MedPAIR : Measuring *Physicians* and *AI Relevance* Alignment in Medical Question Answering

Yuexing Hao, Kumail Alhamoud, Hyewon Jeong, Haoran Zhang, Isha Puri, Grace Yan,
Philip Torr, Mike Schaekermann, Saleh Kalantari, Ariel D. Stern, Marzyeh Ghassemi

Go visit our project website: https://medpair.csail.mit.edu/

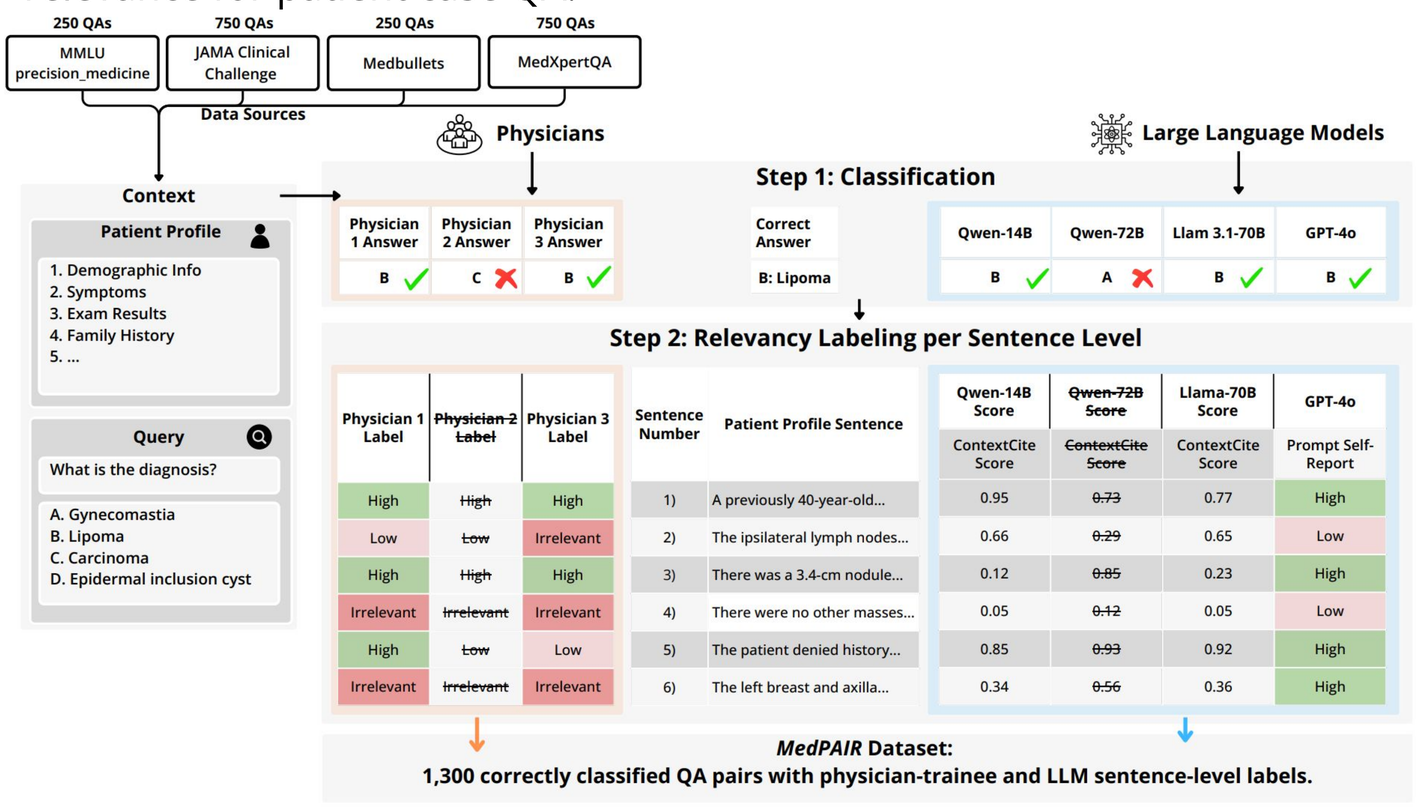
Scan Me

Physicians **think core evidence is ....**
LLMs **think core evidence is .....**
Are they *aligned*??

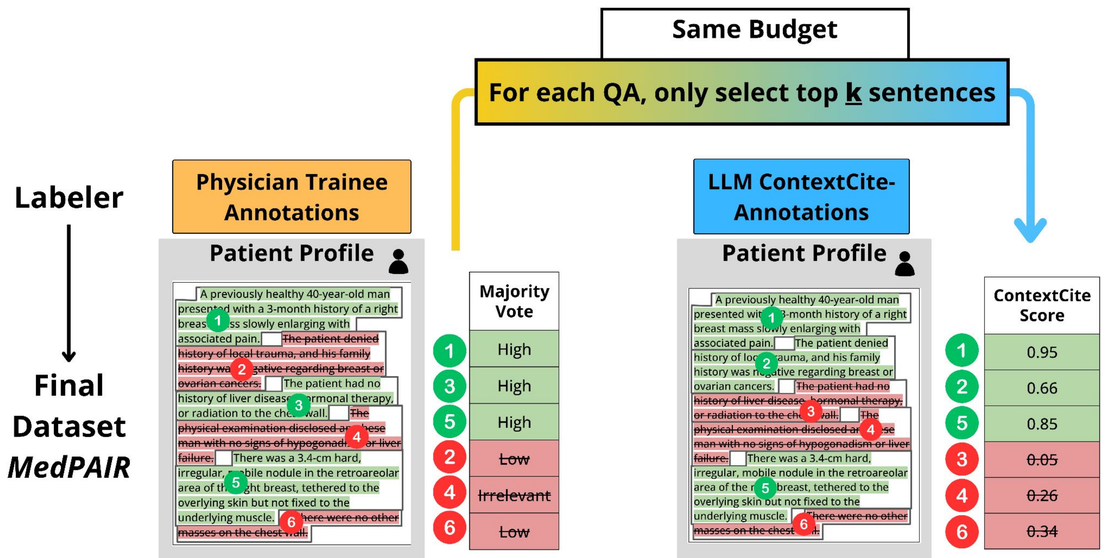
Our benchmark study finds out that the alignment is low (44.9 - 65.9%).

Study Design

We consolidated ***four*** QA datasets into patient profiles and queries. In ***933*** examples, ***52*** physician trainees and ***6*** LLMs selected answers, followed by sentence-level relevance annotations from trainees (excluding those tied to incorrect answers). Majority vote produced binary labels. We release the first benchmark and open dataset of physician trainee-annotated relevance for patient case QA.



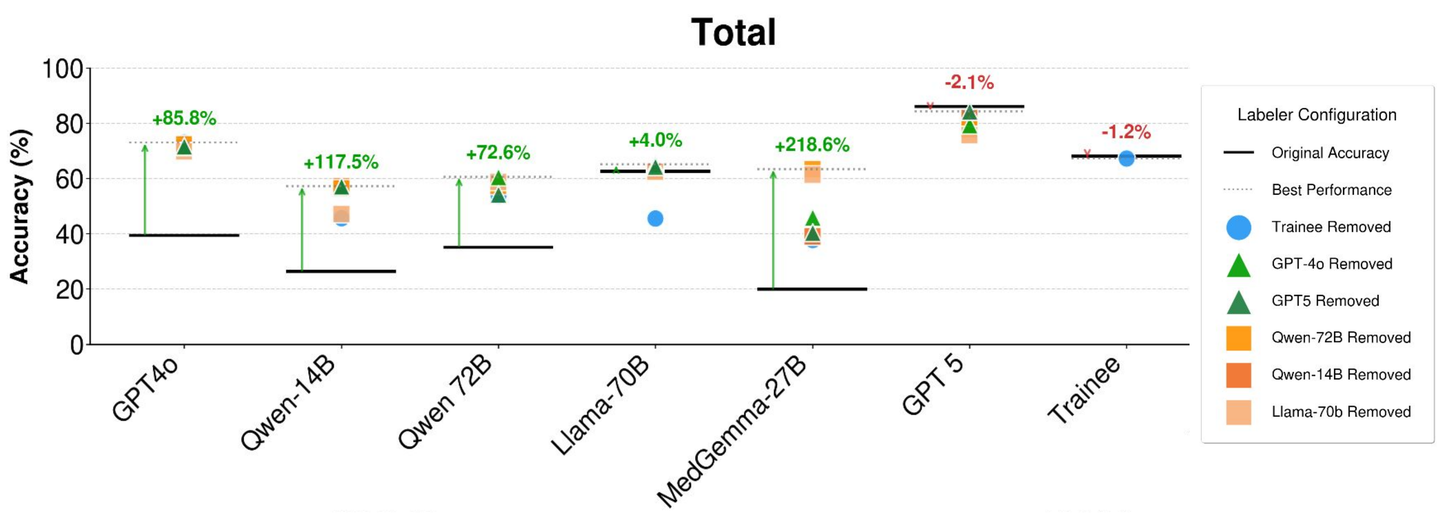
Evaluation Design



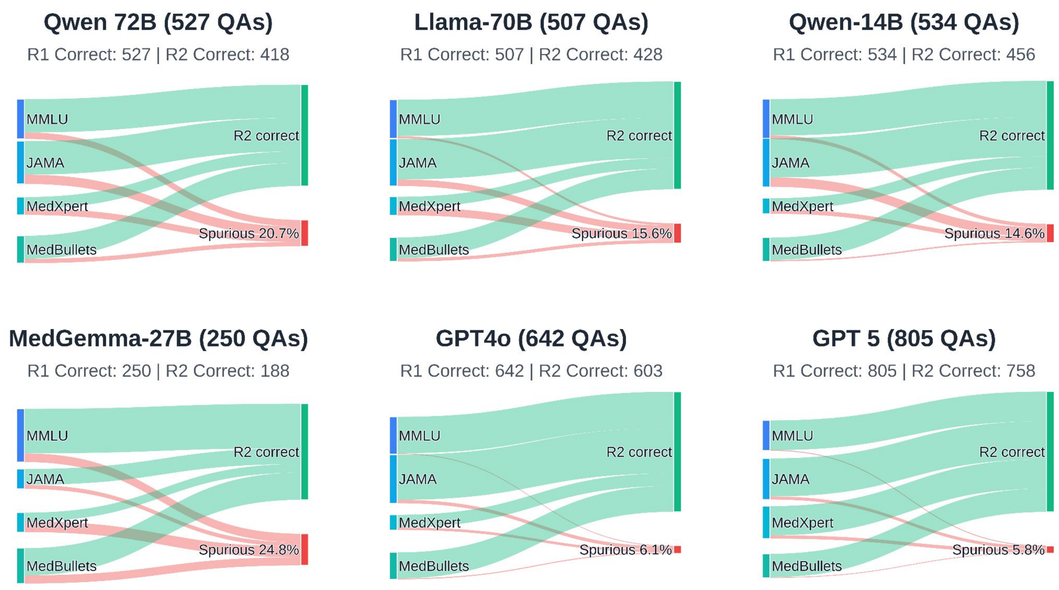
Evaluation metrics to align physician trainees' annotations with LLM ContextCite raw scores using an identical input context budget.
We then selected the k sentences with the highest raw ContextCite scores and labeled them "*high relevance*." The remaining sentences were ranked and assigned to "*low relevance*" or "*irrelevant*" based on their score order. This alignment creates a direct mapping between LLM attributions and human judgments, allowing us to assess how well the model's sentence rankings match expert annotations.

Results

MedPAIR reveals fundamental differences between physician trainees and LLMs in identifying clinically relevant information, with the highest model concordance (Llama-70B) reaching only **65.9%**.



Effect of Filtering Context on Final Performance
Frontier models show minimal sensitivity to low relevant and irrelevant sentence removals. GPT-5 is unaffected as its original accuracy matching its best across every benchmark. GPT-4o and the Qwen models gain small improvements. However, weaker and domain specialized models, show substantial improvements.



Spurious Rate of removing physician-identified low-relevance and irrelevant sentences out of 933 QAs.

Conclusion

The MedPAIR benchmark evaluates alignment between LLM relevance judgments and physician-trainee annotations in medical QA. Using relevance pairs, it **identifies key context** and **reveals mismatches** in relevancy estimates.