MediEval: A Unified Medical Benchmark for Patient-Contextual and Knowledge-Grounded Reasoning in LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly applied to medicine, yet their adoption is limited by concerns over reliability and safety. Existing evaluations either test factual medical knowledge in isolation or assess patient-level reasoning without verifying correctness, leaving a critical gap. We introduce MediEval, a benchmark that links MIMIC-IV electronic health records (EHRs) to a unified knowledge base built from UMLS and other biomedical vocabularies. MediEval generates diverse factual and counterfactual medical statements within real patient contexts, enabling systematic evaluation across a 4-quadrant framework that jointly considers knowledge grounding and contextual consistency. Using this framework, we identify critical failure modes, including hallucinated support and truth inversion, that current proprietary, open-source, and domain-specific LLMs frequently exhibit. To address these risks, we propose Counterfactual Risk-Aware Fine-tuning (CoRFu), a DPO-based method with an asymmetric penalty targeting unsafe confusions. CoRFu improves by +16.4 macro-F1 points over the base model and eliminates truth inversion errors, demonstrating both higher accuracy and substantially greater safety.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse domains, with medicine being among the most high-impact areas of application. In clinical contexts, LLMs have been explored for tasks such as summarizing electronic health records (EHRs), generating discharge instructions, providing clinical decision support, and answering medical questions (Rajpurkar et al., 2022; Singhal et al., 2025; Moor et al., 2023; Singhal et al., 2023; Pal et al., 2022). Their appeal lies in the ability to integrate unstructured text with medical knowledge, potentially reducing documentation burden and assisting clinicians in decision-making.

Translating research prototypes into real-world deployment hinges critically on reliability and safety. Unlike generic NLP tasks, medical reasoning requires not only factual correctness but also contextual grounding in patient-specific data while adhering to verified medical knowledge. Errors in this setting are not mere degradations in performance but risks that can directly translate into patient harm (Thirunavukarasu et al., 2023; Yang et al., 2023; Haltaufderheide and Ranisch, 2024).

A critical challenge is that LLMs often fail to apply medical knowledge consistently within the heterogeneous and noisy context of patient records (Zhou et al., 2025b). For example, a model may state that metformin, a therapy for type 2 diabetes, is contraindicated in severe renal impairment, yet fail to apply this knowledge when the condition appears in a noisy and heterogeneous patient record. The cause of such errors may be that current models are trained to recall facts in isolation rather than to integrate them with diverse patient information. Such inconsistencies expose a critical gap between *knowing* medical facts and *using* them safely.

Existing evaluation paradigms only partially address this gap. Medical data are uniquely challenging because patient records are heterogeneous with free-text clinical notes and tabular entries coded in different systems for diagnoses, procedures, and medications. Medical knowledge is hierarchical and ontology-driven (e.g., UMLS, SNOMED CT, RxNorm), but its large scale, noise, and limited cross-vocabulary connectivity make consistent reasoning difficult. Benchmarks based on EHRs assess the ability to extract or reason over structured data, but often reduce the task to retrieval or serialization without verifying medical soundness (Lovón-Melgarejo et al., 2025). In contrast, knowledge-based evaluations test whether LLMs can handle logical transformations of medical facts (Zhou et al., 2025a; Sung et al., 2021), but do not connect reasoning to real patient contexts. The field

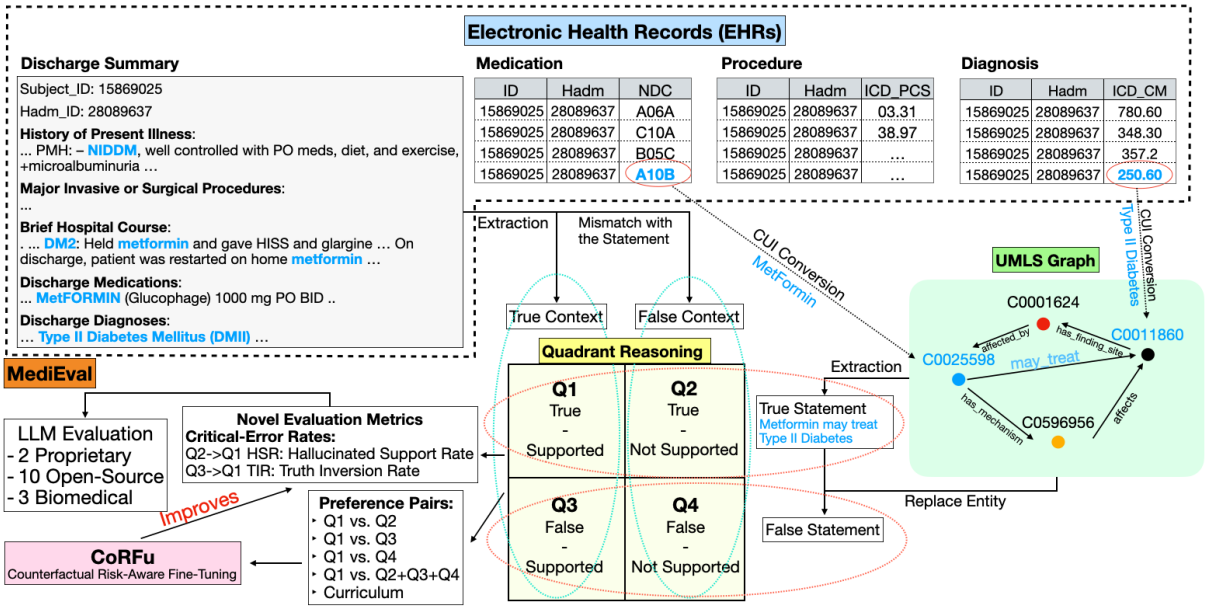

Figure 1: Overview of the current work with a real example; texts in blue indicate the extracted sample.

thus lacks a unified framework that probes whether LLMs can (i) remain faithful to medical knowledge and (ii) apply it consistently to individual patient records.

In this paper, we address this gap with **MediEval** (Figure 1), a benchmark and evaluation framework that links real patient records (MIMIC-IV) (Johnson et al., 2023) with a unified biomedical knowledge base built from UMLS (Bodenreider, 2004), SNOMED CT (Donnelly et al., 2006), and RxNorm (Liu et al., 2005; Nelson et al., 2011). To ensure rigorous construction, MediEval constructs evaluation statements by applying graph-guided substitutions and recombinations within biomedical ontologies, with plausibility checks to ensure that true cases remain clinically valid while false ones are realistic and challenging. These statements are embedded in real patient contexts, enabling evaluation of quadrant-level reasoning, defined as *True/False* with respect to medical knowledge and *Supported/Unsupported* with respect to the patient record. To capture safety-critical errors, we introduce two new metrics: Hallucinated Support Rate (HSR) and Truth Inversion Rate (TIR). We then conduct a comprehensive evaluation of proprietary, open-source (general-purposes), and biomedical LLMs under this unified protocol, revealing consistent weaknesses across quadrants.

To mitigate the identified risks, we introduce **Counterfactual Risk-Aware Fine-tuning (CoRFu)**, a DPO-based method that uses quadrant-structured preference pairs and an asymmetric penalty to push models to prefer "True+Supported" responses over safety-critical confusions (e.g., "False+Supported"). This risk-aware, counterfactual approach targets precisely the failure modes that conventional preference optimization overlooks. Our code are publicly available[1]. Our key contributions are:

- **MediEval Benchmark.** A unified benchmark linking EHRs with biomedical ontologies to test both factual grounding and patient consistency, with novel safety evaluation metrics.
- **Comprehensive Evaluation.** Cross-model evaluation of proprietary, open-source, and biomedical LLMs under identical settings, revealing systematic quadrant-level weaknesses and safety-critical error patterns.
- **Counterfactual Risk-Aware Fine-Tuning (CoRFu).** A risk-aware extension of DPO with asymmetric penalties and quadrant-aware approach that reduces safety-critical errors in medical reasoning tasks.

2 Related Work

A growing body of work investigates whether LLMs can handle complex information in EHRs. One study focused on data serialization and retrieval, with performance sensitive to prompt design and feature selection (Lovón-Melgarejo et al., 2025). Another work has explored long-context

[1] https://anonymous.4open.science/r/MediEval-5FA5/

modeling, showing some architectures can process entire patient timelines with >10k events and achieve greater robustness to irregular temporal patterns (Wornow et al., 2025). Other approaches integrate structured and unstructured data, for instance, by using small auxiliary models as "knowledge triggers" to support tabular prediction (Yan et al., 2025) or by designing code-aware representations that capture semantic and temporal structure in diagnosis prediction (Tan et al., 2025). These studies demonstrate progress in surfacing and encoding patient information, but whether model outputs are consistent with established clinical knowledge remains underexplored.

Another line of research investigates the reliability of the medical knowledge encoded in LLMs. Dynamic probing methods show that models often achieve low joint accuracy when facts are rephrased (Zhou et al., 2025a), while multifaceted evaluations reveal brittle performance on tasks requiring comparison, verification, or rectification (Zhou et al., 2024). Broader studies also report steep declines when moving from factual recall to scenario-based reasoning as cognitive complexity increases (Zhou et al., 2025b). However, even when scenarios are considered, these evaluations do not disentangle whether errors arise from factual misunderstanding, failures of contextual grounding, or both.

Several approaches have sought to improve the factual reliability of LLMs by grounding them in structured biomedical knowledge. These include triplet generation with knowledge graph verification (Su et al., 2025), alignment with biomedical knowledge graph embeddings (Sakhovskiy and Tutubalina, 2025), and graph-based retrieval augmentation for medical question-answering (Wu et al., 2025). Scholars have also critiqued the evaluation paradigms: some argue that multiple-choice questions reward shallow pattern recognition rather than genuine medical knowledge (Griot et al., 2025), while others stress that medical benchmarks must prioritize construct validity to remain clinically meaningful (Alaa et al., 2025). Automatic evaluation frameworks, such as AutoMedEval (Zhang et al., 2025), further highlight the need for scalable yet reliable assessment pipelines.

3 Methodology

3.1 MediEval Framework: Data Construction

MediEval is a benchmark constructed from the MIMIC-IV database (Medical Information Mart for Intensive Care IV) (Johnson et al., 2023) in combination with biomedical ontologies such as UMLS (Bodenreider, 2004). It is designed to evaluate whether large language models (LLMs) can perform clinical inference by verifying whether medical statements are factually correct and properly grounded in patient records. MediEval integrates structured EHR tables, unstructured clinical notes, and ontology-derived knowledge into a unified dataset of context–statement pairs annotated with correctness and support labels (see Figure 1).

Electronic Health Records (EHRs) as Foundation. MIMIC-IV is a large-scale publicly available collection of de-identified electronic health records (EHR) comprising 546,028 hospital admissions for 223,452 unique individuals. Two modalities are utilized: (i) structured tabular data (lists of diagnoses, procedures, and medications recorded with ICD-CM, ICD-PCS, and NDC codes), and (ii) unstructured discharge summaries, which typically contain about 22 sections such as *Brief Hospital Course*, *History of Present Illness*, *Major Surgical or Invasive Procedures*, and *Discharge Diagnoses*.

For each admission a, we define the structured event set and the discharge summary section set as

$$S_a = \{\, \mathcal{Z}^{\text{diag}}_{i_a},\ \mathcal{Z}^{\text{proc}}_{j_a},\ \mathcal{Z}^{\text{med}}_{k_a} \,\}, \tag{1}$$

$$DS_a = \{\, s_a^{(1)},\ s_a^{(2)},\ \ldots,\ s_a^{(m_a)} \,\}. \tag{2}$$

Here $\mathcal{Z}^{\text{diag}}_{i_a}$, $\mathcal{Z}^{\text{proc}}_{j_a}$, and $\mathcal{Z}^{\text{med}}_{k_a}$ denote the lists of diagnosis, procedure, and medication codes for admission a, containing i_a, j_a, and k_a elements respectively, while DS_a represents the set of m_a segmented sections from the discharge summary. See Figure 1 for a real sample from MIMIC-IV.

Step 1: Context Extraction. Discharge summaries are divided into multiple sections, not all of which are equally informative. For MediEval, we define the evaluation context $\mathcal{C}_a \subset DS_a$ as the subset of discharge summary sections that best capture patient trajectory and clinical decision-making. In practice, we select sections covering diagnoses, procedures, treatments, medical history, and hospital course, as these jointly provide the most comprehensive evidence. Importantly, the reasoning that connects these entities is not explicitly structured but embedded in free-text narratives, requiring models to infer relations such as diagnosis–medication from natural language.

Step 2: Semantic Normalization of Medical Codes. Electronic health records employ heterogeneous coding systems: diagnoses and procedures in MIMIC-IV are stored using ICD-9/10, while medications are represented with National Drug Codes (NDC). To reason across modalities, these codes must be projected into a unified semantic space. We achieve this by leveraging the Unified Medical Language System (UMLS) (Bodenreider, 2004), which integrates over 60 families of biomedical vocabularies and provides explicit mappings between them. Specifically, ICD codes are aligned with SNOMED CT concepts, NDC codes are aligned with RxNorm drug identifiers, and all are normalized into Concept Unique Identifiers (CUIs) that serve as universal nodes for cross-system integration. Formally,

$$f : z \mapsto \text{CUI}(z),\ U_a = \{f(z) \mid z \in S_a\}. \quad (3)$$

Here, z denotes a raw code that is being mapped to a CUI(z), and U_a is the set of normalized CUIs for admission a. This normalization ensures that diagnoses, procedures, and medications from different coding systems can be compared, linked, and reasoned over in a unified space.

Step 3: Ontology Graph Construction and Relation Extraction. From the normalized concepts across the cohort, we construct a semantic graph $G = (V, E)$, where V are CUIs and E are relations drawn from the UMLS Metathesaurus, including *treated_by*, *has_associated_procedure*, and *is_a* etc. A dictionary mapping synonyms and abbreviations to CUIs supports robust entity matching in text. Many clinically valid associations are not encoded as direct links but instead mediated through intermediate nodes such as therapeutic classes, anatomical structures, or related conditions. To capture these, we allow multi-hop traversal in G, limited to paths of length at most three in order to preserve clinical plausibility.

Formally, for concepts $h, t \in V$, let $d(h, t)$ be their hop distance; we define their relations as

$$R(h, t) = \{\, r \mid (h, r, t) \in G,\ d(h, t) \leq 3 \,\} \quad (4)$$

Direct relations. The semantic graph captures clinically meaningful associations as single-edge links. For instance, it encodes that the antihypertensive drug *Lisinopril* (CUI C0023861) is linked to the condition *Hypertension* (CUI C0020538) through a *treated_by* relation. Similarly, it represents that the procedure *Coronary Angioplasty* (CUI C0001979) is associated with the diagnosis *Coronary Artery Disease* (CUI C0010054) via a *has_associated_procedure* relation.

Multi-hop inference. Not all clinically valid relations are encoded as direct edges between specific entities in UMLS. For example, many drug–disease links are only represented at the therapeutic class level rather than for individual drugs. As a result, the relation between *Lisinopril* and *Hypertension* may absent as a direct edge. Instead, the graph encodes that Lisinopril (CUI C0023861) belongs to the class *Angiotensin-Converting Enzyme Inhibitors* (CUI C0003028) via an *is_a* relation, and that this class as a whole is linked to *Hypertensive disease* (CUI C0020538) via a *treats* relation:

$$(\text{C0023861}) \xrightarrow{\text{is_a}} (\text{C0003028}) \xrightarrow{\text{treats}} (\text{C0020538}).$$

By traversing this two-hop path, the system can infer the treatment relation that is missing at the leaf level. Such multi-hop reasoning is essential for capturing clinically valid associations that are indirectly encoded in UMLS.

In summary, U_a provides the admission-specific set of normalized concepts, while G encodes both direct and inferred relations among these concepts. Together they form the foundation for structured fact extraction in MediEval.

Step 4: Quadrant-based Statement Construction. For each fact (h, r, t), we generate a set of samples covering the four MediEval quadrants:

$$\mathcal{Y}_a(h, r, t) = \{y^{Q1}, y^{Q2}, y^{Q3}, y^{Q4}\}. \quad (5)$$

Quadrants are defined by the cross of factual correctness (true vs. false) and contextual grounding (supported vs. unsupported), as summarized in Figure 1. This design ensures that evaluation covers both straightforward comprehension and more subtle failure modes such as plausible contradictions, relational mismatch, and erroneous information.

The quadrants are constructed as follows. *Q1 (True–Supported, Supported Fact)* represents the baseline case: a correct fact directly attested in the patient context $\mathcal{C}_a$. *Q2 (True–Unsupported, Plausible Contradiction)* is generated by replacing an entity with a closely related alternative from the ontology graph G (e.g., another entity sharing the same parent concept) that is absent from $\mathcal{C}_a$, yielding a statement that is medically correct in general but unsupported for this patient (see Figure 2 for an example). *Q3 (False–Supported, Relational Mismatch)* recombines entities from distinct true facts,

creating a false relation even though all entities appear in $\mathcal{C}_a$, thereby testing whether a model can resist misleading but superficially plausible associations. Finally, *Q4 (False–Unsupported, Erroneous Information)* introduces a semantically distant distractor entity from G, producing a statement that is both false and unsupported.

This procedure requires ontology-guided substitutions, recombinations, and plausibility checks to guarantee that true statements remain clinically valid while false ones are both plausible and challenging. The resulting dataset integrates curated narrative context, ontology-based code normalization, and graph reasoning into a unified framework for rigorous evaluation of factual reliability and context adherence. See Figure 2 and Appendix A for real samples of each quadrant.

Context:

History of Present Illness:
... patient with metastatic NSCLC to bone, admitted after Hct 23 with dizziness and fatigue. Reports poor sleep, anxiety about prognosis ...

Past Medical History:
... Hypertension, osteoporosis, GERD, tobacco dependence ...

Brief Hospital Course:
... Anemia: transfused 2 units PRBC → Hct improved. ... Pain: continued home oxycontin and oxycodone PRN. ... Anxiety: continued ativan 0.5 mg PRN. ... GERD: continued omeprazole (PPI). ... HTN: continued atenolol 25 mg ...

Discharge Medications:
... Omeprazole 20 mg daily
... Oxycodone ER 60 mg q12h
... Oxycodone 5 mg q6h PRN
... Lorazepam 0.5 mg q4h PRN
... Atenolol 25 mg daily
... Sucralfate 1 g QID ...

Discharge Diagnoses:
... Metastatic NSCLC, anemia, hypertension, osteoporosis, GERD ...

Statement:
Gastroesophageal reflux disease may be treated by aluminum hydroxide, which is a type of Antacids

Label:
Q2 True-Unsupported

Figure 2: Example of statement verification against patient records (Quadrant 2: True-Unsupported). The statement is medically correct, yet in this case, GERD is treated with omeprazole, a different class of medication, so the statement is not supported by the context.

3.2 MediEval Framework: LLM Evaluation

Quadrant Formulation as a Four-Way NLI Task. Given a patient context c and a candidate medical statement y, the evaluation task is to assign y to one of four quadrants $q \in \{Q1, Q2, Q3, Q4\}$. This formulation extends natural language inference (NLI) into a clinically grounded setting: rather than only judging whether a statement is correct, the model must also decide whether it is grounded in the patient record. By framing correctness and support jointly in a 2×2 structure, MediEval captures safety-critical confusions that cannot be identified when these dimensions are assessed separately. Although motivated by medicine, this setup provides a general template for domains where factual validity and contextual grounding are both essential. To ensure a fair comparison, all non-proprietary models are fine-tuned under an identical supervised setup with LoRA adapters and a classification head on the four quadrants, ensuring that differences in performance reflect model capabilities rather than variations in training.

Baseline Metrics. We evaluate models on a test set $\mathcal{D}_{\text{test}} = \{(c_i, y_i, q_i)\}$ with gold quadrant labels q_i and predictions $\hat{q}_i$. Overall accuracy is defined as

$$\text{Accuracy} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_i \mathbb{I}(\hat{q}_i = q_i). \tag{6}$$

In addition, we report macro-averaged Precision, Recall, and F1 across quadrants, together with per-quadrant F1:

$$F1(Q_j) = \frac{2\,\text{Prec}(Q_j)\,\text{Rec}(Q_j)}{\text{Prec}(Q_j) + \text{Rec}(Q_j)}, \quad j = 1, \ldots, 4. \tag{7}$$

These metrics capture overall discrimination ability and expose potential biases, such as over-predicting "supported" statements.

Critical Error Rates. Beyond global performance, clinical safety requires avoiding specific misclassifications. A model may state a correct medical fact but wrongly claim it is supported by the patient record, or elevate an incorrect statement to seemingly valid evidence. To capture these entangled risks, we introduce two error rates:

Hallucinated Support Rate (HSR). The fraction of true but unsupported statements (Q2) misclassified as true and supported (Q1):

$$\text{HSR} = \frac{\sum_i \mathbb{I}(q_i = Q2 \wedge \hat{q}_i = Q1)}{\sum_i \mathbb{I}(q_i = Q2)}. \tag{8}$$

A high HSR indicates that the model hallucinates evidence, presenting correct medical knowledge as if it were grounded in the record.

Truth Inversion Rate (TIR). The proportion of false but superficially supported statements (Q3) misclassified as true and supported (Q1):

$$\text{TIR} = \frac{\sum_i \mathbb{I}(q_i = Q3 \wedge \hat{q}_i = Q1)}{\sum_i \mathbb{I}(q_i = Q3)}. \tag{9}$$

A high TIR indicates that the model elevates incorrect or unsafe claims to seemingly valid evidence, a critical failure mode in clinical contexts.

Interpretation of Metrics. These metrics disentangle complementary aspects of model behavior: *Overall accuracy* and *macro-F1* capture balanced classification performance. *Per-quadrant F1* identifies systematic weaknesses (e.g., under-recognition of unsupported truths). *HSR* and *TIR* directly target the most safety-critical confusions between quadrants: hallucinating support and inverting truth.

3.3 Counterfactual Risk-Aware Fine-Tuning

Large language models are typically fine-tuned in two broad ways. The most direct approach is *supervised fine-tuning* (SFT), where the model is trained to predict gold-standard labels with a cross-entropy loss. While simple and effective, SFT treats each label independently. In contrast, *preference-based optimization* exploits comparisons between a preferred and a dispreferred output. A prominent example is *Direct Preference Optimization* (DPO), which updates the model to increase the relative likelihood of preferred responses while staying close to a reference model (Rafailov et al., 2023).

Building on these foundations and the supervised fine-tuning results established in our evaluation task, we propose **CoRFu (Counterfactual Risk-aware Fine-tuning)**, a contrastive fine-tuning strategy tailored to the clinical safety challenges revealed by MediEval. Training proceeds on preference triplets (c, y_w, y_l), where c is the clinical context, y_w is a factually supported statement, and y_l is a statement drawn from the counterfactual quadrants of MediEval. The key innovation is the **CoRFu loss function**, which augments the standard DPO objective with an asymmetric penalty that more severely punishes safety-critical errors.

Let S denote the DPO preference margin:

$$S(c; y_w, y_l) = \beta \log[\frac{\pi_\theta(y_w|c)\, \pi_{\text{ref}}(y_l|c)}{\pi_{\text{ref}}(y_w|c)\, \pi_\theta(y_l|c)}]. \quad (10)$$

and define the CoRFu loss as:

$$\mathcal{L}_{\text{CoRFu}} = -\mathbb{E}\big[\log \sigma(S)\big] + \lambda \cdot \mathbb{E}\big[\mathbb{I}(S < 0) \cdot S^2\big]. \quad (11)$$

The first term is the standard DPO objective, which encourages the model π_θ to prefer y_w over y_l relative to a reference policy π_{ref}. The second term introduces an asymmetric penalty that activates only when the model ranks y_l above y_w ($S < 0$). Scaling this penalty quadratically in S disproportionately punishes high-confidence mistakes, aligning optimization pressure with clinical safety by discouraging unsafe misinterpretations of evidence.

4 Experimental Setup

4.1 Dataset

The final MediEval dataset is derived from MIMIC-IV v3.1 and consists of 2,015 unique hospital admissions, linked to 8,350 medical statements and yielding 37,144 samples (see Appendix B for distribution of relation types). The context length varies from 340 to 2,827 tokens. To prevent data leakage across splits, we retain only one admission per patient, since each patient may have multiple hospitalizations. The dataset is partitioned into training, validation, and test sets in an 80/10/10 ratio, balanced across the four quadrants. To ensure correctness, 200 test samples were randomly selected for human evaluation (details in Appendix C).

4.2 Large Language Models and Baselines

To enable systematic comparison across model sizes and training regimes, we selected 15 LLMs spanning proprietary, open-source, and biomedical domain-specific families. These include **GPT** (OpenAI) (OpenAI et al., 2024), **LLaMA** (Meta) (Grattafiori et al., 2024), **Mistral** (Mistral AI) (Jiang et al., 2024), **Qwen** (Alibaba) (Yang et al., 2025), and the biomedical models **Meditron** (Chen et al., 2023), **Med42** (Christophe et al., 2024), and **ClinicalCamel** (Toma et al., 2023). For implementing CoRFu, we used the **Llama-3.1–8B-Instruct** and **Qwen3-8B** models, with direct preference optimization (DPO) serving as a baseline.

4.3 Training Setups and Evaluation Metrics

LLM evaluation with supervised fine-tuning. All non-proprietary models are fine-tuned under an identical supervised setup to ensure comparability. Each model is equipped with a sequence classification head and trained on the four MediEval quadrants using weighted cross-entropy loss. LoRA adapters are applied for parameter-efficient fine-tuning, with uniform hyperparameters across models (batch size, learning rate, and number of epochs). Proprietary models are evaluated in their base zero-shot generative form without fine-tuning, and their results are shown in Appendix E.

Evaluation of CoRFu. We construct preference pairs where Q1 (supported truth) is always the preferred class. We explore three regimes:

- **Pairwise:** contrast Q1 with a specific error type (Q2, Q3, or Q4).

- **Mixed:** contrast Q1 against a pooled set of Q2/3/4 in a single stage.
- **Curriculum:** train sequentially on Q1 vs. Q2, then Q1 vs. Q3, and finally Q1 vs. Q4, reusing the updated model at each stage.

The mixed setup exposes the model to diverse error types simultaneously, while the curriculum regime introduces them in stages of increasing difficulty, reducing interference from heterogeneous signals. By encoding MediEval's safety perspective directly into both the loss and the training setups, CoRFu aligns model fine-tuning not only with preference correctness but also with clinical reasoning.

Evaluation Metrics. Performance is measured by accuracy, macro-precision, macro-recall, macro-F1, and per-quadrant F1. Furthermore, critical error rates are quantified by the hallucinated support rate (HSR) and the truth inversion rate (TIR), for which lower values indicate better safety.

5 Results and Discussion

5.1 MediEval Benchmarking Results

Overall Performance. Table 1 reports macro metrics (accuracy, precision, recall, F1) and safety scores (HSR, TIR) on MediEval. Overall performance is limited given the clinical reasoning requirements, with accuracies ranging from 59.3% to 73.9% and macro F1 from 50.3% to 70.7% across base models. The strongest model is `Llama-3.3-70B-Instruct`, which achieves the highest overall results in accuracy (73.9%), precision (73.3%), recall (73.2%), and F1 (70.7%). Among smaller models, `Llama-3.1-8B-Instruct` is competitive, reaching 61.5% macro F1. Although pretrained on clinical corpora, biomedical LLMs do not dominate. These results suggest that lexical familiarity and domain exposure alone are insufficient; fine-grained reasoning to disentangle factual correctness from contextual grounding requires explicit supervision.

Per-Quadrant Performance. Performance is uneven across the four quadrants, but consistent with their expected difficulty. **Q1** (Supported Fact) is easiest (best 86.7 by `Llama-3.3-70B`), consistent with models' strength on recognized truthful patterns in-context. **Q2** (Plausible Contradiction) is harder, with the best model below 70%, showing that LLMs often treat true but irrelevant statements as supported. **Q3** (Relational Mismatch) stresses relation-level reasoning; the best score reaches only 69.3, showing limited resistance to superficially coherent recombinations. Domain-specific models excel on **Q4** (Erroneous Information) (74.9 to 76.4%), likely because clinical pretraining gives them stronger priors for rejecting clearly false medical statements, though their broader reasoning across quadrants remains limited. Notably, even within a single model (e.g., `Llama-3.1-8B-Instruct`), performance varies sharply across quadrants, showing that strength in one type of reasoning does not transfer to others. This unevenness highlights persistent weaknesses in how models combine factuality with patient grounding.

Safety-Critical Errors. Safety-critical metrics expose risks not captured by aggregate F1. `Mixtral-8x7B` consistently achieves the lowest error rates, with HSR (Q2→Q1) of 20.5% and TIR (Q3→Q1) of 15.3%, whereas `Llama-3.2-3B` exhibits the highest risk on both metrics (HSR 40.9%, TIR 31.6%). This indicates that Mixtral is the safest model, even though it does not attain the highest macro F1. In contrast, `Llama-3.3-70B` achieves the best macro F1 (70.7%) yet fails to minimize HSR/TIR (21.2/21.1), demonstrating that higher accuracy does not necessarily translate into safer clinical reasoning. Overall, the divergence between F1 and safety metrics shows that models can perform well on average while still making risky errors.

5.2 CoRFu Evaluation Results

Overall Effects. Across both backbones, CoRFu outperforms the base SFT and DPO models, but the strongest gains concentrate in a single pairing. On `Llama-3.1-8B-Instruct`, **Q1 vs. Q2** achieves the best results at 76.8% accuracy and 77.9% macro F1 (+16.4 F1 over SFT), while other variants are lower (e.g., Q1 vs. Q3 at 72.9% F1, Q1 vs. mix at 69.6%, curriculum at 70.6%). For `Qwen3-8B`, the pattern is sharper: only **Q1 vs. Q2** yields a substantial aggregate improvement (70.7% Acc, 71.0% F1; +11.1 F1 over base), whereas Q1 vs. Q3, Q1 vs. Q4, and curriculum underperform on macro metrics. This shows that contrasting supported with unsupported truths provides the most reliable signal for improving overall reasoning, while other pairings mainly target specific weaknesses.

Per-Quadrant Improvements. CoRFu lifts per-quadrant F1 scores, but the biggest gains come from targeted pairings rather than curriculum. With the `Llama-3.1-8B-Instruct` backbone, **Q1 vs.**

Table 1: MediEval benchmarking results (best results underlined) and CoRFu results (best results in **bold**).

		Overall Performance				Per-Quadrant F1-Scores				Critical Error Rates	
Type	**Model**	**Acc.**	**Prec.**	**Rec.**	**F1**	**F1_Q1**	**F1_Q2**	**F1_Q3**	**F1_Q4**	**HSR**	**TIR**
Open-source (with SFT)	Llama-3.1-8B-Instruct	67.8	66.4	66.8	61.5	83.6	64.0	38.2	60.0	28.2	21.1
	Llama-3.2-1B-Instruct	64.1	57.6	64.3	50.3	80.9	0.0	68.1	52.2	27.3	15.8
	Llama-3.2-3B-Instruct	64.1	61.8	63.8	61.6	77.1	53.8	44.8	70.9	40.9	31.6
	Llama-3.3-70B-Instruct	73.9	73.3	73.2	70.7	86.7	70.0	65.9	60.0	21.2	21.1
	Mixtral-8x7B-Instruct-v0.1	65.4	69.6	66.0	63.8	51.4	68.3	69.3	66.4	20.5	15.3
	Mistral-7B-Instruct-v0.3	60.5	61.0	60.7	59.7	67.8	55.6	46.2	69.2	29.1	26.3
	Vicuna-13B-v1.5	59.3	59.3	59.2	59.3	62.6	56.1	66.8	51.6	22.7	15.5
	Qwen3-4B	61.7	61.3	61.7	61.4	71.7	49.0	63.3	61.6	31.8	26.3
	Qwen3-8B	63.7	64.4	63.7	59.9	70.0	58.6	50.0	60.8	28.2	31.1
	Qwen3-32B	62.9	64.6	63.8	62.1	72.9	62.6	51.4	61.3	28.2	21.1
Domain-specific (with SFT)	Meditron-70B	64.4	62.3	65.2	68.0	66.2	64.3	66.7	74.9	24.8	26.3
	Med42-70B	65.6	63.3	65.3	62.8	58.8	55.8	58.5	76.4	23.9	25.6
	ClinicalCamel-70B	62.0	61.7	62.6	62.0	55.4	54.3	63.3	75.0	24.8	25.6
Llama-3.1-8B-Ins.	+ DPO (Q1 vs. Q2)	65.6	59.7	64.2	59.5	53.5	55.1	56.7	72.7	32.7	27.9
	+ CoRFu (Q1 vs. Q2)	**76.8**	**77.2**	**77.0**	**77.9**	73.2	**76.6**	**78.9**	**79.9**	18.2	**0.0**
	+ CoRFu (Q1 vs. Q3)	72.7	66.8	74.2	72.9	**84.2**	69.1	71.1	67.1	20.1	11.5
	+ CoRFu (Q1 vs. Q4)	64.5	74.7	63.6	69.0	70.6	67.4	70.0	68.1	19.9	17.7
	+ CoRFu (Q1 vs. mix)	69.1	70.2	69.9	69.6	67.7	68.6	72.1	71.7	23.1	12.9
	+ CoRFu (Curriculum)	64.5	75.0	63.6	70.6	70.6	70.5	72.3	68.8	19.1	3.5
Qwen3-8B	+ DPO (Q1 vs. Q2)	51.0	51.6	50.9	51.1	51.2	49.1	56.7	47.4	22.7	4.4
	+ CoRFu (Q1 vs. Q2)	70.7	71.5	71.1	71.0	65.1	66.7	78.0	74.3	21.8	**0.0**
	+ CoRFu (Q1 vs. Q3)	61.1	52.0	59.9	53.0	52.9	55.3	49.3	54.4	22.3	8.3
	+ CoRFu (Q1 vs. Q4)	60.3	55.2	59.6	50.7	54.3	50.1	51.3	46.7	18.9	7.6
	+ CoRFu (Q1 vs. mix)	59.3	60.7	60.3	59.3	53.3	52.4	68.9	62.7	19.3	13.3
	+ CoRFu (Curriculum)	48.2	46.0	48.1	47.7	40.8	45.3	41.1	49.2	**17.3**	14.5

Q2 achieves the best balance across quadrants (Q1/Q2/Q3/Q4 = 73.2/76.6/78.9/79.9), improving both factual grounding (Q2) and overall accuracy. Targeted objectives produce predictable effects: training on (Q1 vs. Q2) substantially boosts Q2 F1 (+12.6 over base), while (Q1 vs. Q3) increases Q3 F1. These results show that targeted pairings not only correct specific weaknesses but also provide complementary improvements across quadrants.

Safety-Critical Reductions. Improvements in accuracy do not automatically translate into safety, but CoRFu reduces both HSR and TIR compared to SFT and DPO. On `Llama-3.1-8B`, the (Q1 vs. Q2) run achieves the lowest HSR at 18.2% and eliminates TIR entirely. On `Qwen-8B`, curriculum yields the lowest HSR at 17.3%, while (Q1 vs. Q2) again minimizes TIR to 0.0. These results show that different contrastive objectives selectively mitigate different risks, confirming that aggregate accuracy and safety-critical errors must be disentangled and addressed with tailored supervision.

Ablation Study. We conducted an ablation study for the value of λ. When $\lambda = 0$, the method reduces to vanilla DPO, with performance reported in Table 1. Values between 0.5 and 1.0 yield the best macro-F1 while significantly lowering HSR and TIR. For $\lambda > 1.0$, over-penalization occurs, leading to a decrease in macro-F1 and an increase in error rates. See Appendix F & G for more analyses.

Overall Insight. Taken together, the results show that contrastive objectives shape model behavior in complementary ways. Q1–Q2 training provides the strongest overall gains, improving accuracy, macro F1, and grounding while driving HSR to its lowest levels. Q1–Q3 supervision instead strengthens supported reasoning and sharply lowers TIR, though with smaller aggregate benefits. Curriculum is less competitive on headline metrics but can reduce HSR further for some backbones. Overall, these results demonstrate that contrastive fine-tuning can be directed to improve specific weaknesses, and that MediEval makes such effects visible by disentangling accuracy from safety-critical errors.

6 Conclusion

We introduced **MediEval**, a benchmark that jointly evaluates factual verification and patient-contextual reasoning by linking biomedical ontologies with real-world patient records. Our results show that current LLMs struggle with quadrant-level reasoning and make safety-critical errors not reflected in aggregate accuracy. To mitigate these risks, we proposed **Counterfactual Risk-Aware Fine-tuning (CoRFu)**, which reduces unsafe misclassifications through asymmetric preference optimization.

For future work, MediEval could be extended to other clinical tasks such as question answering, adapted to safety-critical domains beyond medicine, and integrated with methods such as retrieval-augmented generation or models designed for temporal patient reasoning. We hope this work provides a foundation for more rigorous and risk-aware evaluation of LLMs.

Limitations

Dataset size and coverage. MediEval is constructed from approximately 2k admissions in MIMIC-IV, yielding around 37k statements. This scale reflects a deliberate design choice to prioritize clinical plausibility, careful quality control, and strict prevention of patient-level information leakage, which are particularly important in safety-critical medical evaluation. The dataset construction pipeline is fully automated and released with the codebase, making MediEval readily extensible to larger cohorts, additional institutions, or alternative clinical subsets as future work.

Annotation and validation. A subset of the test set was independently reviewed by medically trained annotators, with GPT-5 used as an auxiliary reference to inspect disagreements. High agreement suggests that the ontology-grounded construction leads to largely unambiguous labels. While broader expert validation could further strengthen confidence, the reliance on structured biomedical knowledge and deterministic generation reduces subjectivity compared to fully manual annotation.

Data and ontology noise. As with any benchmark grounded in real-world clinical data and large biomedical ontologies, MediEval inherits noise and inconsistencies from MIMIC-IV and resources such as UMLS, SNOMED CT, and RxNorm. We mitigate these effects through ontology-guided plausibility checks, constrained multi-hop traversal, and explicit grounding in patient context. Further advances in ontology curation and clinical data quality are likely to directly benefit benchmarks such as MediEval.

References

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Position: Medical large language model benchmarks should prioritize construct validity. In *Forty-second International Conference on Machine Learning Position Paper Track*.

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

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, and 1 others. 2024. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.

Kevin Donnelly and 1 others. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maxime Griot, Jean Vanderdonckt, Demet Yuksel, and Coralie Hemptinne. 2025. Pattern recognition or medical knowledge? the problem with multiple-choice questions in medicine. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5321–5341.

Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *NPJ digital medicine*, 7(1):183.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. 2005. Rxnorm: prescription for electronic drug information exchange. *IT professional*, 7(5):17–23.

Jesús Lovón-Melgarejo, Martin Mouysset, Jo Oleiwan, Jose G Moreno, Christine Damase-Michel, and Lynda Tamine. 2025. Evaluating llm abilities to understand tabular electronic health records: A comprehensive study of patient data extraction and retrieval. In *European Conference on Information Retrieval*, pages 153–168. Springer.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. 2011. Normalized names for clinical drugs: Rxnorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4):441–448.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 6 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 53728–53741.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. Ai in health and medicine. *Nature medicine*, 28(1):31–38.

Andrey Sakhovskiy and Elena Tutubalina. 2025. Bali: Enhancing biomedical language representations through knowledge graph and language model alignment. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1152–1164.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2025. KGARevion: An AI agent for knowledge-intensive biomedical QA. In *The Thirteenth International Conference on Learning Representations*.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanchao Tan, Hang Lv, Yunfei Zhan, Guofang Ma, Bo Xiong, and Carl Yang. 2025. BoxLM: Unifying structures and semantics of medical concepts for diagnosis prediction in healthcare. In *Forty-second International Conference on Machine Learning*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Re, Sanmi Koyejo, and Nigam Shah. 2025. Context clues: Evaluating long context models for clinical prediction tasks on EHR data. In *The Thirteenth International Conference on Learning Representations*.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, Vienna, Austria. Association for Computational Linguistics.

Jiahuan Yan, Jintai Chen, Chaowen Hu, Bo Zheng, Yaojun Hu, Jimeng Sun, and Jian Wu. 2025. Small models are LLM knowledge triggers for medical tabular prediction. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Xiechi Zhang, Zetian Ouyang, Linlin Wang, Gerard De Melo, Zhu Cao, Xiaoling Wang, Ya Zhang, Yanfeng Wang, and Liang He. 2025. AutoMedEval: Harnessing language models for automatic medical capability evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6272–6285, Vienna, Austria. Association for Computational Linguistics.

Yuxuan Zhou, Xien Liu, Chen Ning, and Ji Wu. 2024. Multifaceteval: multifaceted evaluation to probe llms in mastering medical knowledge. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6669–6677.

Yuxuan Zhou, Xien Liu, Chen Ning, Xiao Zhang, and Ji Wu. 2025a. Reliable and diverse evaluation of LLM medical knowledge mastery. In *The Thirteenth International Conference on Learning Representations*.

Yuxuan Zhou, Xien Liu, Chenwei Yan, Chen Ning, Xiao Zhang, Boxun Li, Xiangling Fu, Shijin Wang, Guoping Hu, Yu Wang, and Ji Wu. 2025b. Evaluating LLMs across multi-cognitive levels: From medical knowledge mastery to scenario-based problem solving. In *Forty-second International Conference on Machine Learning*.

Appendices

A Real Samples for Each Quadrant

This appendix provides representative examples of MediEval samples from each of the four quadrants (see Figure 2 for Q2). Each example consists of a patient context extracted from a real MIMIC-IV discharge summary, a constructed medical statement, and its corresponding quadrant label.

Context:

History of Present Illness:
... patient with metastatic NSCLC to bone, admitted after Hct 23 with dizziness and fatigue. Reports poor sleep, anxiety about prognosis ...

Past Medical History:
... Hypertension, osteoporosis, GERD, tobacco dependence ...

Brief Hospital Course:
... Anemia: transfused 2 units PRBC → Hct improved. ... Pain: continued home oxycontin and oxycodone PRN. ... Anxiety: continued ativan 0.5 mg PRN. ... GERD: continued omeprazole (PPI). ... HTN: continued atenolol 25 mg ...

Discharge Medications:
... Omeprazole 20 mg daily
... Oxycodone ER 60 mg q12h
... Oxycodone 5 mg q6h PRN
... Lorazepam 0.5 mg q4h PRN
... Atenolol 25 mg daily
... Sucralfate 1 g QID ...

Discharge Diagnoses:
… Metastatic NSCLC, anemia, hypertension, osteoporosis, GERD …

Statement:
Gastroesophageal reflux disease may be treated by Omeprazole, which is a type of Proton Pump Inhibitor

Label:
Q1 True-Supported

Figure 3: Example of statement verification against patient records (Quadrant 1: True-Supported). The statement is medically correct, and GERD is indeed treated with omeprazole, making it true and supported.

Context:

History of Present Illness:
... patient with metastatic NSCLC to bone, admitted after Hct 23 with dizziness and fatigue. Reports poor sleep, anxiety about prognosis ...

Past Medical History:
... Hypertension, osteoporosis, GERD, tobacco dependence ...

Brief Hospital Course:
... Anemia: transfused 2 units PRBC → Hct improved. ... Pain: continued home oxycontin and oxycodone PRN. ... Anxiety: continued ativan 0.5 mg PRN. ... GERD: continued omeprazole (PPI). ... HTN: continued atenolol 25 mg ...

Discharge Medications:
... Omeprazole 20 mg daily
... Oxycodone ER 60 mg q12h
... Oxycodone 5 mg q6h PRN
... Lorazepam 0.5 mg q4h PRN
... Atenolol 25 mg daily
... Sucralfate 1 g QID ...

Discharge Diagnoses:
… Metastatic NSCLC, anemia, hypertension, osteoporosis, GERD …

Statement:
Gastroesophageal reflux disease may be treated by Atenolol

Label:
Q3 False-Supported

Figure 4: Example of statement verification against patient records (Quadrant 3: False-Supported). The crafted statement is medically incorrect, as Atenolol is indicated for hypertension. Since both GERD and Atenolol appear in the patient context, the statement may seem supported, even though it is false.

Context:

History of Present Illness:
... patient with metastatic NSCLC to bone, admitted after Hct 23 with dizziness and fatigue. Reports poor sleep, anxiety about prognosis ...

Past Medical History:
... Hypertension, osteoporosis, GERD, tobacco dependence ...

Brief Hospital Course:
... Anemia: transfused 2 units PRBC → Hct improved. ... Pain: continued home oxycontin and oxycodone PRN. ... Anxiety: continued ativan 0.5 mg PRN. ... GERD: continued omeprazole (PPI). ... HTN: continued atenolol 25 mg ...

Discharge Medications:
... Omeprazole 20 mg daily
... Oxycodone ER 60 mg q12h
... Oxycodone 5 mg q6h PRN
... Lorazepam 0.5 mg q4h PRN
... Atenolol 25 mg daily
... Sucralfate 1 g QID ...

Discharge Diagnoses:
... Metastatic NSCLC, anemia, hypertension, osteoporosis, GERD ...

Statement:
Gastroesophageal reflux disease may be treated by insulin

Label:
Q4 False-Unsupported

Figure 5: Example of statement verification against patient records (Quadrant 4; False-Unsupported). The statement is medically incorrect, since insulin is not indicated for GERD. Moreover, insulin does not appear in the patient context, making the statement false and unsupported.

B Distribution of Relation Types

This appendix reports the distribution of biomedical relation types used in MediEval. Relations are derived from UMLS and associated vocabularies after semantic normalization and ontology-guided traversal. The distribution reflects the clinical diversity of the benchmark, covering treatment, diagnostic, pharmacologic, and associative relations.

Table 2: Distribution of relation types in the dataset

Relation Type	Count
treats	17,540
may_be_treated_by	4,552
related_to	4,008
classified_as	3,774
contraindicated_class_of	1,250
subset_includes	1,568
has_pharmacologic_class	854
diagnoses	738
prevents	748
causes	808
associated_with	588
authorized_value	282
co-occurs_with	130
defining_characteristic_of	230
manifestation_of	56
sign_or_symptom_of	12
may_be_diagnosed_by	4
part_of	2
Total	**37,144**

C Human Evaluation

We conducted human validation on 200 balanced samples. Two medically trained annotators labeled each item independently. In cases of disagreement, GPT-5 was consulted as an auxiliary signal for sanity checking, but final labels were determined by the human annotators. Agreement was 97% (194/200), with Cohen's $kappa \approx 0.96$, indicating strong agreement. All disagreements occurred between Q2 and Q3, which is one of the subtle distinctions that the benchmark is designed to probe. Because our extraction pipeline is fully ontology-driven and deterministic, the errors stem from noise or inconsistencies in the underlying ontologies or EHR data.

D Implementation Details

All training was performed on an NVIDIA L40S GPU with 46 GB memory. All reported results are averaged over 3 random seeds (42, 43, 44). For the MediEval supervised baselines, we fine-tuned models for 3 epochs with per-device batch size 4, gradient accumulation 8, maximum sequence length 4096, learning rate 2×10^{-5}, weight decay 0.01, and warmup ratio 0.03. When enabled, LoRA used rank $r = 8$, $\alpha = 16$, and dropout 0.05. For the CoRFu experiments, we fine-tuned with maximum sequence length 4096, per-device batch size 1, gradient accumulation 16 (effective batch size 16), and trained for 1 epoch. We used the AdamW optimizer with learning rate 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01, warmup ratio 0.05, and gradient clipping at 1.0. LoRA was configured with rank $r = 16$, $\alpha = 32$, and dropout 0.1. The CoRFu loss used $\beta = 0.1$ and $\lambda = 0.5$.

E Evaluation of Proprietary GPT Models

GPT-4o and GPT-5 are evaluated under a realistic clinical setting, following their typical usage: zero-shot inference without task-specific fine-tuning. This reflects practical deployment constraints, where adapting proprietary models is often infeasible. Our goal is not a direct head-to-head comparison, but rather to highlight two complementary observations: (1) the task remains challenging even for frontier models without adaptation, and (2) smaller models equipped with targeted supervision (SFT/CoRFu) can outperform general-purpose LLMs on safety-critical dimensions.

For proprietary models, the prompt consists of the following components: (1) context, (2) state-

Table 3: MediEval benchmarking results of Proprietary GPT Models.

	Overall Performance				Per-Quadrant F1-Scores				Critical Error Rates	
Model	**Acc.**	**Prec.**	**Rec.**	**F1**	**F1_Q1**	**F1_Q2**	**F1_Q3**	**F1_Q4**	**HSR**	**TIR**
GPT-5	45.7	36.4	46.7	30.4	68.3	53.3	0.0	0.0	19.3	0.0
GPT-4o	49.3	31.6	50.0	29.1	66.3	50.0	0.0	0.0	46.7	0.0

ment, (3) definitions of the four quadrants, (4) an instruction to select exactly one label, and (5) the list of candidate labels.

Overall, proprietary models struggle on this task. GPT models achieve only 29.1% (GPT-4o) and 30.4% (GPT-5) macro-F1. Since both models are evaluated in their base generative form without task-specific adaptation, this performance reflects both the difficulty of MediEval and the lack of alignment to the evaluation protocol.

In terms of safety errors, HSR (Q2→Q1) ranges from 19.3% for GPT-5 to 46.7% for GPT-4o. Notably, proprietary models never predict Q3, resulting in zero TIR. This behavior indicates miscalibration rather than genuine safety awareness.

F Ablation Study

Figure 6 illustrates the effect of the regularization coefficient λ on performance and error rates, under Q1 vs. Q2 CoRFu training. For both models, moderate values of λ (around 0.5–1.0) achieve the best trade-off: macro-F1 is maximized, while HSR and TIR are substantially reduced compared to $\lambda = 0$, which corresponds to vanilla DPO. As λ increases further, performance degrades and error rates rise, indicating over-penalization of negative preference margins. This ablation is necessary because λ directly controls the strength of the corrective penalty introduced by CoRFu, balancing preference optimization against error suppression. In contrast, we fix the DPO temperature β, since it only rescales the preference margin and its effect can be absorbed into λ.

G Qualitative Evaluation

We further analyze whether model behavior correlates with input characteristics such as context length or relation type, across all evaluated settings, including SFT and CoRFu. We observe no consistent correlation between any of the reported evaluation metrics and context length, nor systematic performance differences across relation types for any model. These results indicate that the observed performance trends are not driven by input

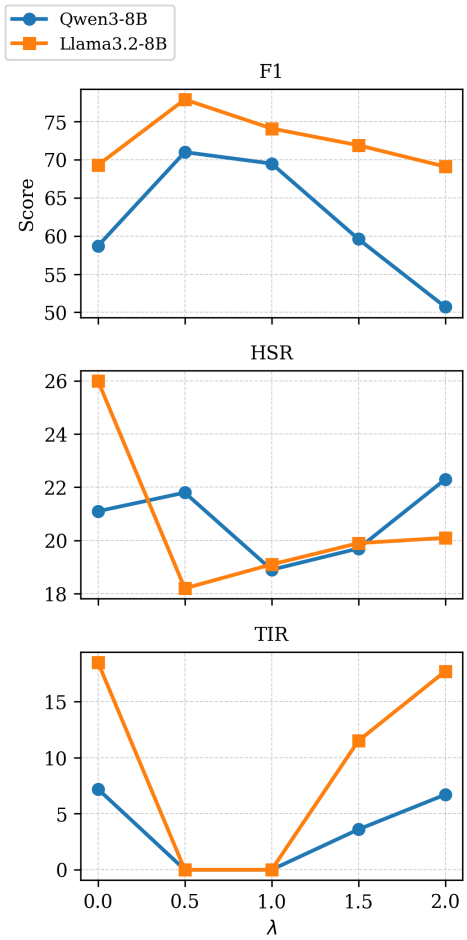

Figure 6: Effect of the regularization coefficient λ on macro-F1, HSR, and TIR for Qwen3-8B and Llama-3.2-8B under Q1 vs. Q2 CoRFu training.

complexity or relation semantics, but instead reflect differences in training objectives and supervision strategies.

Instead, errors predominantly arise from confusions between Q2 and Q1, indicating difficulties in distinguishing between partially incorrect and fully incorrect cases. We provide a qualitative analysis of these failure modes below.

Qualitative Example - hadm_id: 24480054

Statement. *Gastroesophageal reflux disease may be treated by aluminum hydroxide, which*

is a type of antacid.

Context.

```
History of Present Illness:
___ is a G4P3 female with a history of
thrombophilia and cerebrovascular accident
who presented to an outside hospital with
worsening epigastric pain without overt
signs of labor. She was admitted in early
labor and subsequently delivered her baby on
the morning of ___. However, her abdominal
pain continued to worsen during
hospitalization.

She has a history of gastroesophageal reflux
disease previously treated effectively with
pantoprazole, later transitioned to
omeprazole due to insurance changes, with
recurrence of symptoms. During the first
trimester of pregnancy, she experienced
worsening epigastric pain with significant
nausea and vomiting. After the first
trimester, nausea and vomiting improved, and
her heartburn and epigastric pain also
improved. In the week prior to delivery, her
symptoms worsened again, with pain primarily
localized to the epigastrium and exacerbated
by movement. She attempted over-the-counter
Tums and Pepcid without relief.

At the outside hospital, she underwent
abdominal ultrasound and CT imaging
demonstrating pancreatitis. She was noted to
have leukocytosis and received imipenem
prior to transfer. She received at least 4L
of normal saline and voided at least 350cc.
She reported bilious emesis but denied
ongoing nausea.

Her delivery was uncomplicated. Her newborn
son, ___, is healthy. No estimated blood
loss was recorded, and delivery was vaginal.
She continued to experience suprapubic pain.

On arrival to the MICU, vital signs were:
temperature 97.9F, blood pressure 156/97,
heart rate 107, respiratory rate 16, and
SpO2 97% on nasal cannula. She reported
abdominal pain that was tolerable and
improved with hydromorphone. She denied
shortness of breath. She had not passed gas,
and her last bowel movement was ___.

Past Medical History:
- Stroke associated with Depo-Provera,
without residual deficits
- Thrombophilia (MTHFR deficiency)
- Hepatitis C (cleared virus; viral load
undetectable, antibody positive; no IVDU;
history of blood transfusion ___)
- Ruptured appendix ___
- Morbid obesity
- Gastroesophageal reflux disease
- Pregnancy in ___ with possible fetal
alcohol syndrome, ADHD, and bipolar
disorder; no issues in ___ or ___ pregnancies

Brief Hospital Course:
Severe Acute Pancreatitis: The patient was
admitted to the MICU and kept NPO. MRCP
demonstrated greater than 30% pancreatic
necrosis with a hemorrhagic component. A
nasojejunal tube was placed, and tube feeds
were initiated. Initial intolerance improved
with cycling. No drainable fluid collections
were identified. Gastroenterology followed
throughout admission. She was discharged on
a clear liquid diet with plans to advance to
a low-fat diet. Persistent left upper
quadrant pain gradually improved.

Thrombophilia: The patient has a history of
MTHFR deficiency and prior stroke. A new
non-occlusive portal vein thrombosis was
identified during admission. Therapeutic
enoxaparin was initiated after stabilization
of hemorrhagic pancreatitis, with a planned
duration of ___ months and hematology
follow-up.

Postpartum Status: The patient had an
uncomplicated vaginal delivery of a healthy
infant on ___. Social work was involved due
to limited home support.

Acute Kidney Injury: An acute kidney injury
was present on admission, likely secondary
to pancreatitis, and resolved with
intravenous fluid resuscitation.

Ileus and Diarrhea: An initial ileus was
treated conservatively. The patient later
developed tube-feed--associated diarrhea.
Clostridioides difficile testing was
negative. Symptoms improved with as-needed
loperamide.

New-Onset Diabetes Mellitus: Hyperglycemia
developed after initiation of tube feeds,
likely secondary to severe pancreatitis. The
patient was started on an insulin sliding
scale. Due to clinical complexity, she was
discharged on a Humalog sliding scale only
and requires outpatient reassessment.

Inactive Issues:
Gastroesophageal reflux disease: continue
pantoprazole.

Discharge Diagnoses:
- Acute severe pancreatitis
- Hyperglycemia, likely new-onset diabetes
mellitus
- Portal vein thrombosis
- Acute kidney injury, resolved
```

Label. True-Supported

The statement is medically correct in general; however, the specific treatment (aluminum hydroxide) is not mentioned in the clinical context. While GERD is documented and treated with pantoprazole and other antacids, the generated statement introduces an unsupported medication, and is there-

fore labeled as Q2.

Almost all models fail on this example due to a combination of lexical bias and treatment-level over-generalization. The presence of strong surface cues such as GERD, reflux, and references to antacid use biases models toward conceptual matching, leading them to infer support even when the specific medication (aluminum hydroxide) is not mentioned.

In addition, models tend to collapse distinctions within therapeutic classes, treating different antacids as interchangeable. This error is further exacerbated by clinical complexity rather than context length: the note is dominated by severe acute pancreatitis and post-partum complications, while GERD appears as a secondary issue, making fine-grained verification of drug-level evidence particularly challenging. Together, these factors reveal a persistent difficulty in jointly verifying medical correctness and strict contextual grounding.