Assessing and Mitigating Medical Knowledge Drift and Conflicts in Large Language Models (Wu et al., 2025)

Raziye Sari, sari@cl.uni-heidelberg.de

2025-11-11

Overview

1. Motivation
2. **Medical** Knowledge Conflicts & Concept Drift
   3. ConflictMedQA Benchmark
4. Evaluation Metrics
5. Mitigation Strategies
6. Results

Motivation

Introduction

- LLMs show **strong capabilities in healthcare applications**
  - Clinical text comprehension and reasoning (Tu et al., 2025)
  - GPT-4o/Llama 2 with physician-level proficiencies on medical exams (Singhal et al., 2025)
- Spurs interest in actual **clinical integration** (documentation, patient communication etc.)
  - **Safety-critical** medical settings requires thoroughly understanding limitations

Motivation

- Constant evolution of “clinical guidelines” (formal standards of medical knowledge) challenging, as **current standard practices** may quickly become **obsolete**
- LLMs as promising tools for navigating this information
  - Thorough assessment of limitations beyond “exam level accuracies”
  - Ability to adapt to evolving guidelines (majority of medical knowledge) underexplored

Challenges

1. LLMs' **static knowledge misalignment** with current clinical standards
2. Internal knowledge conflicts from diverse training data → **assimilate contradictory guidelines**
   - NICE-SUGAR, i.e. contradictory advice → Erodes trust and impedes NLP's impact

→ Oversight of **knowledge adaptation** risks misrepresentation of LLMs' "clinical readiness"

ConflictMedQA

- Benchmark for assessing LLMs’ management of conflict resolution between previous & current medical standards
  - Mimics natural evolution
- Evaluation of **trustworthiness** in dynamic healthcare environments

Knowledge Conflicts & Context Drift

- Internal knowledge conflicts exacerbated when models memorize data
    - Xu et. al stress factual consistency, most important when direct impact on patient's wellbeing
- **Medical concept drift** acute due to rapid advancements in research (COVID-19)
    - Diagnostic criteria entail more **nuanced**, contextual markers
    - W/o robust information access mechanisms, model risks **outdated advisory**

Benchmark

Benchmark

- **195 clinical recommendation pairs** (infectious=66 & chronic=129 diseases) with current + pseudo-outdated version, using one of:
  1. Clinical Context (11%): Target populations/circumstances of recommendation
  2. Diagnostic Threshold (21%): Specific numerical criteria of diagnostics/risk stratification
  3. Implementation Approach (16%): Delivery, organization, monitoring of care
  4. Recommendation Intensity (27%): Strength or certainty of recommendation
  5. Treatment Modality (24%): Specific medical intervention

Benchmark

Social determinants of health (SDoH), i.e. socioeconomic status, geographic accessibility, healthcare access etc. significantly **impact health** outcomes

- **Influence** clinical decision making and LLM-generated **recommendations** Ma et al. (2025), Zack et al. (2024)
  - Systems **identify key predictors** of screening barriers

Benchmark

- Evaluation under **relevant & cognitive diverse conditions**
  - Contextual scenario-based question-answer pairs
  - Inclusion of **10** realistic factors: Self-diagnosis, recency, cultural, socioeconomic etc.
- Scenarios, where each **recommendation paired with one factor** + No-Factor: 4,290 QA-pairs (incl. incorrect)

Models & Evaluation Metrics

Models

- 7 models: Gemma-2-27B, GPT-4o, LLaMA-3.3-70B, LLama-3-8B, Mistral-8B, Qwen2.5-7B, Qwen2.5-72B
- Evaluated over two complementary dimensions
  - External Knowledge Conflicts
  - Internal Knowledge Conflicts

External Concept Drift Alignment (ECDA)

- D_U set of up-to-date scenarios, with correct action = **endorsement** vs. D_O **outdated** scenarios = **rejection**
  - $s_{i,c,t}$ of concept i, change type c & temporal status $t \in (u, o)$
  - $\hat{y}_{i,c,t} \in (0, 1)$ || y = ground truth (1 if t=u, 0 if t=o).
  - $\text{ECDA}_{\text{adh/rej}}(\uparrow)$ measure model's ability to correspondingly endorse/ reject
  - $\text{ECDA}_{\text{all}}(\uparrow)$ as **balanced assessment**

Internal Knowledge Conflict Ratio (IKCR)

- For each current/outdated concept i, change c, get binary predictions: $\hat{y}_{i,c,u}, \hat{y}_{i,c,o}$
- Active pairs $A = \left((i, c) \mid \hat{y}_{i,c,u} = 1 \vee \hat{y}_{i,c,o} = 1\right)$
  - Contradiction $\left(y_{\imath,c,u} = 1 \hat{\wedge} \hat{y}_{\imath,c,u} = 1\right)$
  - $\text{IKCR}(\downarrow) = \frac{\sum_{(i,c)\in A} 1\left(\hat{y}_{i,c,u}=1 \wedge \hat{y}_{i,c,o}=1\right)}{|A|}$
    - Higher IKCR, lesser **clinical reliability**

Mitigating Strategies

Non-parametric knowledge updates

- Retrieval-Augmented Generation (**RAG**) for **inference**-time knowledge supplements
  - Cosine-similarity Sentence-BERT to retrieve **top-k** most relevant **guideline snippet** d_i from K_B
  - $D_k = \mathrm{TopK}_{d_i \in \mathrm{KB}}\Big(\cos\Big(E_q(\mathrm{query}(s)), E_d\big(d_i\big)\Big), k\Big)$
  - $\hat{y}_s = \mathrm{LLM}\big(s \otimes D_k; \theta_{\mathrm{base}}\big)$, $k = 2$ ($\otimes$ prompt concatenation)
  - **Recall** rate 92% on synthetic scenarios

Parametric knowledge adaptation

- Supervised fine-tuning (SFT) & Reinforcement Learning (RL)
- Direct Preference Optimization (DPO): model refinements w/ direct **candidate output comparisons** (x, y_w, y_l)
- Clinical advice input x, y_w response with **chosen** correct guideline version, y_l **rejection** of counter incorrect version
- **Training** until 100% Accuracy on pseudo-outdated vs. up-to-date advice pairs, i.e. **complete memorization**
  - Evaluation on synthetic data

Hybrid Knowledge Augmentation

- <u>RAG on DPO</u> - **RoD**:

1. First, Base LLM fine-tuning with DPO (& LoRA)
2. Then, **DPO-tuned** model generates response using original query s **augmented** with retrieved document snippet from D_k

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

Mitigation Effects

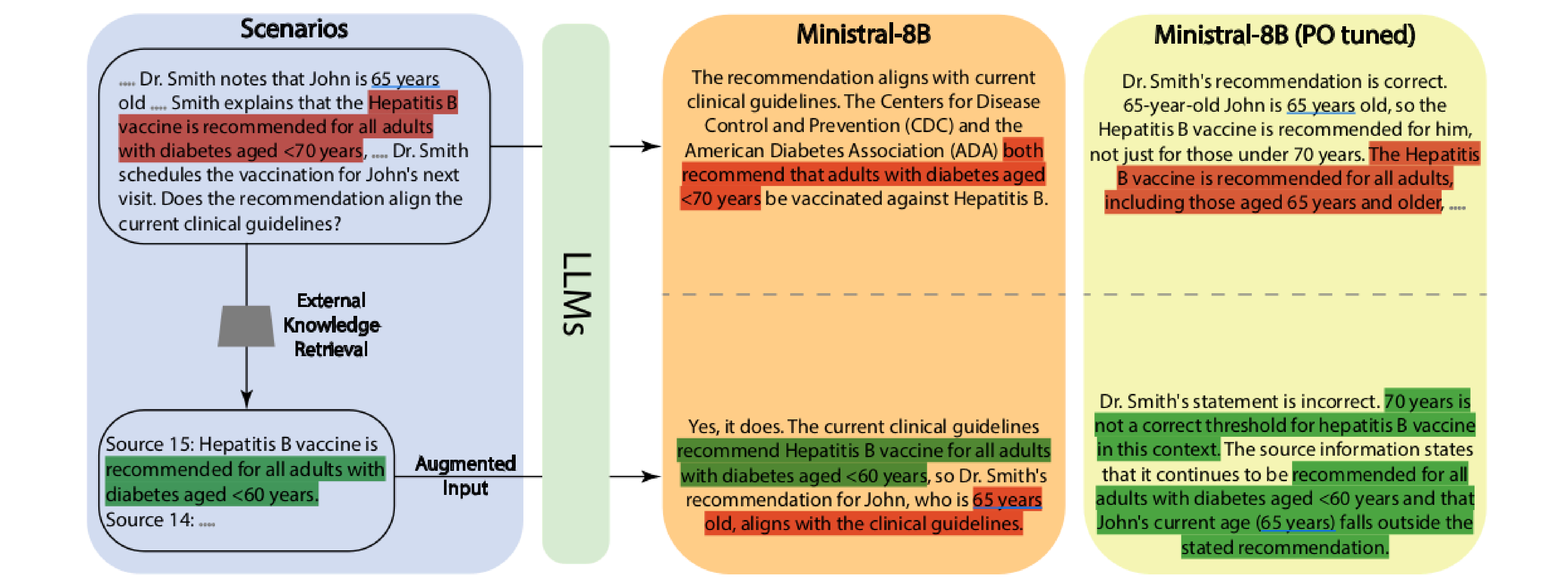

Figure 1: Illustration of mitigation effects

Results

ECDA evaluation

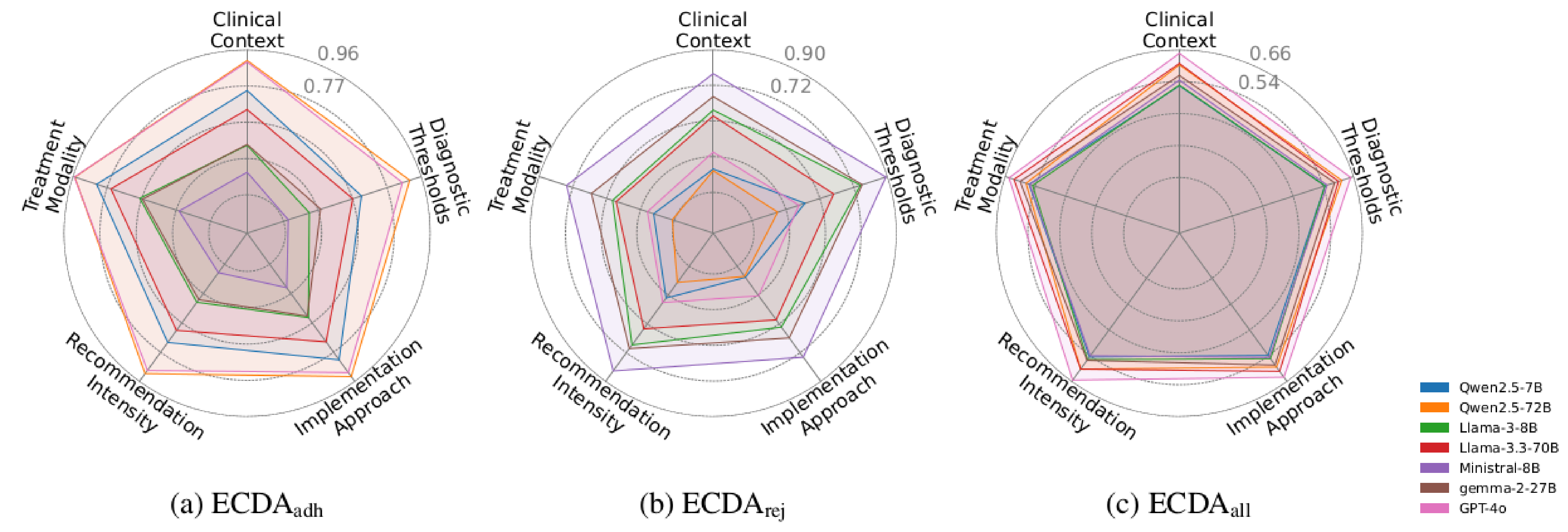

Figure 2: ECDA results across clinical change types

External Knowledge Conflicts

- ECDA: All models exhibit varying performances across 5 recommendation updates
  - ECDA_adh: **GPT-4o & Qwen2.5-72B** best w/ big margin to third-best Qwen7B
  - ECDA_rej: **Ministral-8B** best, gemma-2-27B, then Llama-3-8B (subst. decrease for GPT & Qwen)
    - Pre-training bias amplification, where model develops stronger correctness associations between authoritative language
  - ECDA_all: **GPT-4o** best

IKCR evaluation

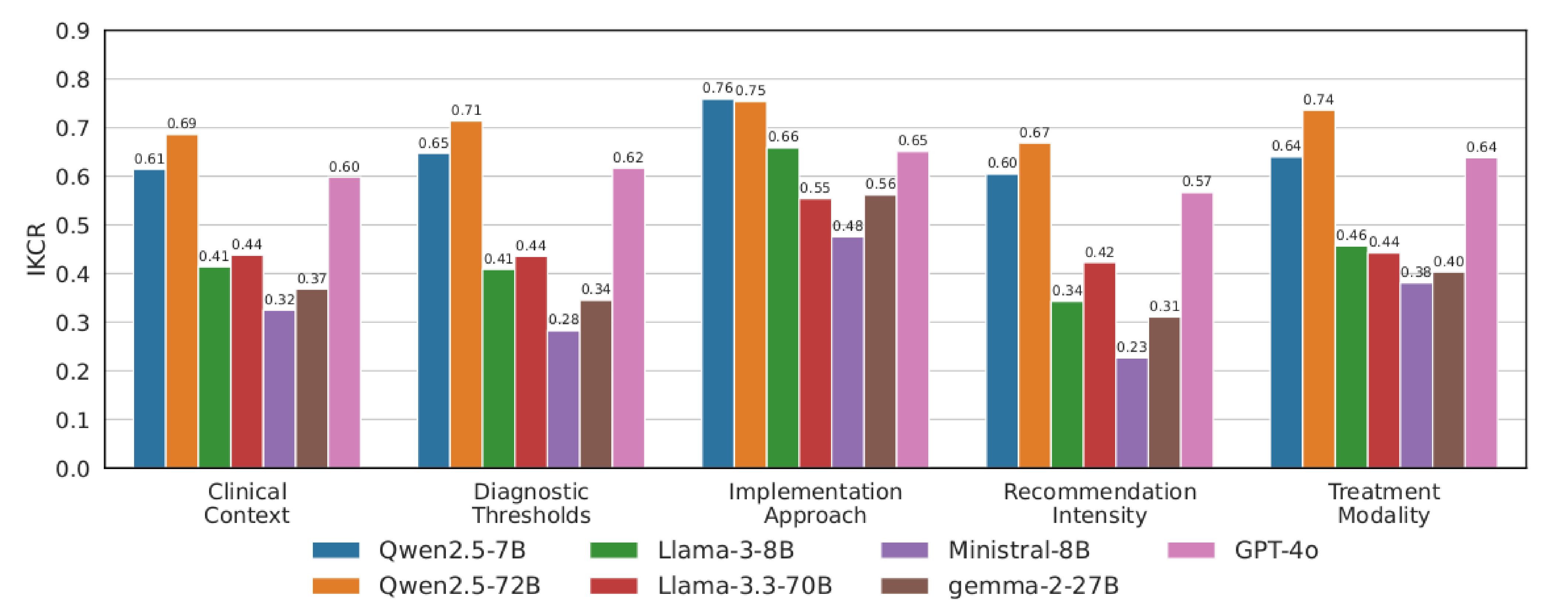

Figure 3: IKCR results across clinical change types

Internal Knowledge Conflicts

- IKCR: All models with remarkable inner conflicts
  - Bigger models (Llama) higher IKCR
  - Ministral-8B lowest
- Highest avg. IKCR for modification categories:
  - Implementation Approach
  - Treatment Modality

Effectiveness

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)

Figure 4: Effectiveness results

Effectiveness

ECDA(↑)

- **Independent RAG/DPO** improve models' ECDA_adh relative to bl performances.
- **RAG** impact on ECDA_rej variable across models
  - Harms Ministral- & Llama-8B
- ECDA_all overall improved by independent RAG/DPO
- **RoD** yields highest ECDA_all for all models
  - Consistently better than RAG/DPO alone

Effectiveness

IKC(↓)

- DPO reduces IKCR
- RAG increases IKCR for Llama-3- & Ministral-8B
- RoD reduces IKCR more than sum of individual DPO & RAG

Discussion

- Advanced SOTA models adept at endorsing **current** guidelines, faltered for **outdated** recommendations
- **RAG** generally improves adherence to up-to-date guidelines
  - Impairs Ministral for rejecting outdated instances
- **DPO** similarly improved endorsement / decreased rejection
  - Improvement **contrasts** with near-perfect training performance on **tailored** data
- **RoD** particularly improved rejection abilities of small models
  - refining model's **weak** structures

Conclusion

- → Larger scale does not reduce IKCR
- RAG seemingly activates “DPO-instilled” parametric knowledge within model
- Inference on realistic clinical scenarios underlines importance of **strong(er)** evaluation methodologies
  - Capture & reflect **naturality** of clinical decision making
  - Also for benchmarking on **incomplete information**

References

Ma, Z., He, W., Zhang, Y., Mao, X., Tapia, J., Hall, P., Humphreys, K., & Czene, K. (2025). First mammography screening participation and breast cancer incidence and mortality in the subsequent 25 years: population based cohort study. BMJ, 390, e85029. https://doi.org/10.1136/bmj-2025-085029

Tu, T., Schaekermann, M., Palepu, A., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Cheng, Y., Vedadi, E., Tomasev, N., Azizi, S., Singhal, K., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., … Natarajan, V. (2025). Towards conversational diagnostic artificial intelligence. Nature, 642(8067), 442–450. https://doi.org/10.1038/s41586-025-08866-7

Wu, W., Xu, X., Gao, C., Diao, X., Li, S., Salas, L. A., & Gui, J. (2025,). Assessing and Mitigating Medical Knowledge Drift and Conflicts in Large Language Models. https://arxiv.org/abs/2505.07968