten text.
{{input text}}

Guarded Prompt for the Revision task

You are a patient education assistant helping rewrite clinical text into plain-language information that a patient or family member can understand.
Your task is to rewrite the source text in clear, patient-friendly language without changing its meaning.
The output will be read by a patient or caregiver. Use plain, direct, non-technical language where possible. Preserve the clinical meaning, but make the wording easier to understand. Keep the tone respectful, calm, and supportive.
Use only information that appears in the source text. Do not add new facts, interpretations, diagnoses, actions, recommendations, timelines, or causes. Do not convert uncertain or hedged statements into definite statements. Preserve the original meaning and level of detail as closely as possible. Improve clarity, grammar, fluency, and organization. Keep the tone professional and neutral. Output only the rewritten text.
{{input text}}

Appendix C. Prompts in RQ2 Experiments

Direct Uncertainty Classification Prompt

You will read a short clinical text snippet. Focus only on the target concept given below. Ignore uncertainty language that applies to other concepts in the snippet.
Uncertainty level definitions:
UL1: the concept is explicitly absent or ruled out; this is a hard negative.
UL2: the author favors the concept, but does not assert it definitively.
UL3: the concept is plausible, but the author shows weaker commitment than UL2.
UL4: the text explicitly says the concept is unresolved, indeterminate, or cannot currently be decided.
UL5: the concept is mentioned as a query, rule-out target, workup target, follow-up target, or patient-specific future contingency, rather than as a current conclusion about the patient's present state.
Task: 1. Find the shortest quote in the snippet that signals the uncertainty status of the target concept. 2. Assign the correct uncertainty level for that target.
Output rules: Return JSON only. Do not wrap the JSON in markdown code fences. Do not include any explanation outside the JSON. Copy the evidence quote exactly from the snippet. Judge only the target concept, not the whole sentence. If several phrases are present, choose the one that most directly determines the uncertainty level for the target. Quote the uncertainty signal, not the target phrase itself. Use only the text in the snippet. Do not use outside clinical knowledge. Use only one of these labels: UL1, UL2, UL3, UL4, UL5. Use UL5 narrowly: if the text expresses a current possibility or probability about the target, choose UL2 or UL3 instead of UL5. If the text says the target cannot currently be resolved, choose UL4 instead of UL5. If the best evidence is punctuation such as "?", copy that exact symbol as the evidence quote.
Output schema: {{"target_id": "...", "target_text": "...", "evidence_quote": "...", "uncertainty_level": "UL1 or UL2 or UL3 or UL4 or UL5"}}
Target: target_id: {target_id}, target_text: {target_text}
{{input text}}

Multi-Signal Uncertainty Classification and Ranking Prompt

You will read multiple short clinical text signals. Each signal contains one target concept and one short snippet. Your job has two parts.
For each signal, find the shortest quote in the snippet that determines the uncertainty status of the target, and assign the correct uncertainty level.
After you have labeled all signals, rank them from strongest to weakest author commitment about the target as a current patient-state claim.
Uncertainty level definitions:
UL1: the concept is explicitly absent or ruled out; this is a hard negative.
UL2: the author favors the concept, but does not assert it definitively.
UL3: the concept is plausible, but the author shows weaker commitment than UL2.
UL4: the text explicitly says the concept is unresolved, indeterminate, or cannot currently be decided.
UL5: the concept is mentioned as a query, rule-out target, workup target, follow-up target, or patient-specific future contingency, rather than as a current conclusion about the patient's present state.
Ranking order: For ranking, use this fixed order of author commitment, strongest to weakest: UL1 > UL2 > UL3 > UL4 > UL5. Rank strength of author commitment about the patient's current state. Do not rank disease severity, clinical importance, or treatment priority. Judge each signal separately before ranking the full set. Use only the text in each snippet. Do not use outside clinical knowledge. Focus only on the anchored target for each signal. Ignore uncertainty language that applies to other concepts in the same snippet.
Output rules: Return JSON only. Do not wrap the JSON in markdown code fences. Do not include any explanation outside the JSON. Copy each evidence quote exactly from the corresponding snippet. Use the shortest quote that most directly determines the uncertainty level. Use only one of these labels: UL1, UL2, UL3, UL4, UL5. Use UL5 narrowly: if the text expresses a current possibility or probability about the target, choose UL2 or UL3 instead of UL5. If the text says the target cannot currently be resolved, choose UL4 instead of UL5. If the best evidence is punctuation such as "?", copy that exact symbol as the evidence quote. The signals list in your output must contain every input signal exactly once. The rank_strongest_to_weakest list must contain every signal_id exactly once, with no duplicates and no omissions.
Output schema: {{"signals": [{{"signal_id": "S1", "target_text": "...", "evidence_quote": "...", "uncertainty_level": "UL1 or UL2 or UL3 or UL4 or UL5"}}], "rank_strongest_to_weakest": ["S1", "S2", "S3"]}}
{{input text}}

Appendix D. Additional Analysis Results

Table 7: BERTScore-F1 between source documents and LLM-generated outputs in RQ1. All values are percentages. Despite high semantic similarity across all conditions, LLMs do not preserve uncertainty cues correctly (see Section 4.1).

Model	Task / Condition	BERTScore F1
`claude-haiku-4.5`	Revision / Baseline	82.00
	Revision / Guarded	83.67
	Summarization / Baseline	83.45
	Summarization / Guarded	85.93
`gemini-2.5-flash`	Revision / Baseline	82.20
	Revision / Guarded	83.94
	Summarization / Baseline	84.65
	Summarization / Guarded	86.46
`gpt-oss-120b`	Revision / Baseline	82.06
	Revision / Guarded	84.04
	Summarization / Baseline	82.34
	Summarization / Guarded	84.64

Table 8: Document-clustered bootstrap 95% confidence intervals for URR and CAR. URR and CAR are reported as percentages, and guarded-prompt gains are reported as percentage-point differences relative to the corresponding baseline condition. (see Section 4.1).

Model	Task	Baseline URR, 95% CI	Baseline CAR, 95% CI	Guarded ΔURR, 95% CI	Guarded CAR, 95% CI
Haiku 4.5	Revision	33.6 [31.2, 36.3]	45.0 [41.7, 48.0]	+9.6 [7.0, 12.3]	40.1 [36.4, 42.7]
	Summarization	43.3 [41.4, 45.2]	41.9 [39.6, 44.3]	+19.2 [17.2, 21.1]	24.6 [22.9, 26.4]
gemini-2.5-flash	Revision	39.0 [36.6, 41.2]	42.7 [40.6, 44.8]	+7.1 [5.0, 9.2]	37.0 [35.0, 39.6]
	Summarization	45.5 [43.5, 47.5]	38.1 [35.8, 40.4]	+16.9 [15.1, 18.6]	23.6 [22.0, 25.3]
gpt-oss-120b	Revision	37.8 [35.9, 40.2]	42.8 [40.6, 44.7]	+9.9 [7.8, 11.9]	34.4 [32.4, 36.5]
	Summarization	44.4 [42.4, 46.0]	37.9 [35.8, 39.9]	+11.8 [10.1, 13.7]	29.3 [27.6, 30.9]
Sonnet 4.6	Revision	37.2 [35.1, 39.6]	38.9 [36.8, 41.6]	—	—
	Summarization	46.3 [44.5, 48.2]	38.7 [36.5, 40.6]	—	—
gpt-5.5	Revision	48.2 [46.1, 50.8]	32.1 [30.1, 34.2]	—	—
	Summarization	53.2 [51.0, 55.2]	30.1 [27.6, 32.3]	—	—
medgemma-27b	Revision	37.7 [34.1, 41.5]	44.8 [40.8, 48.3]	—	—
	Summarization	41.4 [39.3, 43.7]	39.7 [37.2, 42.3]	—	—

Table 9: Aggregate RQ2 confusion matrix across all four evaluated RQ2 models and N=1–5. Rows are gold labels and columns are predicted labels.

Gold / Pred.	**UL1**	**UL2**	**UL3**	**UL4**	**UL5**
UL1	1463	39	9	9	0
UL2	4	1299	171	3	43
UL3	52	292	1001	25	150
UL4	11	288	115	1064	42
UL5	53	130	133	179	1025

Table 10: Example of proposition-level scoring in RQ1. The target is retained, but the output removes the source uncertainty cue, so the pair contributes to TRR and CAR but not URR.

Field	**Example**
Source cue-target pair	Source text: "CXR raised concern for R mediobasal parenchymal opacity, *suggestive of* **pneumonia**." Cue: "suggestive of"; target: "pneumonia"; source level: UL2.
Output match	Output text: "Your family member came to the hospital very sick with an infection in the lungs (**pneumonia**) that made it hard to breathe." Match class: exact surface match.
Assigned output level	No local uncertainty cue is attached to the retained target in the output, so the output level is assigned as a definite assertion.
Metric contribution	Target retained: yes; uncertainty retained: no; certainty assertion: yes. The pair contributes to TRR and CAR, but not URR.