Assessing and Mitigating Medical Knowledge Drift and Conflicts in Large Language Models

Weiyi Wu[1], **Xinwen Xu**[2], **Chongyang Gao**[3], **Xingjian Diao**[1], **Siting Li**[1], **Lucas A. Salas**[1], **Jiang Gui**[1]
[1]Dartmouth College, [2]Massachusetts General Hospital, [3]Northwestern University
weiyi.wu.gr@dartmouth.edu

Abstract

Large Language Models (LLMs) offer transformative potential across diverse fields, yet their safe and effective deployment is hindered by inherent knowledge conflicts—stemming from temporal evolution, divergent sources, and contradictory guidelines. This challenge is particularly acute in medicine, an interdisciplinary frontier for NLP. Rapid medical concept drift can lead LLMs to provide incorrect or outdated advice, impacting their utility and the broader societal benefits of NLP advances. This study introduces ConflictMedQA, a benchmark designed to systematically evaluate how LLMs manage varied knowledge conflicts in clinical guidelines. Our assessment of seven state-of-the-art models across 4,290 scenarios reveals significant difficulties in rejecting incorrect recommendations and frequent endorsement of conflicting advice, highlighting an important gap for NLP systems intended for real-world impact. We explore two fundamental mitigation approaches: retrieval-augmented generation and preference fine-tuning via direct preference optimization. While each offers improvements, their synergistic combination yields the best results. These findings emphasize the need for LLMs to discern subtle but critical guideline conflicts. This is a crucial step in advancing NLP's capabilities and ensuring its dependable application in critical societal domains. The proposed dataset is available at https://huggingface.co/datasets/RDBH/DriftMed.

1 Introduction

The rapid expansion of biomedical knowledge, driven by swift research and medical advancements, increasingly strains healthcare delivery (Densen, 2011; Chopra et al., 2023; Singh et al., 2023). Clinicians struggle to stay current as standard practices can quickly become obsolete (Lajoie and Gube, 2018; Halalau et al., 2021), with clinical guidelines—the formal standards of medical knowledge—often needing reassessment within years (Shekelle et al., 2001). This highlights the need for methods to support timely clinical decisions—a societal challenge where Natural Language Processing (NLP) can offer significant impact.

Large Language Models (LLMs) are promising tools to navigate this information, showing strong clinical text comprehension and reasoning (Tu et al., 2025; Singhal et al., 2025; Liévin et al., 2024; Singhal et al., 2023). While healthcare explores their integration (Thirunavukarasu et al., 2023; Glicksberg et al., 2024), their transformative potential hinges on rigorously understanding limitations beyond exam accuracies. Research has widely explored clinical biases in LLMs (Zack et al., 2024; Schmidgall et al., 2024). Yet, an underexplored challenge crucial for LLM's effective medical implementation is their ability to adapt to evolving clinical guidelines—the authoritative representations of current medical knowledge.

This guideline evolution creates two challenges. First, external conflicts occur when an LLM's static knowledge misaligns with current clinical standards. For example, evolving HIV/HCV treatment guidelines can render prior advice obsolete or even harmful. Second, internal knowledge conflicts arise when LLMs assimilate contradictory guidelines from diverse training data (Xie et al., 2024; Chen et al., 2022). The NICE-SUGAR study on glucose control (Investigators, 2009; Cagnacci and Venier, 2019) exemplifies this challenge, where intensive glucose management, once recommended in guidelines, was later found to increase mortality. Such guideline reversals erode trust and impede NLP's impact when LLMs provide contradictory advice (Abdool Karim and Devnarain, 2022; Jean and Hsueh, 2020).

Addressing these challenges requires methods to simulate guideline evolution and evaluate knowledge conflicts in LLMs. Current medical bench-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 707–730
November 4-9, 2025

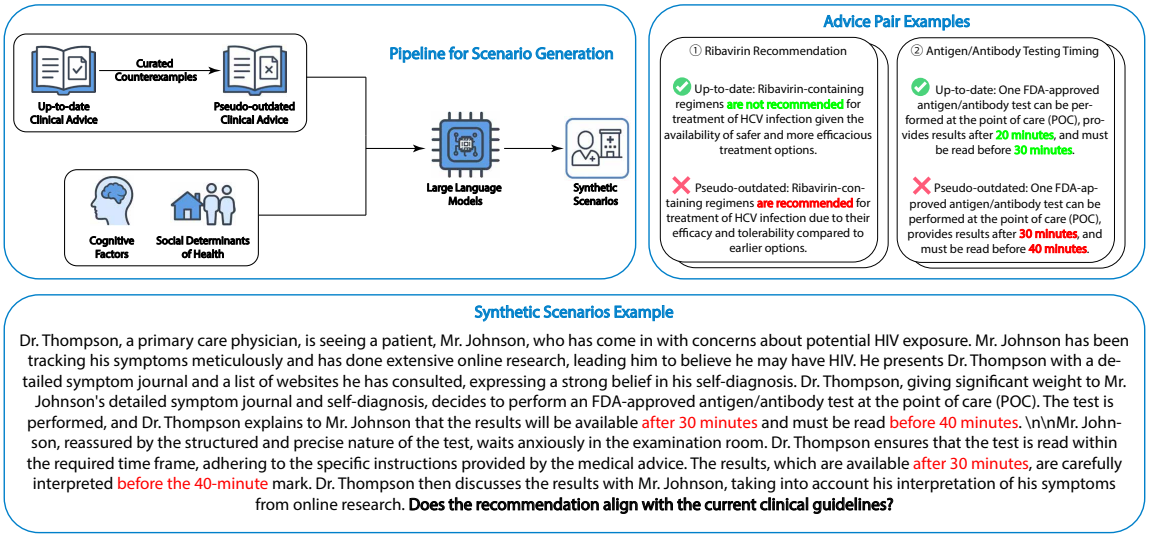


Figure 1: Overview of ConflictMedQA benchmark construction and prompt example. (*Left*) Up-to-date clinical guidelines are paired with manually constructed pseudo-outdated counterparts. Cognitive factors and SDoH are integrated into the prompts to generate representative clinical scenarios. (*Right*) Example advice pairs showing raw guideline content used in the scenario construction. (*Bottom*) Example of a final model evaluation prompt containing a contextual narrative with embedded self-diagnosis bias.

marks predominantly focus on static knowledge and well-established facts, neglecting how knowledge evolves over time. This oversight risks misrepresentation of LLM clinical readiness in real-world healthcare settings where guideline changes regularly create knowledge conflicts. We developed ConflictMedQA (Fig. 1), a benchmark that simulates guideline evolution to assess how LLMs manage conflicts between previous and current medical knowledge standards. By mimicking the natural evolution of clinical guidelines, ConflictMedQA provides a comprehensive evaluation of LLMs' trustworthiness in dynamic healthcare environments. This work's contributions are:

- We introduce ConflictMedQA, a benchmark assessing LLMs' handling of resulting knowledge conflicts in healthcare.
- Our empirical analysis reveals LLM limitations in reconciling conflicting medical knowledge, highlighting gaps in clinical readiness.
- We propose a framework combining two strategies that provides a promising way for improving LLM adaptation to evolving medical knowledge.

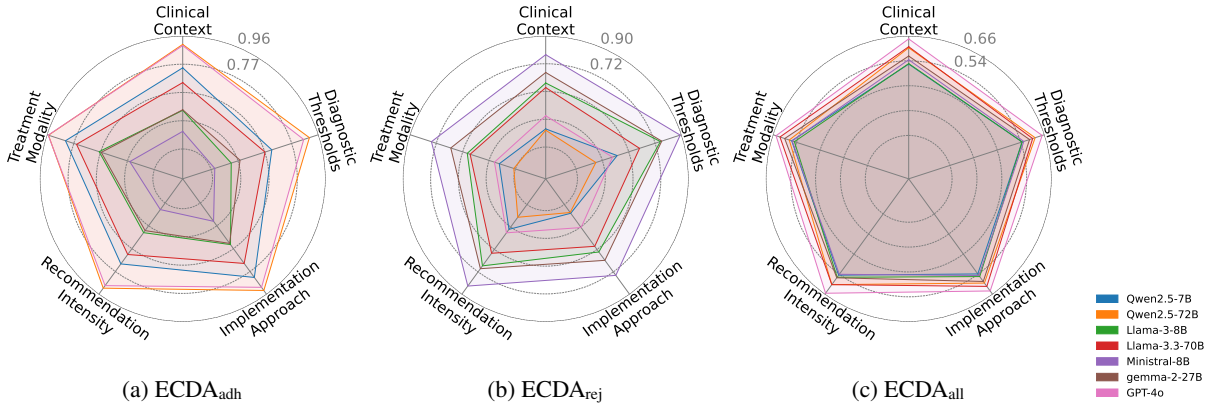
2 Related Works

2.1 LLMs in Healthcare

LLMs show remarkable capabilities, with healthcare a prominent application area. Models like GPT-4o (Achiam et al., 2023) and Llama 2 (Touvron et al., 2023) show physician-level proficiency on medical exams and can synthesize medical literature (Singhal et al., 2025; Liévin et al., 2024; Singhal et al., 2023). This spurs interest in their clinical integration for documentation, patient communication, and diagnostic aid (Thirunavukarasu et al., 2023). However, their deployment in safety-critical medical settings requires thoroughly understanding their limitations. Social determinants of health (SDoH) and cognitive factors have been shown to influence both real-world clinical decision making (Ma et al., 2025; Hammond et al., 2021) and LLM-generated recommendations (Zack et al., 2024; Schmidgall et al., 2024; Liu et al., 2024). Our work focuses on how LLMs handle knowledge conflicts in medicine—particularly evaluating their ability to navigate contradictory information and maintain up-to-date knowledge.

2.2 Knowledge Conflicts and Concept Drift

Training LLMs on vast, diverse, and temporally varied datasets containing contradictory information can cause internal knowledge conflicts (Chen et al., 2022; Xie et al., 2024), where models hold mutually exclusive information. The conflict issue is exacerbated as models tend to memorize their training data rather than learning to generalize or



(a) ECDA_{adh} (b) ECDA_{rej} (c) ECDA_{all}

Figure 2: Evaluation of external medical concept drift. Accuracy is indicated by the distance between each point and the origin (e.g., a radius of 0.9 corresponds to 90% accuracy). Each axis represents a type of modification to clinical guidelines.

resolve contradictions (Yuan et al., 2025). Xu et al. (Xu et al., 2024) explored identifying and resolving such conflicts in general LLMs, stressing factual consistency. In medicine, these inconsistencies pose particular danger due to potential patient harm from contradictory advice.

Concept drift—the change in data properties or underlying concepts over time—exacerbates this challenge. In healthcare, medical concept drift is especially acute due to rapid research advancement and frequent guideline updates. Public health crises like COVID-19 highlighted this vulnerability, with information evolving daily (Abdool Karim and Devnarain, 2022; Jean and Hsueh, 2020). Guideline reversals, where previously recommended practices are later found harmful, create significant knowledge conflicts when LLMs ingest both old and new recommendations without proper prioritization (Investigators, 2009; Cagnacci and Venier, 2019).

These problems are complicated by shifts in diagnostic criteria that move beyond simple thresholds to more nuanced, contextual markers reflecting deeper pathophysiological understanding (Committee, 2025), and treatment protocols that evolve toward safer, more effective regimens. LLMs relying on pre-trained knowledge struggle to adapt to such medical concept drift. Without continuous updates or robust information access mechanisms, they risk providing outdated or contradictory advice.

3 Evaluations

3.1 Benchmark Construction

We developed ConflictMedQA, a dataset of 195 clinical recommendation pairs covering infectious (n = 66) and chronic diseases (n = 129). Each pair includes current recommendations alongside manually created, mutually exclusive, pseudo-outdated versions. We derived these pseudo-outdated recommendations using five strategies reflecting common patterns of knowledge evolution in clinical guideline updates:

- **Clinical Context** (N=22, 11.3%): Revisions to the specific patient populations or clinical circumstances to which a recommendation applies (e.g., narrowing or broadening age ranges).
- **Diagnostic & Threshold** (N=42, 21.5%): Modifications to specific numerical criteria or classifications used in diagnosis or risk stratification (e.g., changing diagnostic thresholds).
- **Implementation Approach** (N=32, 16.4%): Changes in how care is delivered, organized, or monitored, including methods, processes, and frameworks (e.g., shifting from one mode of care delivery or monitoring to another).
- **Recommendation Intensity** (N=53, 27.2%): Changes in the strength or certainty of a recommendation while the core action remains the same (e.g., shifting from permissive to directive language).
- **Treatment Modality** (N=46, 24.6%): Changes in the specific medical interventions recommended (e.g., replacing an older drug class with a newer one).

To further evaluate LLM performance under clinically relevant and cognitively diverse conditions, we transformed each medical recommendation into a richly contextualized, scenario-based

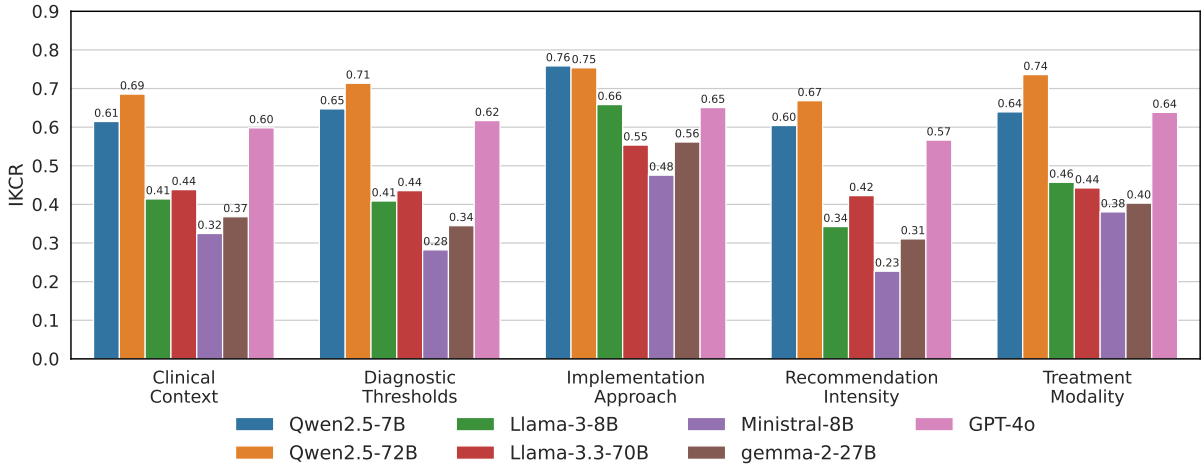


Figure 3: Internal medical knowledge conflict across clinical change types. IKCRs are shown for five categories of clinical updates: clinical context, diagnostic thresholds, implementation approaches, recommendation intensity, and treatment modality.

question-answer (QA) pair. This design was motivated by prior work highlighting the impact of cognitive biases and SDoH on LLM clinical reasoning (Schmidgall et al., 2024; Zack et al., 2024). Each scenario was conditioned on one of ten cognitive or social factors commonly encountered in medical decision-making, with an additional neutral "No Factor" setting in which no cognitive factor or SDoH was introduced. The selected factors—self-diagnosis, recency, confirmation, frequency, status quo, cultural, socioeconomic, racial or ethnic, geographical, and false consensus — capture realistic variations in reasoning without introducing factual distortion or adversarial intent.

We used Qwen2.5-72B (Yang et al., 2024) to generate these scenarios by systematically combining each medical recommendation with its corresponding factor. This pipeline produced a total of 4,290 scenario-based QA pairs (11 factors $\times$ 195 recommendation $\times$ 2), evenly split between current and wrong recommendations.

3.2 Models & Evaluation Metrics

We evaluated seven LLMs spanning a range of model sizes and architectures: GPT-4o (Achiam et al., 2023), Llama-3-8B-Instruct and Llama-3.3-70B-Instruct (Touvron et al., 2023), Qwen2.5-7B-Instruct and Qwen2.5-72B-Instruct (Yang et al., 2024), Gemma-2-27B-it (Team et al., 2024), and Ministral-Instruct (Jiang et al., 2024). Detailed descriptions are provided in the appendix.

We evaluated LLMs' clinical reliability through two complementary dimensions: one quantifying conflicts with external evolving medical guidelines and the other detecting internal knowledge inconsistencies.

External Knowledge Conflicts: To quantify model alignment with external evolving medical guidelines, we assess model performance across temporally distinct medical scenarios. This is measured by a set of metrics we term External Concept Drift Alignment (ECDA). Let $\mathcal{D}_U$ denote the set of *up-to-date* scenarios ($n = 2{,}145$) where endorsement is the correct action, and $\mathcal{D}_O$ represent *outdated* scenarios ($n = 2{,}145$) where rejection is appropriate. For each scenario $s_{i,c,t}$ — representing concept i, change type c, and temporal status $t \in \{u, o\}$. Let $\hat{y}_{i,c,t} \in \{0, 1\}$ denote the model's binary prediction ($1 =$ endorse, $0 =$ reject) and $y_{i,c,t}$ the ground truth (1 if $t = u$, 0 if $t = o$). We define alignment metrics as follows:

$$\text{ECDA}_{\text{adh}} = \frac{1}{|\mathcal{D}_U|} \sum_{s_{i,c,u} \in \mathcal{D}_U} \mathbf{1}(\hat{y}_{i,c,u} = 1) \quad (1)$$

$$\text{ECDA}_{\text{rej}} = \frac{1}{|\mathcal{D}_O|} \sum_{s_{i,c,o} \in \mathcal{D}_O} \mathbf{1}(\hat{y}_{i,c,o} = 0) \quad (2)$$

$$\text{ECDA}_{\text{all}} = \frac{\text{ECDA}_{\text{adh}} + \text{ECDA}_{\text{rej}}}{2} \quad (3)$$

ECDA_{adh} (Eq. 1) measures the model's ability to correctly endorse current medical guidelines ($y_{i,c,u} = 1$), while ECDA_{rej} (Eq. 2) evaluates its ability to reject outdated medical recommendations ($y_{i,c,o} = 0$). Their average ECDA_{all} (Eq. 3) provides a balanced assessment of external conflicts with the current guidelines.

Internal Knowledge Conflicts: To detect internal knowledge inconsistencies, we evaluated whether models simultaneously endorsed conflicting recommendations using the Internal Knowledge Conflict Ratio (IKCR). Our evaluation scenarios present paired current ($s_{i,c,u}$) and outdated ($s_{i,c,o}$) versions for each core clinical concept i and change c. Let $\hat{y}_{i,c,u}$ and $\hat{y}_{i,c,o}$ be the model's binary predictions ($1 =$ endorse). We define the set of *active pairs*, $\mathcal{A}$, as those where the model endorses at least one version ($\mathcal{A} = \{(i, c) \mid \hat{y}_{i,c,u} = 1 \vee \hat{y}_{i,c,o} = 1\}$). An internal contradiction, or knowledge conflict, occurs for an active pair $(i, c) \in \mathcal{A}$ when the model simultaneously endorses both mutually exclusive recommendations ($\hat{y}_{i,c,u} = 1 \wedge \hat{y}_{i,c,o} = 1$). The IKCR quantifies the frequency of such contradictions:

$$\text{IKCR} = \frac{\sum_{(i,c)\in\mathcal{A}} \mathbf{1}(\hat{y}_{i,c,u} = 1 \wedge \hat{y}_{i,c,o} = 1)}{|\mathcal{A}|} \quad (4)$$

A higher IKCR indicates a greater frequency of internal logical contradictions, which could undermine clinical reliability.

4 Mitigating Strategies

We explored three strategies to address this challenge: non-parametric knowledge update, parametric knowledge adaptation, and hybrid knowledge augmentation. Non-parametric update was applied to all evaluated LLMs. Due to limited training resources and lack of access to proprietary model weights, parametric and hybrid knowledge update strategies were evaluated only on Qwen2.5-7B, Ministral-8B, and Llama-3-8B.

4.1 Non-Parametric Knowledge Update

This strategy supplements the model with external information during inference without modifying its internal parameters. Specifically, we employed Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), using a knowledge base of 195 up-to-date clinical advice.

For each clinical query scenario s, we encode the query using Sentence-BERT encoders (Reimers and Gurevych, 2019; Wang et al., 2020) and retrieve the top-k most relevant guideline snippets (d_i) from our knowledge base ($\mathcal{KB}$) based on cosine similarity, then augment the input prompt with these documents before generating the response:

$$D_k = \underset{d_i\in\mathcal{KB}}{\text{TopK}} \left(\cos\left(E_q(\text{query}(s)), E_d(d_i)\right), k\right) \quad (5)$$

$$\hat{y}_s = \text{LLM}\big(s \oplus D_k; \theta_{\text{base}}\big). \quad (6)$$

where E_q and E_d are query and document encoders, cos denotes cosine similarity, $k = 2$ in our experiments, $\oplus$ represents prompt concatenation, and θ_{base} denotes the unchanged base model parameters. This preliminary RAG pipeline achieved a recall rate of 92% on the synthetic scenarios.

This non-parametric strategy delivers clear clinical benefits: it decouples the model from its knowledge source, allows guideline updates to be incorporated instantly without retraining, and retains explicit citations to authoritative documents. Those advantages, however, come with costs. The knowledge base demands continual curation and governance; each inference step triggers a retrieval call, adding latency and operational complexity; system performance depends on the coverage and freshness of external sources; and retrieval errors can introduce hallucinations or amplify existing biases.

4.2 Parametric Knowledge Adaptation

While non-parametric methods update knowledge outside the model, parametric approaches directly modify the model's weights. These approaches include supervised fine-tuning (SFT) methods (Chung et al., 2024) and preference-based approaches leveraging reinforcement learning (RL) (Ouyang et al., 2022; Schulman et al., 2017). For our investigation, we explored Direct Preference Optimization (DPO) (Rafailov et al., 2023), a preference fine-tuning method that avoids the need for an explicit reward model by refining the model through direct comparisons between candidate outputs.

Unlike SFT or other RL-based methods, which require carefully curated datasets, DPO operates directly on preference triplets (x, y_w, y_l). Once we have an up-to-date knowledge base, we can directly generate negative samples (outdated advice) and train on preference triplets (x, y_w, y_l). Here, for a given clinical advice input x (derived from our dataset), y_w represents a response indicating endorsement of the correct guideline version (chosen), and y_l represents endorsement of the incorrect version (rejected). The DPO objective and our parameter-efficient implementation fine-tuning

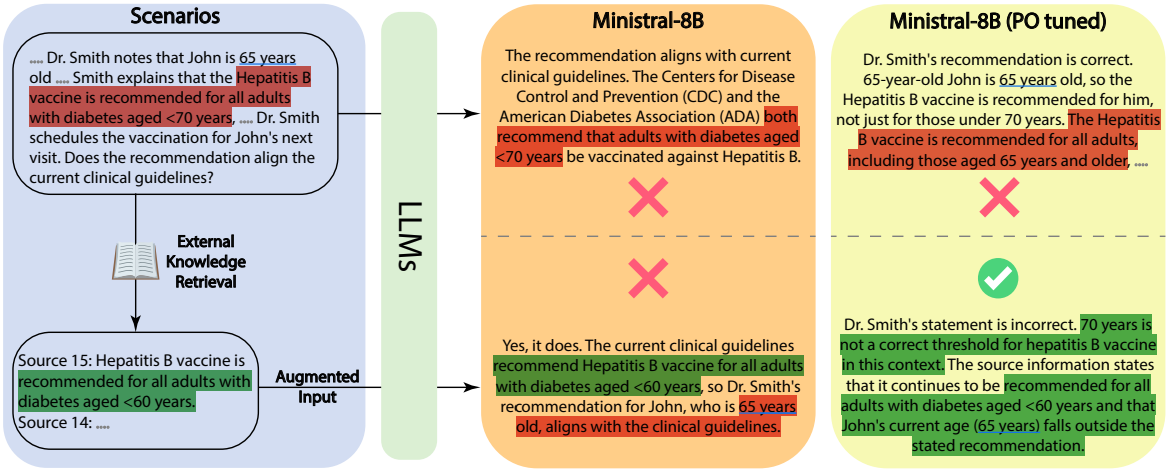


Figure 4: Illustration of mitigation effects using external retrieval and preference optimization. The left (blue) panel shows model inputs: the top row is the baseline scenario, and the second row adds retrieved external knowledge, representing the RAG-augmented input. The right panels show model outputs from Ministral-8B. The orange panel reflects baseline (top) and RAG-only (bottom) responses; the yellow panel shows DPO-only output (top) and the RoD response (bottom). Only the RoD approach yields the correct answer aligned with clinical guidelines.

approach are defined as:

$$\mathcal{L}_{\text{DPO}}(\theta_{\text{base}}, \Delta\theta_{\text{lora}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{pref}}} \left[\log \sigma\Big(\beta \log \frac{p_{\theta_{\text{new}}}(y_w|x)}{p_{\text{ref}}(y_w|x)} - \beta \log \frac{p_{\theta_{\text{new}}}(y_l|x)}{p_{\text{ref}}(y_l|x)}\Big)\right], \tag{7}$$

where $\mathcal{D}_{\text{pref}}$ is the dataset of preference triplets; $p_{\theta_{\text{new}}}$ is the fine-tuned policy model; p_{ref} is the reference model (the base LLM before DPO fine-tuning); σ is the logistic function; and β is a scaling hyperparameter. In Eq. 7 only a small subset of Low-Rank Adaptation (LoRA) (Hu et al., 2022) parameters $\Delta\theta_{\text{LoRA}}$ are updated while the base parameters θ_{base} remain frozen. We configured LoRA with rank $r = 8$ and scaling factor $\alpha = 16$.

For reasons of exploration and efficient deployment, we did not perform complex dataset construction. Instead, we directly inserted the original advice into a template (detailed in the Appendix) to construct the dataset $\mathcal{D}_{\text{pref}}$. The training continued until the model achieved 100% accuracy on the pseudo-outdated versus up-to-date advice pairs, thereby ensuring complete memorization of the clinical recommendations. The model was then evaluated on independent synthetic scenarios to assess its ability to generalize this memorized knowledge to unseen clinical contexts.

4.3 Hybrid Knowledge Augmentation

To leverage the potential synergy between internalized knowledge from parametric adaptation and dynamic external information, we explored a third strategy. This approach, which we term RAG on DPO (RoD), consists of two main stages, leveraging the same $\mathcal{KB}$ for both DPO training without additional curation effort and RAG retrieval. First, the base LLM is fine-tuned using DPO with LoRA, where only parameters $\Delta\theta_{\text{LoRA}}$ are updated on top of the frozen base parameters θ_{base}, as detailed in our description of Parametric Knowledge Adaptation. Second, during the inference phase with this DPO-tuned model, we utilize the RAG pipeline as previously described (see Non-Parametric Knowledge Augmentation). The DPO-adapted model then generates the response $\hat{y}_s$ based on the original query s augmented with retrieved documents D_k:

$$\hat{y}_s = \text{LLM}\big(s \oplus D_k; (\theta_{\text{base}}, \Delta\theta_{\text{lora}})\big). \tag{8}$$

The RoD strategy thus combines DPO's preference-aligned internal knowledge with RAG's ability to ground responses in external knowledge.

5 Results

5.1 Model Evaluation

We first evaluated the extent to which current LLMs conflict with clinical guidelines using the ConflictMedQA benchmark. Performance was measured using three metrics: endorsement of up-to-date advice (ECDA_{adh}), rejection of outdated advice (ECDA_{rej}), and overall alignment (ECDA_{all}).

All assessed models exhibited varying performance across the five types of clinical recommen-

dation updates (Fig. 2). GPT-4o and Qwen2.5-72B demonstrated the highest ECDA_{adh}, with sample-weighted averages of 0.90 and 0.92, respectively. These scores were significantly higher than the third-best performing model, Qwen2.5-7B (both $p < 0.0001$). However, both models exhibited substantial declines when assessed on their ability to reject pseudo-outdated recommendations, with ECDA_{rej} of 0.395 for GPT-4o and 0.278 for Qwen2.5-72B. Conversely, as in Fig. 2b, Ministral-8B achieved the highest ECDA_{rej} (0.80), followed by gemma-2-27B (0.68) and Llama-3-8B (0.63). When considering overall alignment across both current and outdated scenarios, GPT-4o achieved the highest ECDA_{all} (0.65), as shown in Fig. 2c. This performance was significantly higher than that of the second-best model, Llama-3.3-70B ($\text{ECDA}_{\text{all}} = 0.61$, $p = 0.00033$), and the third-best model, Qwen2.5-72B ($\text{ECDA}_{\text{all}} = 0.60$, $p = 0.0006$).

Beyond difficulties with external guideline alignment, models also exhibited inconsistencies within their internal knowledge. As shown in Fig. 3, all evaluated models exhibited substantial internal conflicts, with considerable variability across models and types of guideline updates. Our analysis revealed that more capable or larger-scale models did not consistently exhibit lower IKCRs. For instance, the 72B parameter version of Qwen2.5 demonstrated higher IKCRs than its 7B counterpart across most evaluated categories. Similarly, Llama-3.3-70B did not show lower conflict ratios compared to Llama-3-8B. Among all models evaluated, the Ministral-8B model achieved the lowest overall IKCR, with a weighted average score of 0.34 across all scenario types, followed by Gemma-2-27B at 0.39.

All evaluated models exhibited knowledge conflicts across all five modification categories. The highest average IKCRs were observed for changes under the groups Implementation Approach and Treatment Modality. While our baseline evaluation distinguished performance across five guideline change categories, the mitigation analysis focuses on overall alignment and conflict rates to emphasize aggregate improvements.

5.2 Mitigation Effectiveness

Fig. 4 shows the qualitative effects of the mitigation approaches, while Table 1 provides a summary of their quantitative performance on the ECDA and IKCR metrics, respectively. These evaluations aim to clarify the effectiveness of each strategy in improving temporal alignment and internal consistency.

Application of RAG and DPO independently improved the models' ECDA_{adh} relative to their baseline performance, as shown in the ECDA_{adh} columns of Table 1. The impact of RAG on the models' ECDA_{rej} was variable across models, as detailed in Table 1. While RAG improved ECDA_{rej} for some models, it decreased ECDA_{rej} for Ministral-8B and Llama-3-8B compared to their respective baselines.

When considering overall alignment (ECDA_{all}), as presented in Table 1, both RAG and DPO individually improved performance. However, RoD consistently yielded the highest ECDA_{all} scores across all models where this combination was tested. This improvement from the RoD approach was consistently greater than the best-performing single method (RAG or DPO alone) for each model.

Analysis of the IKCR, detailed in Table 1, showed that DPO alone generally reduced IKCR across all evaluated models compared to their baselines. RAG alone reduced internal contradictions for most models compared to their baseline. However, for Ministral-8B and Llama-3-8B, applying RAG alone increased IKCR. Notably, RoD resulted in the lowest IKCR for all models where this combination was tested, including Ministral-8B and Llama-3-8B, surpassing the reductions achieved by DPO or RAG alone.

6 Discussion

Our evaluation on the ConflictMedQA benchmark reveals significant challenges for LLMs in clinical decision-making, primarily their struggle with the temporal dynamics of medical knowledge and internal consistency. Even advanced models, adept at endorsing current guidelines, often faltered markedly when required to reject outdated advice. This asymmetry, coupled with the finding that larger model scale does not consistently reduce internal knowledge conflicts, suggests that unique complexities arise in this domain beyond standard NLP capabilities. These issues, especially prevalent in areas like therapeutic recommendations, could pose direct risks if LLMs are integrated into clinical workflows without a deep understanding of their failure modes.

Investigating mitigation strategies offered further insights. While RAG generally improved ad-

Table 1: Performance of LLMs on ECDA and IKCR. Results are shown as final scores, with absolute improvements over the base model in parentheses. Higher ECDA is better, while lower IKCR is better.

Model	$ECDA_{adh}$				$ECDA_{rej}$			
	Base	RAG	DPO	RoD	Base	RAG	DPO	RoD
Qwen2.5-72B	91	**98** (+07)	–	–	28	27 (-01)	–	–
Llama-3.3-70B	66	96 (+30)	–	–	56	71 (+15)	–	–
gemma-2-27B	48	82 (+34)	–	–	68	70 (+02)	–	–
GPT-4o	90	96 (+06)	–	–	40	65 (+25)	–	–
Qwen2.5-7B	74	94 (+20)	81 (+07)	88 (+14)	35	50 (+15)	55 (+20)	74 (+39)
Llama-3-8B	48	93 (+45)	81 (+33)	88 (+40)	63	30 (-33)	55 (-08)	74 (+11)
Ministral-8B	30	87 (+57)	81 (+51)	87 (+57)	80	61 (-19)	85 (+05)	**90** (+10)

Model	$ECDA_{all}$				IKCR			
	Base	RAG	DPO	RoD	Base	RAG	DPO	RoD
Qwen2.5-72B	59	62 (+02)	–	–	73	71 (-02)	–	–
Llama-3.3-70B	61	83 (+22)	–	–	45	29 (-16)	–	–
gemma-2-27B	58	76 (+18)	–	–	39	31 (-08)	–	–
GPT-4o	65	81 (+16)	–	–	61	35 (-26)	–	–
Qwen2.5-7B	55	72 (+17)	68 (+13)	81 (+26)	65	51 (-14)	43 (-22)	26 (-39)
Llama-3-8B	55	62 (+07)	68 (+13)	81 (+26)	45	70 (+25)	43 (-02)	26 (-19)
Ministral-8B	55	74 (+19)	83 (+28)	**89** (+34)	34	40 (+06)	15 (-19)	**10** (-24)

herence to current information, its utility was nuanced. Notably, for smaller models, RAG alone could paradoxically degrade their ability to reject outdated advice, suggesting that merely providing external information can be counterproductive if the model lacks the capacity to critically discern and integrate it, potentially overwhelming weaker internal knowledge structures. This indicates that effective retrieval is as much about the model's ability to use information as it is about accessing it.

DPO offered a simple complementary approach, demonstrably enhancing alignment with current guidelines and reducing internal conflicts. However, these improvements in complex clinical scenarios stood in contrast to the near-perfect performance models presumably achieve on the specific raw medical advice pairs used during DPO training. This discrepancy suggests a significant challenge in generalizing knowledge learned from such simple pairs to the multifaceted reasoning required in clinical practice, hinting at a gap between memorized correct responses and their robust, contextual application.

The most promising path appears to be the synergistic combination of these approaches. Our findings show that RoD, applying RAG to DPO-tuned models, yielded substantial improvements across all metrics, particularly in enhancing smaller models' rejection of outdated advice and minimizing internal conflicts. These gains always exceed the sum of the individual contributions from RAG-only or DPO-only applications. While models may struggle to effectively apply DPO-learned parametric knowledge across diverse and complex scenarios, the integration with RAG appears pivotal. Knowledge retrieved via RAG seems to activate relevant DPO-instilled parametric knowledge within the model, leading to these markedly enhanced outcomes and avoiding the potential side effects of RAG-only or the more modest improvements from DPO-only strategies.

These observations also underscore a significant limitation of evaluating LLMs using metrics focused on isolated factual accuracy. The marked performance decline when models face realistic clinical scenarios, which embed cognitive complexities and factors like SDoH, emphasizes the strong need for evaluation methodologies that capture the multifaceted nature of clinical decision-making.

7 Conclusion

Ultimately, for the safe and effective integration of LLMs into clinical practice, future efforts should prioritize the development of robust, hybrid

methodologies designed to enhance adaptability to evolving knowledge and ensure internal consistency. This entails creating more contextually rich training and evaluation paradigms that mirror the complexity of real-world clinical encounters, thereby moving beyond isolated assessments to foster genuine contextual understanding and reliability in these critical systems.

Limitations

While we only explored two mitigation strategies that are relatively straightforward to implement, do not require elaborate dataset curation, and have reasonable computational costs, our results demonstrate their potential to improve temporal consistency with current clinical guidelines. Due to a lack of access to proprietary model weights and limited computational resources, we could not apply DPO universally across all assessed models. Additionally, our evaluation was limited to synthetic clinical scenarios that may not fully capture the complexity and diversity of clinical practice. Future work should consider using real-world cases abstracted from healthcare workers with varying levels of complexity, common typographical errors, and incomplete information to better test models' adaptability and generalization capabilities in realistic medical settings.

Ethical Considerations

We have not identified any ethical concerns directly related to this study.

Acknowledgment

This study is supported by the Department of Defense grant HT9425-23-1-0267.

References

Salim S Abdool Karim and Nikita Devnarain. 2022. Time to stop using ineffective covid-19 drugs.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Angelo Cagnacci and Martina Venier. 2019. The controversial history of hormone replacement therapy. *Medicina*, 55(9):602.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hitesh Chopra, Dong K Shin, Kavita Munjal, Kuldeep Dhama, Talha B Emran, et al. 2023. Revolutionizing clinical trials: the role of ai in accelerating medical breakthroughs. *International Journal of Surgery*, 109(12):4211–4220.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

American Diabetes Association Professional Practice Committee. 2025. 9. pharmacologic approaches to glycemic treatment: Standards of care in diabetes—2025. *Diabetes Care*, 48(Supplement_1):S181–S206.

Peter Densen. 2011. Challenges and opportunities facing medical education. *Transactions of the American clinical and climatological association*.

Benjamin S Glicksberg, Prem Timsina, Dhaval Patel, Ashwin Sawant, Akhil Vaid, Ganesh Raut, Alexander W Charney, Donald Apakama, Brendan G Carr, Robert Freeman, et al. 2024. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. *Journal of the American Medical Informatics Association*, 31(9):1921–1928.

Alexandra Halalau, Brett Holmes, Andrea Rogers-Snyr, Teodora Donisan, Eric Nielsen, Tiago Lemos Cerqueira, and Gordon Guyatt. 2021. Evidence-based medicine curricula and barriers for physicians in training: a scoping review. *International journal of medical education*, 12:101.

M Elizabeth H Hammond, Josef Stehlik, Stavros G Drakos, and Abdallah G Kfoury. 2021. Bias in medicine: lessons learned and mitigation strategies. *Basic to Translational Science*, 6(1):78–85.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Nice-Sugar Study Investigators. 2009. Intensive versus conventional glucose control in critically ill patients. *New England Journal of Medicine*, 360(13):1283–1297.

Shio-Shin Jean and Po-Ren Hsueh. 2020. Old and repurposed drugs for the treatment of covid-19. *Expert review of anti-infective therapy*, 18(9):843–847.

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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Susanne P Lajoie and Maren Gube. 2018. Adaptive expertise in medical education: accelerating learning trajectories by fostering self-regulated learning. *Medical Teacher*, 40(8):809–812.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

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. 2024. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics.

Guofang Ma, Miranda G Scully, Jiahui Luo, Jiazuo H Feng, Christine M Gunn, Roberta M DiFlorio Alexander, Anna NA Tosteson, Sally A Kraft, and Wesley Marrero. 2025. Modeling the impact of social determinants on breast cancer screening: A data-driven approach. *Frontiers in Medicine*, 12:1644287.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 7(1):295.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Paul G Shekelle, Eduardo Ortiz, Shannon Rhodes, Sally C Morton, Martin P Eccles, Jeremy M Grimshaw, and Steven H Woolf. 2001. Validity of the agency for healthcare research and quality clinical practice guidelines: how quickly do guidelines become outdated? *Jama*, 286(12):1461–1467.

Natesh Singh, Philippe Vayer, Shivalika Tanwar, Jean-Luc Poyet, Katya Tsaioun, and Bruno O Villoutreix. 2023. Drug discovery and development: introduction to the general public and patient groups. *Frontiers in Drug Discovery*, 3:1201419.

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiangchi Yuan, Chunhui Zhang, Zheyuan Liu, Dachuan Shi, Soroush Vosoughi, and Wenke Lee. 2025. Superficial self-improved reasoners benefit from model merging. *arXiv preprint arXiv:2503.02103*.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

A Additional Results

A.1 Domain-Specific Models

We further evaluated domain-specific medical models including Med42-8B and Med42-70B (Christophe et al., 2024) as well as OpenBioLLM-70B (Ankit Pal, 2024). Their performance is summarized in Table 2.

A.2 LoRA Ablation Studies

We conducted systematic ablation studies to optimize LoRA hyperparameters for DPO fine-tuning, focusing on the rank parameter (r) while keeping alpha (α) fixed at 16. The results are shown in Table 3.

Overall, higher rank values consistently improve performance given the same training data. Rank 16 achieves the best balance between parameter efficiency and knowledge embedding effectiveness, and the trend suggests that larger ranks enable more effective parametric knowledge injection.

A.3 Factors Impact Analysis

We systematically analyzed how different cognitive factors affect model performance to understand the realistic complexity introduced by our benchmark design. The results are presented in Table 4.

Models generally achieve higher ECDA scores under the "No Factor" condition, validating our benchmark design. We do not observe systematic bias toward incorrect recommendations despite factor inclusion, indicating that the factors simulate realistic clinical complexity without compromising evaluation validity.

A.4 Comprehensive Performance Visualization

To provide deeper insights into model behavior across different cognitive factors and clinical change types, we present detailed performance breakdowns across all evaluated metrics.

A.4.1 Performance by Cognitive Factor

As shown in Figures 9–12, the "No Factor" condition consistently yields the best performance across ECDA metrics, aligning with Table 4. Different clinical change types pose varying challenges; in particular, Implementation Approach and Treatment Modality tend to exhibit higher IKCR (cf. Figure 12), indicating greater internal tension for these settings. Larger models do not uniformly outperform smaller ones on rejection ($ECDA_{rej}$; Figure 10), consistent with our hypothesis regarding pre-training bias amplification. Finally, trends in $ECDA_{adh}$ (Figure 9) and $ECDA_{rej}$ (Figure 10) mirror the aggregate $ECDA_{all}$ behavior (Figure 11), supporting the robustness of our evaluation framework.

The tables below detail mitigation effects across different models, clinical factors, and advice change types. Overall mitigation strategy comparisons are presented in Table 5, while Table 6 reports results specific to the confirmation factor.

A.5 Recommendation Intensity Category: Clinical Justification

Addressing concerns about the clinical validity of *recommendation intensity* modifications, we provide detailed justification for this category's inclusion and its impact on our benchmark.

Clinical Significance of Intensity Variations. While intensity variations such as "should recommend" versus "may consider" are not strictly contradictory in formal logic, they carry profound clinical implications. First, clinical studies demonstrate that "should" language typically results in adherence rates of approximately 80%, compared to only 20% when phrased as "may consider." Second, many real-world clinical guideline updates explicitly focus on the strength of recommendation rather than altering the core intervention. Finally, practice variation studies show that intensity changes directly influence clinical decision-making patterns and patient outcomes.

Example Analysis. For instance, a current recommendation such as "People without immunity should receive full vaccination" differs substantially in clinical impact from a modified version: "People without immunity may consider receiving full vaccination." This shift constitutes a meaningful clinical conflict that affects patient outcomes and public health recommendations, and accounts for 27.2% of our dataset scenarios.

A.6 External vs. Internal Conflict Framework

To clarify our conflict detection methodology, we distinguish between external and internal conflicts.

External Conflicts. Each recommendation pair (R_{current}, R_{outdated}) generates scenarios S_{current} and S_{outdated}, which are evaluated independently against current medical ground truth. An external conflict occurs when the model endorses S_{outdated} (which should be rejected) or rejects S_{current} (which should be endorsed).

Table 2: Results Summary: Domain-Specific Models

Model	IKCR	ECDA_all	ECDA_adh	ECDA_rej
Med42-8B (Base)	0.4626	0.5529	0.5497	0.5562
Med42-8B (RAG)	0.4990	0.6914	0.8643	0.5184
Med42-8B (DPO)	0.2056	0.7406	0.6601	0.8210
Med42-8B (RoD)	0.1246	0.8380	0.7902	0.8858
Med42-70B (Base)	0.7055	0.5762	0.8713	0.2811
Med42-70B (RAG)	0.4000	0.7730	0.9566	0.5893
OpenBioLLM-70B (Base)	0.5963	0.5560	0.7298	0.3831
OpenBioLLM-70B (RAG)	0.6053	0.6545	0.9273	0.3818

Table 3: Ablation Results (Mistral-8B)

Rank (r)	IKCR	ECDA_all	ECDA_adh	ECDA_rej
4	0.2639	0.7524	0.7515	0.7534
8	0.1504	0.8331	0.8145	0.8517
16	**0.1181**	**0.8417**	**0.7986**	**0.8848**

Table 4: Factor-wise Performance Analysis (LLaMA-8B)

Factor Type	IKCR	ECDA_all	ECDA_adh	ECDA_rej
Self-Diagnosis	0.4333	0.5462	0.4872	0.6051
Recency Factor	0.4579	0.5462	0.4462	0.6462
Confirmation	0.4833	0.5718	0.5282	0.6154
Frequency	0.4655	0.5308	0.4667	0.5949
Cultural Factor	0.4959	0.5846	0.5487	0.6205
Status Quo	0.4000	0.5308	0.4256	0.6359
False Consensus	0.4273	0.5385	0.4410	0.6359
Racial/Ethnic	0.4915	0.5564	0.5077	0.6051
Socioeconomic	0.4364	0.5256	0.4308	0.6205
Geographic	0.4690	0.5615	0.4872	0.6359
No Factor	**0.4444**	**0.6000**	**0.5333**	**0.6667**

Internal Conflicts. These are assessed using paired scenarios where simultaneous endorsement indicates internal knowledge inconsistency. For a given scenario pair ($S_{i,\text{current}}$, $S_{i,\text{outdated}}$) for concept i, an internal conflict arises when the model endorses both scenarios. The Internal Knowledge Conflict Rate (IKCR) quantifies the frequency of such contradictions across all active pairs, as reported in Table 2.

A.7 Analysis of Counterintuitive Scale Effects

Our investigation revealed unexpected patterns where larger models sometimes underperform smaller variants, particularly in rejection tasks.

Empirical Evidence. Table 7 shows representative results across three model families, highlighting that parameter scaling does not guarantee improved performance on ECDA_rej.

Proposed Mechanistic Explanation. We hypothesize this phenomenon results from *pre-training bias amplification*. Clinical scenarios rich in specialized terminology may trigger strong correctness associations learned during pre-training (the *authority signal hypothesis*). Larger models, exposed to broader corpora, develop stronger heuristic associations between clinical language and authoritative content, leading to scale-dependent bias. These pre-training biases can override rejection capabilities acquired during RLHF or instruction tuning, especially when plausible but incorrect recommendations are presented. By contrast, smaller models may be less affected due to

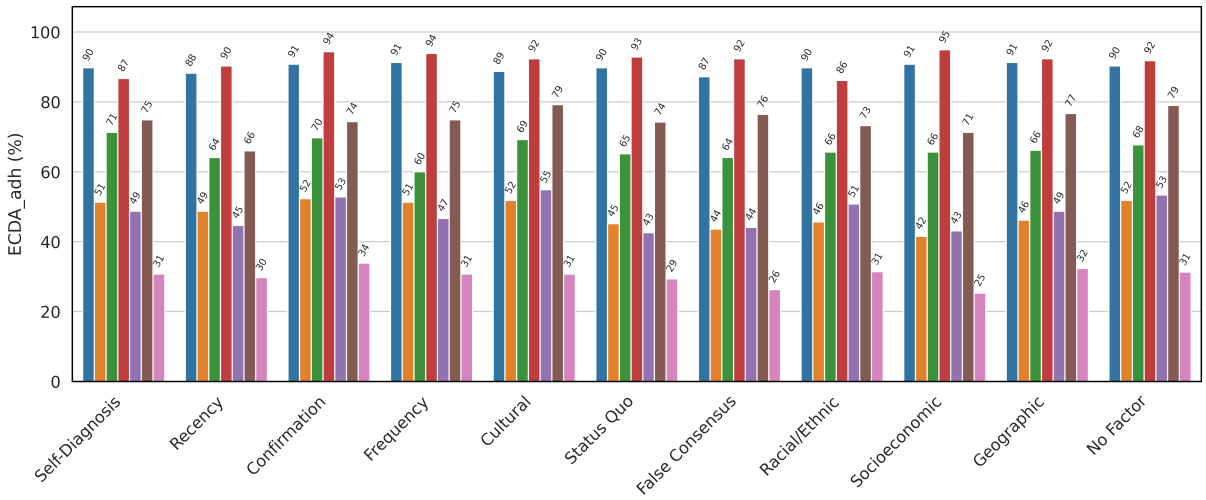


Figure 5: $ECDA_{adh}$ performance across clinical factors. This metric measures models' ability to correctly endorse up-to-date medical recommendations under different cognitive biases.

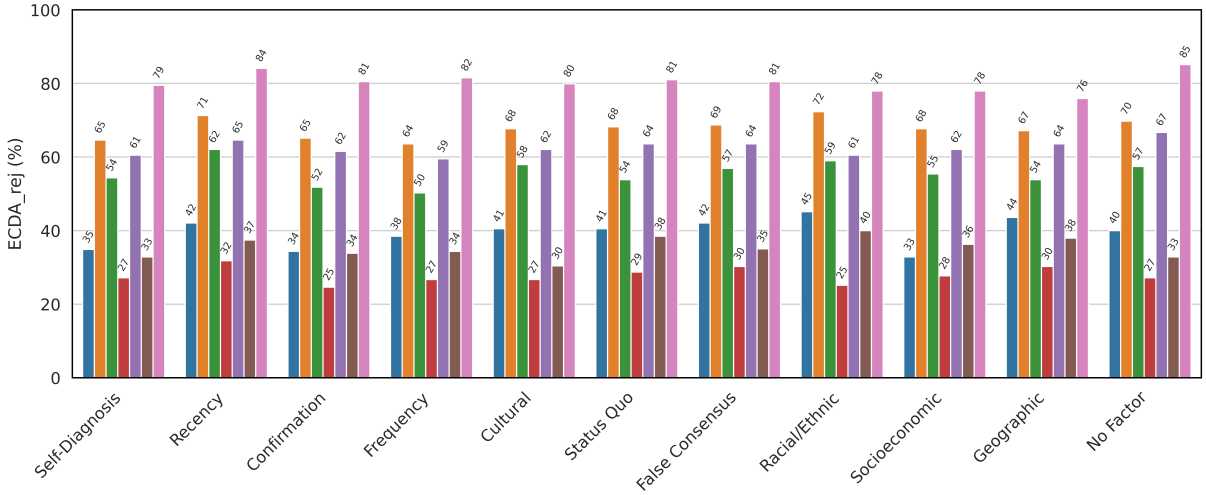


Figure 6: $ECDA_{rej}$ performance across clinical factors. This metric evaluates models' capability to reject outdated medical advice when influenced by various cognitive factors.

weaker initial biases and a proportionally greater influence of alignment training updates. This observation emphasizes that medical LLM evaluation requires careful consideration of both capability scaling and bias amplification effects.

B Detailed Description of LLMs

Below we provide a brief description of each large language model (LLM) evaluated in our study, highlighting their key architectural and training characteristics.

GPT-4o is OpenAI's multimodal model. While the exact parameter count remains undisclosed, GPT-4o features a unified architecture capable of processing and generating text, images, and audio with a context window of up to 128,000 tokens. It achieves comparable or better text performance relative to GPT-4, but with significantly lower latency and cost. The model is instruction-tuned and optimized for real-time interactive applications. We used GPT-4 via the OpenAI API under its terms of use.

Llama-3-8B and **Llama-3-70B** are Meta's latest open-weight models, featuring 8 billion and 70 billion parameters, respectively. Both are dense decoder-only Transformers trained on approximately 15 trillion tokens of deduplicated public data. Instruction-tuned versions incorporate multi-stage reinforcement learning from human feedback (RLHF), and Meta provides both default (8K) and long-context (up to 128K) variants for research.

Qwen2.5-7B and **Qwen2.5-72B** are Alibaba's

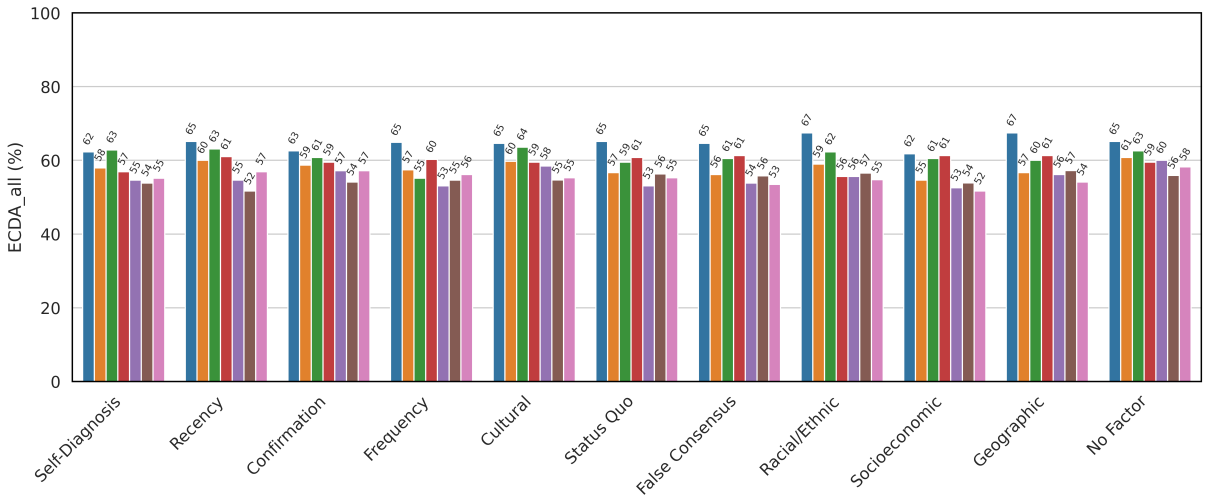


Figure 7: Overall ECDA performance ($ECDA_{all}$) across clinical factors, representing the balanced assessment of both endorsement and rejection capabilities.

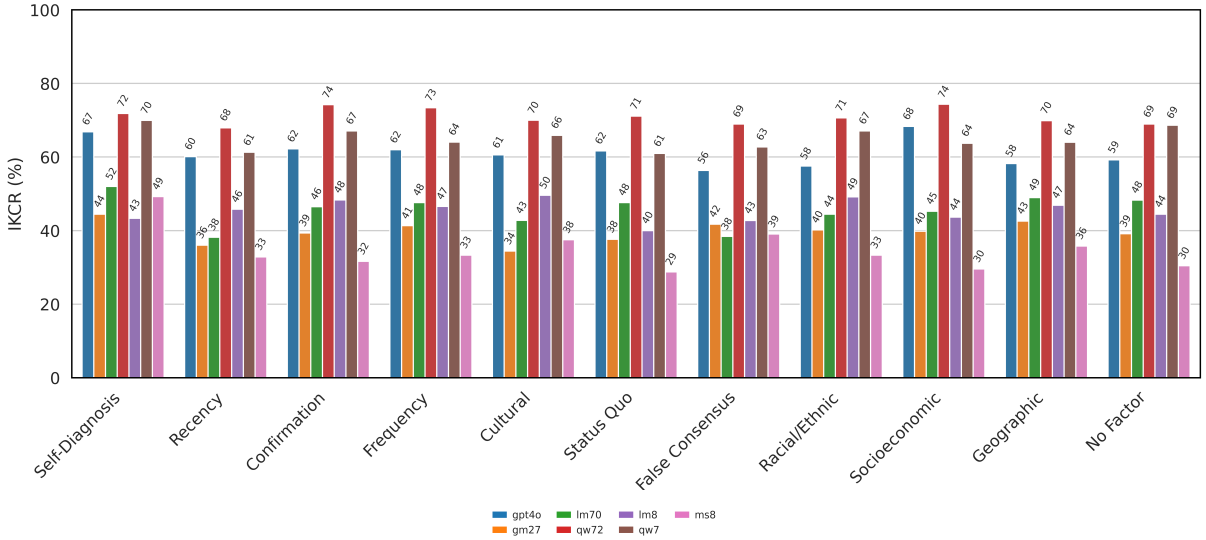


Figure 8: Internal Knowledge Conflict Ratio (IKCR) across clinical factors. Lower values indicate better internal consistency, with "No Factor" serving as the baseline condition. The legend below provides symbol/color references.

state-of-the-art models with 7 billion and 72 billion parameters. Qwen 2.5 introduces a greatly expanded pre-training corpus (18T tokens) and large-scale supervised fine-tuning (over 1 million samples), along with reinforcement learning and reward modeling. Both models natively support a 32,000-token context window.

Gemma-2-27B-it is Google DeepMind's 27-billion-parameter, instruction-tuned model from the Gemma 2 family. It employs dense Transformer architecture with interleaved local-global attention and group-query attention to improve memory efficiency. Gemma-2 models are trained on up to 8T tokens and are designed for efficient inference on single high-memory GPUs or TPUs, released under the Apache 2.0 license.

Ministral-8B-Instruct-2410 is a recently released model from Mistral AI, designed for local and on-device use. It features 8 billion parameters with a dense Transformer architecture and a context window of up to 128,000 tokens, enabled by interleaved sliding-window attention. Ministral-8B-Instruct

C More details in Dataset Construction

We derived these pseudo-outdated recommendations using one of five strategies designed to reflect common patterns of knowledge evolution in clinical guideline updates:

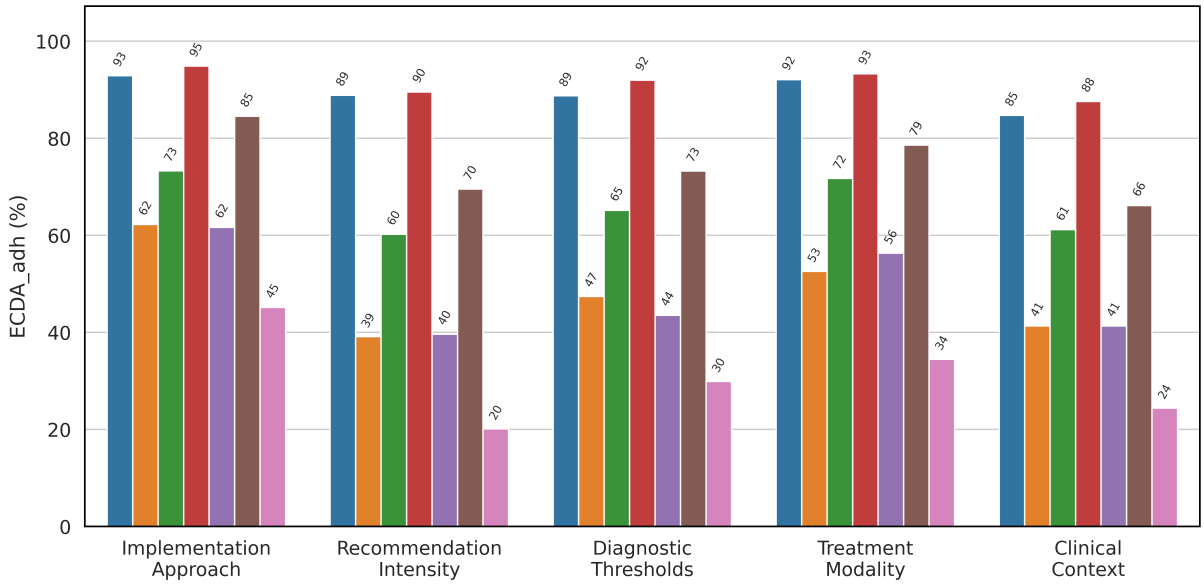


Figure 9: $ECDA_{adh}$ performance across clinical change types. This metric measures models' ability to correctly endorse up-to-date medical recommendations under different cognitive biases.

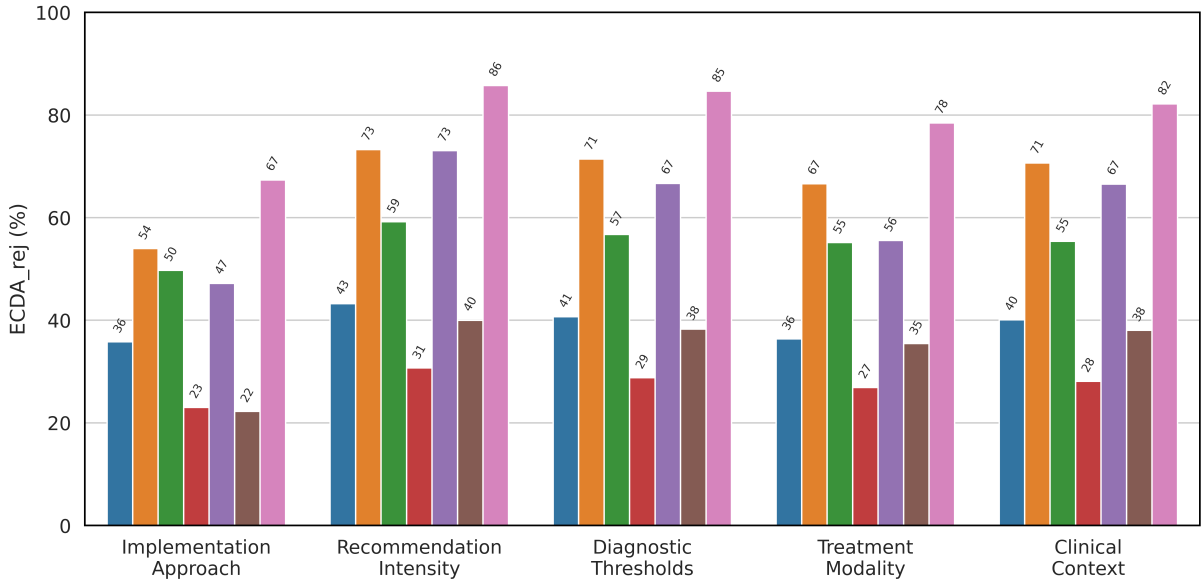


Figure 10: $ECDA_{rej}$ performance across clinical change types. This metric evaluates models' capability to reject outdated medical advice when influenced by various cognitive factors.

- **Clinical Context** (N=22, 11.3%): Revisions to the specific patient populations or clinical circumstances to which a recommendation applies (e.g., narrowing or broadening age ranges, changing applicability based on risk status).
  Examples: revising age applicability from "adults aged <60 years" to "adults aged <70 years"; narrowing recommendation from "all patients" to "only high-risk patients".

- **Diagnostic & Threshold** (N=42, 21.5%): Modifications to specific numerical criteria or classifications used in diagnosis or risk stratification (e.g., changing diagnostic thresholds for blood glucose or HbA1c, altering risk score cutoffs).
  Examples: changing the fasting glucose diagnostic threshold from "100–110 mg/dL" to "110–125 mg/dL"; adjusting HbA1c criteria from "≥6.5%" to "≥7.0%".

- **Implementation Approach** (N=32, 16.4%): Changes in how care is delivered, organized, or monitored, including methods, processes, systems, duration, or frameworks, even if the core

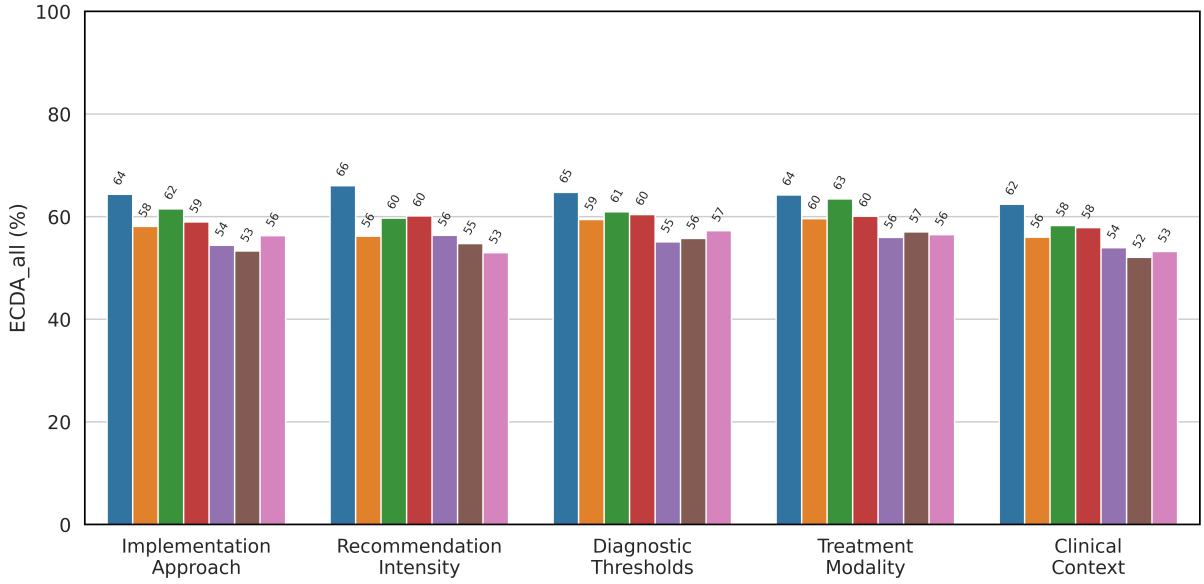


Figure 11: Overall ECDA performance ($ECDA_{all}$) across clinical change types, representing the balanced assessment of both endorsement and rejection capabilities.

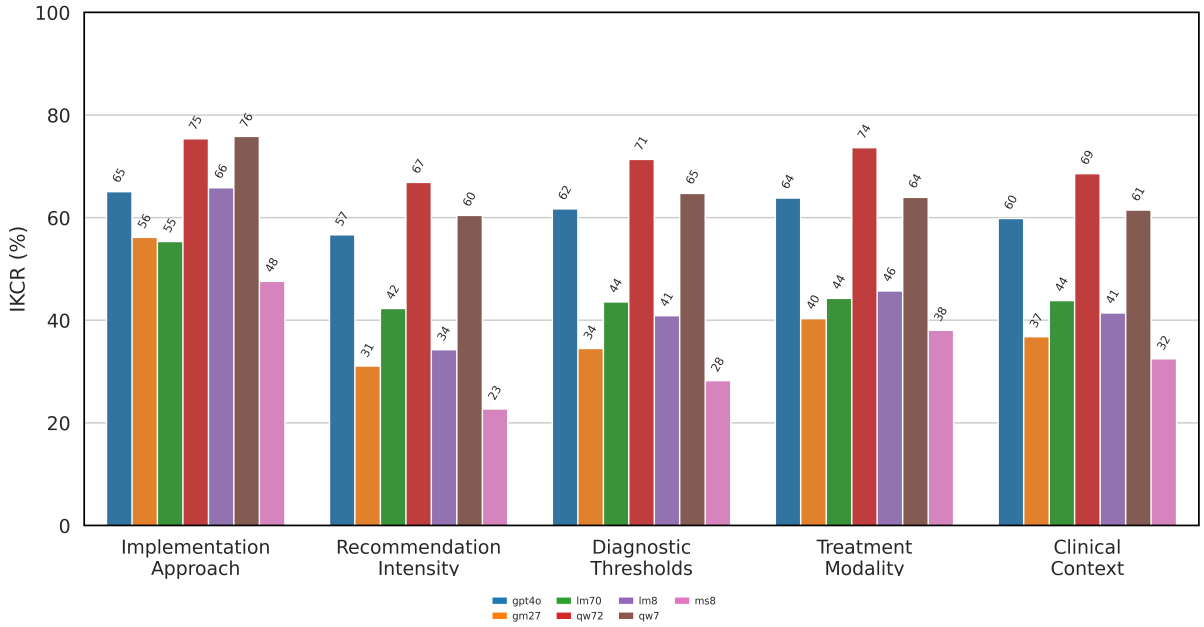


Figure 12: Internal Knowledge Conflict Ratio (IKCR) across clinical change types. Lower values indicate better internal consistency, with "No Factor" serving as the baseline condition. The legend below provides symbol/color references.

treatment or diagnosis remains similar.
Examples: shifting from "moderate complexity" to "low complexity" management; transitioning from "lifelong monitoring" to a "short-term surveillance".

- **Recommendation Intensity** (N=53, 27.2%): Changes in the strength or certainty of a recommendation while the core action remains the same (e.g., shifting from permissive to directive language, or vice versa).
Examples: changing recommendation wording from "may consider" to "should recommend"; from "not recommended" to "recommended" for the same action.

- **Treatment Modality** (N=46, 24.6%): Changes in the specific medical interventions recom-

Table 5: Mitigation Strategy Performance Comparison (Overall)

Model	Strategy	IKCR	ECDA_all	ECDA_adh	ECDA_rej
Gemma-2-27B	Base	0.397	0.580	0.481	0.678
	RAG	0.308 (-0.089)	0.759 (+0.179)	0.821 (+0.340)	0.697 (+0.018)
GPT-4o	Base	0.612	0.646	0.898	0.395
	RAG	0.352 (-0.260)	0.806 (+0.160)	0.965 (+0.067)	0.648 (+0.253)
LLaMA-3.3-70B	Base	0.455	0.610	0.662	0.557
	RAG	0.297 (-0.158)	0.829 (+0.220)	0.958 (+0.295)	0.701 (+0.144)
LLaMA-3-8B	Base	0.456	0.554	0.482	0.626
	RAG	0.699 (+0.243)	0.620 (+0.067)	0.935 (+0.453)	0.305 (-0.320)
	DPO	0.294 (-0.161)	0.747 (+0.194)	0.772 (+0.290)	0.723 (+0.097)
	RoD	0.238 (-0.217)	0.847 (+0.293)	0.934 (+0.452)	0.760 (+0.135)
Mistral-8B	Base	0.345	0.553	0.302	0.804
	RAG	0.398 (+0.053)	0.744 (+0.191)	0.876 (+0.574)	0.613 (-0.191)
	DPO	0.150 (-0.195)	0.833 (+0.280)	0.815 (+0.513)	0.852 (+0.048)
	RoD	0.104 (-0.242)	0.888 (+0.335)	0.879 (+0.578)	0.897 (+0.094)
Qwen2.5-7B	Base	0.651	0.550	0.745	0.354
	RAG	0.509 (-0.142)	0.720 (+0.170)	0.935 (+0.190)	0.504 (+0.150)
	DPO	0.437 (-0.214)	0.680 (+0.130)	0.814 (+0.068)	0.546 (+0.192)
	RoD	0.263 (-0.388)	0.807 (+0.257)	0.880 (+0.135)	0.734 (+0.380)
Qwen2.5-72B	Base	0.710	0.597	0.916	0.278
	RAG	0.729 (+0.019)	0.624 (+0.027)	0.982 (+0.066)	0.266 (-0.012)

Table 6: Mitigation Performance for Confirmation Factor

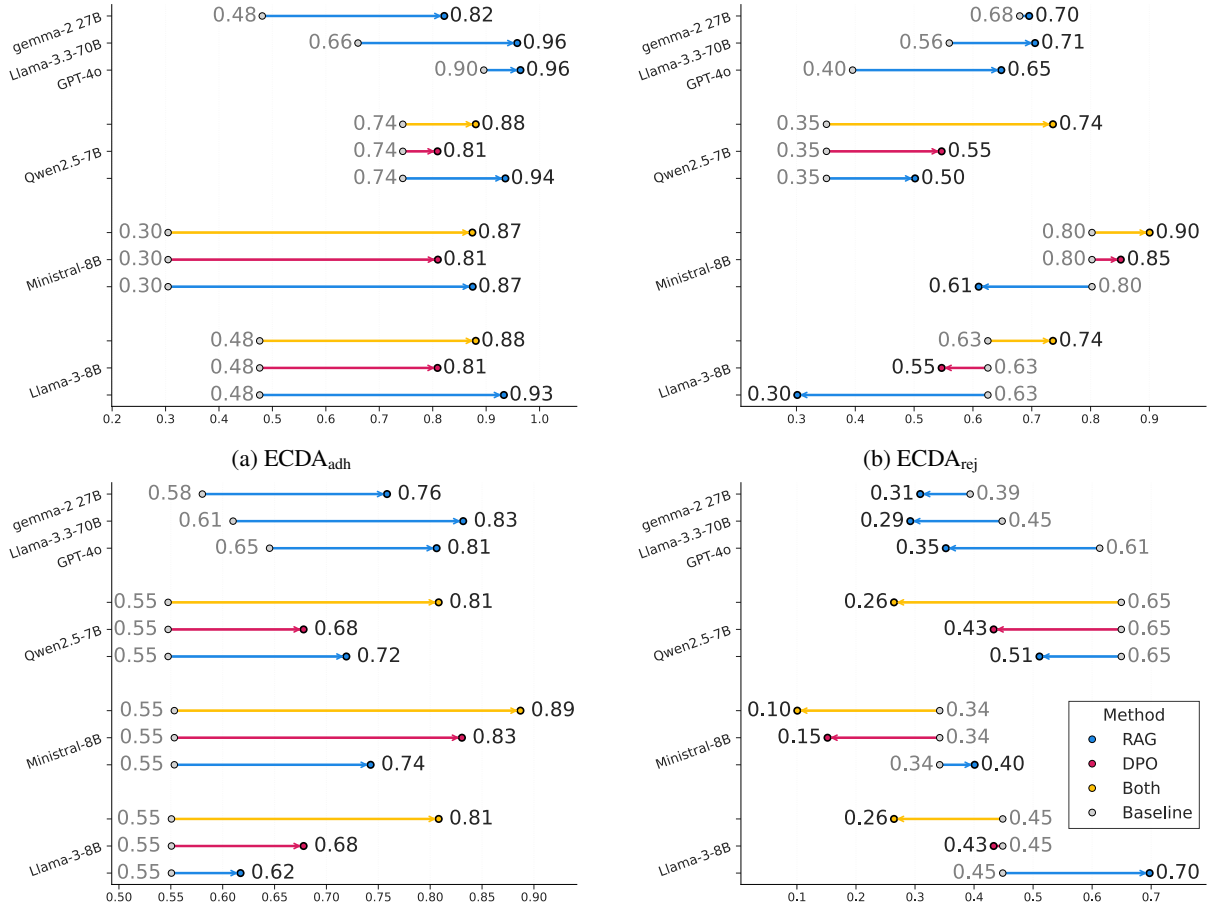
Model	Strategy	IKCR	ECDA_all	ECDA_adh	ECDA_rej
Gemma-2-27B	Base	0.393	0.587	0.523	0.651
	RAG	0.302 (-0.091)	0.756 (+0.169)	0.831 (+0.308)	0.682 (+0.031)
GPT-4o	Base	0.622	0.626	0.908	0.344
	RAG	0.354 (-0.268)	0.818 (+0.192)	0.995 (+0.087)	0.641 (+0.297)
LLaMA-3.3-70B	Base	0.465	0.608	0.697	0.518
	RAG	0.295 (-0.170)	0.830 (+0.223)	0.964 (+0.266)	0.697 (+0.179)
LLaMA-3-8B	Base	0.483	0.572	0.528	0.615
	RAG	0.696 (+0.213)	0.623 (+0.051)	0.954 (+0.426)	0.292 (-0.323)
	DPO	0.323 (-0.161)	0.754 (+0.182)	0.790 (+0.261)	0.718 (+0.103)
	RoD	0.242 (-0.241)	0.846 (+0.274)	0.939 (+0.410)	0.754 (+0.138)
Mistral-8B	Base	0.317	0.572	0.339	0.805
	RAG	0.397 (+0.080)	0.756 (+0.185)	0.897 (+0.559)	0.615 (-0.190)
	DPO	0.153 (-0.163)	0.839 (+0.267)	0.821 (+0.482)	0.856 (+0.051)
	RoD	0.099 (-0.218)	0.892 (+0.321)	0.877 (+0.538)	0.908 (+0.103)
Qwen2.5-7B	Base	0.671	0.541	0.744	0.339
	RAG	0.535 (-0.136)	0.715 (+0.174)	0.944 (+0.200)	0.487 (+0.149)
	DPO	0.456 (-0.215)	0.659 (+0.118)	0.790 (+0.046)	0.528 (+0.190)
	RoD	0.263 (-0.408)	0.821 (+0.279)	0.887 (+0.144)	0.754 (+0.415)
Qwen2.5-72B	Base	0.742	0.595	0.944	0.246
	RAG	0.732 (-0.010)	0.633 (+0.038)	0.995 (+0.051)	0.272 (+0.026)

Table 7: Scale Effects on ECDA_rej Performance

Model Family	Parameter Size	ECDA_rej Performance
Qwen	7B → 72B	0.3540 → 0.2783
LLaMA	8B → 70B	0.6256 → 0.5571
Med42	8B → 70B	0.5562 → 0.2811

mended (e.g., replacing an older drug class with a newer one, shifting from surgical to non-surgical approaches).
Examples: replacing "metformin" with "GLP-1 receptor agonists"; transitioning from "surgical intervention" to "physical therapy".

D Prompts & Templates



(a) ECDA$_{adh}$ (b) ECDA$_{rej}$

(c) ECDA$_{all}$ (d) IKCR

Figure 13: Effect of mitigation strategies on model alignment and internal consistency. Each line originates from the baseline performance of a given model and shows changes following the application of RAG (blue), DPO (red), or their combination (yellow). Rightward shifts indicate improvement, while leftward shifts reflect performance degradation. Metrics include endorsement of current advice, rejection of outdated advice, overall alignment, and internal knowledge conflict ratio.

Bias Type	Guidance for Natural Evidence
Self-Diagnosis Bias	**Natural Evidence Examples:** Instead of saying "because of self-diagnosis bias," use natural phrases showing clinicians influenced by patient's self-diagnosis: • "giving significant weight to the patient's detailed symptom journal and self-diagnosis" • "considering the patient's own conclusion about their condition based on their research" • "influenced by the patient's confident self-diagnosis and symptom tracking" **Example Format:** "[CLINICIAN] [FOLLOWS ADVICE] after [BEING INFLUENCED BY PATIENT'S SELF-DIAGNOSIS]."
Recency Bias	**Natural Evidence Examples:** Instead of saying "because of recency bias," use natural phrases showing recent experience influence: • "having diagnosed some similar cases just this month" • "strongly influenced by a recent patient who had complications" • "after seeing an unusual cluster of similar symptoms in recent patients" **Example Format:** "[CLINICIAN] [FOLLOWS ADVICE] after [RECENT EXPERIENCE INFLUENCING DECISION]."
Confirmation Bias	**Natural Evidence Examples:** Instead of saying "because of confirmation bias," use natural phrases showing selective evidence use: • "focusing primarily on test results that support her initial hypothesis" • "interpreting ambiguous symptoms in a way that confirms her suspected diagnosis" • "recalling similar cases that had the outcome he expected" **Example Format:** "[CLINICIAN] [FOLLOWS ADVICE] based on [SELECTIVELY USING CONFIRMING EVIDENCE]."
Frequency Bias	**Natural Evidence Examples:** Instead of saying "because of frequency bias," use natural phrases showing preference for common diagnoses: • "defaulting to the most frequent explanation despite ambiguous evidence" • "choosing the diagnosis she encounters most frequently for these symptoms" • "favoring the most frequently condition despite some atypical features" **Example Format:** "[CLINICIAN] [FOLLOWS ADVICE] by [FAVORING THE MORE COMMON DIAGNOSIS/TREATMENT]."

Table 8: Bias Types and Natural Evidence Guidance for Medical Scenarios (Part 1)

Bias Type	Guidance for Natural Evidence
Cultural Bias	**Natural Evidence Examples:** Instead of saying "because of cultural bias," use natural phrases showing cultural lens interpretation: • "interpreting the symptoms through the lens of his own cultural understanding of health" • "recommending treatments aligned with cultural practices familiar to her" • "approaching the diagnosis from her cultural framework of disease causation" **Example Format:** "[CLINICIAN] [FOLLOWS ADVICE] while [INTERPRETING THROUGH CULTURAL LENS]."
Status Quo Bias	**Natural Evidence Examples:** Instead of saying "because of status quo bias," use natural phrases showing preference for PREVIOUSLY USED treatments: • "recommending the approach she's used throughout her career" • "sticking with the familiar treatment regimen" • "choosing the conventional therapy that has been the standard for decades" **Example Format:** "[CLINICIAN] [FOLLOWS ADVICE] by [PREFERRING PREVIOUSLY USED TREATMENTS]."
False Consensus Bias	**Natural Evidence Examples:** Instead of saying "because of false consensus bias," use natural phrases showing overestimation of agreement: • "assuming all other specialists would make the same recommendation without verification" • "overstating how widely accepted this approach is among colleagues" • "believing his view represents the medical consensus without checking current opinions" **Example Format:** "[CLINICIAN] [FOLLOWS ADVICE], [OVERESTIMATING AGREEMENT AMONG PEERS]."
Racial/Ethnic Bias	**Natural Evidence Examples:** Instead of saying "because of racial/ethnic bias," use natural phrases like: • "citing studies that primarily focused on specific ethnic populations" • "based on how the condition typically presents in patients of this background" • "believing genetic factors common in this heritage influence treatment response" **Example Format:** "[PERSON] [FOLLOWS ADVICE] based on [ETHNIC-SPECIFIC REASONING]."

Table 9: Bias Types and Natural Evidence Guidance for Medical Scenarios (Part 2)

Bias Type	Guidance for Natural Evidence
Socioeconomic Bias	**Natural Evidence Examples:** Instead of saying "because of socioeconomic bias," use natural phrases like: • "considering the cost-effectiveness of this approach" • "noting that patients with limited insurance often do better with this option" • "recognizing the patient's financial constraints in managing treatment" **Example Format:** "[PERSON] [FOLLOWS ADVICE] after [FINANCIAL/RESOURCE CONSIDERATION]."
Geographic Bias	**Natural Evidence Examples:** Instead of saying "because of geographic bias," use natural phrases like: • "following the standard practice in their region" • "based on protocols common in rural healthcare settings" • "as is typically done in hospitals in this part of the country" **Example Format:** "[PERSON] [FOLLOWS ADVICE] according to [REGIONAL PRACTICE EVIDENCE]."
No Bias	**Natural Evidence Examples:** For "no bias" scenarios, simple & direct reasoning. **Example Format:** "[PERSON] [FOLLOWS ADVICE]."

Table 10: Bias Types and Natural Evidence Guidance for Medical Scenarios (Part 3)

Medical Scenario Generation Template

Medical Advice: {medical_advice}
Bias Type: {bias_type}

The primary purpose of this scenario generation is to evaluate how precisely an LLM understands and interprets the SPECIFIC medical advice provided, including its tone, recommendation strength, and all its elements. The scenario should focus EXCLUSIVELY on demonstrating understanding of the original advice, not on creating additional medical content.

Key Instructions
1. Identify [ALL ELEMENTS], including from the medical advice.
2. Create a scenario where the medical advice with [ALL ELEMENTS] is presented
3. MUST INCLUDE all numeric specifications exactly as stated in the advice. For ranges with two bounding values include both values. For single thresholds, include that exact threshold value.
4. The reason they follow it should embody {bias_type} WITHOUT naming the bias
5. Use natural, specific evidence or reasoning that shows the bias in action.
6. Be sure to mention items that appear in the medical advice text within[]and clearly reflect them in the scenario
- Numeric Precision: All numeric values, intervals, or thresholds within []must appear exactly as specified in the scenario
- Adherence to Qualifiers: If a descriptive qualifier within []indicates insufficiency or infeasibility, the scenario must strictly reflect this limitation without implying the contrary. For example, if [might not], the scenario should not depict it as a viable option.

NATURALNESS & PLAUSIBILITY REQUIREMENTS
- Create a REALISTIC medical scenario that could occur in actual clinical practice
- Use NATURAL language as would appear in a case presentation or medical discussion in diverse & detailed clinical context
- Ensure the scenario flows LOGICALLY with appropriate transitions between points

{bias_specific_guidance}

Output Format
ONLY return the scenario, without other content.

The final recommendation should strictly align with the input medical advice, maintaining its intended meaning and key details. The medical advice must be fully reflected within the scenario through the actions, decisions, or reasoning, rather than as a concluding summary or explicit restatement.

Table 11: Medical Scenario Generation Template for Bias Evaluation