ICSPIS2025

ICSPIS2025

11th International conference on Signal Processing and Intelligent Systems

Enhancing Reasoning Skills in Small Persian Medical Language Models Can Outperform Large Scale Data Training

Mehrdad Ghassabi, Zahra Kazemi, et al.

Motivation

Medical question answering requires deliberate, multi-step reasoning, yet most small language models struggle with such “slow thinking,” especially in low-resource languages like Persian.

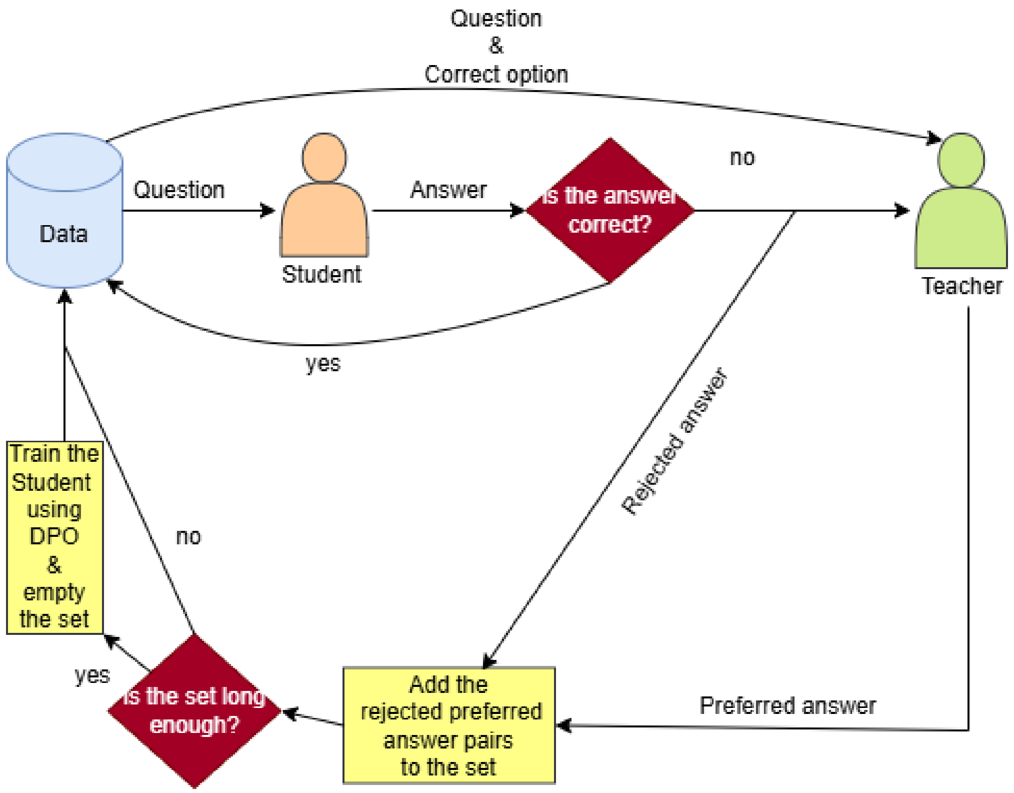
- Medical QA requires multi-step reasoning
- Small LLMs struggle with “slow thinking”
- Persian medical data is extremely limited
- Data scaling and RLHF are costly

Proposed Method

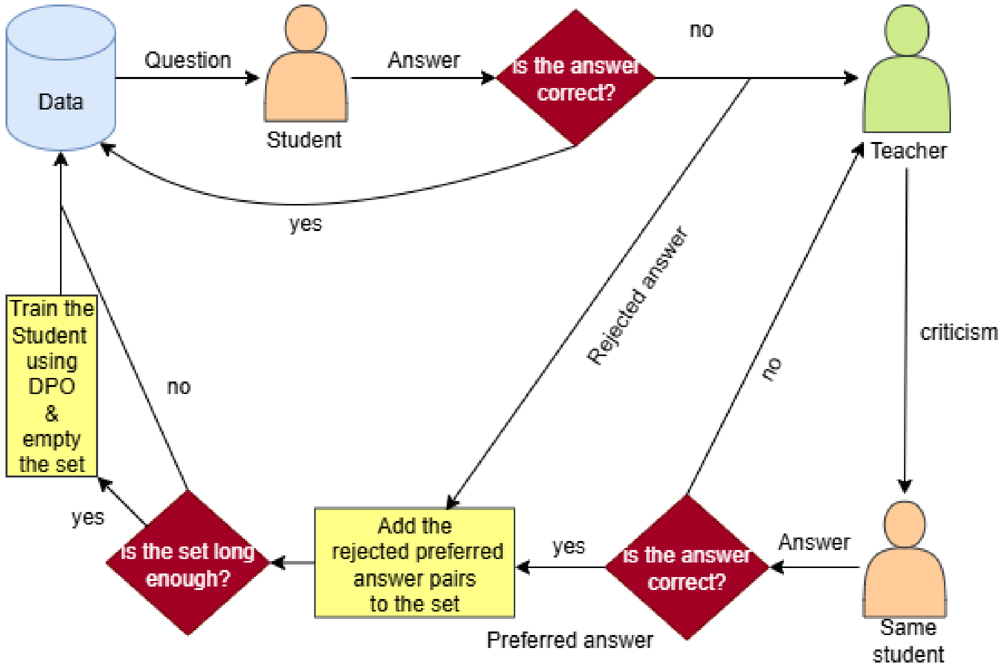
We introduce a reasoning-centered training framework based on RLAIF and DPO, develop gaokerena-R as an 8B Persian medical reasoning model.

The final dataset contains over 18,000 verified Persian medical questions and 11,000 reasoning preference pairs, evaluated on FA_MED_MMLU and IBMSEE benchmarks.

- 11K reasoning preference pairs
- 18K verified Persian medical MCQs
- Benchmarks: FA_MED_MMLU, IBMSEE
- Models: gaokerena-R, gaokerena-V, aya-expanse-8b



Method 1.
Teacher Correction for Reasoning Alignment

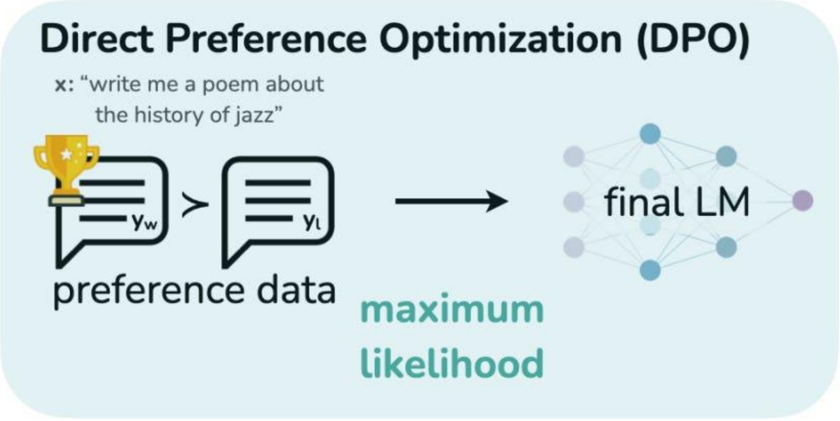


Method 2.
Teacher-Guided Self-Correction

RLAIF + DPO Optimization

Reinforcement Learning with AI Feedback (RLAIF) is used to generate high-quality reasoning preference pairs without human annotation.

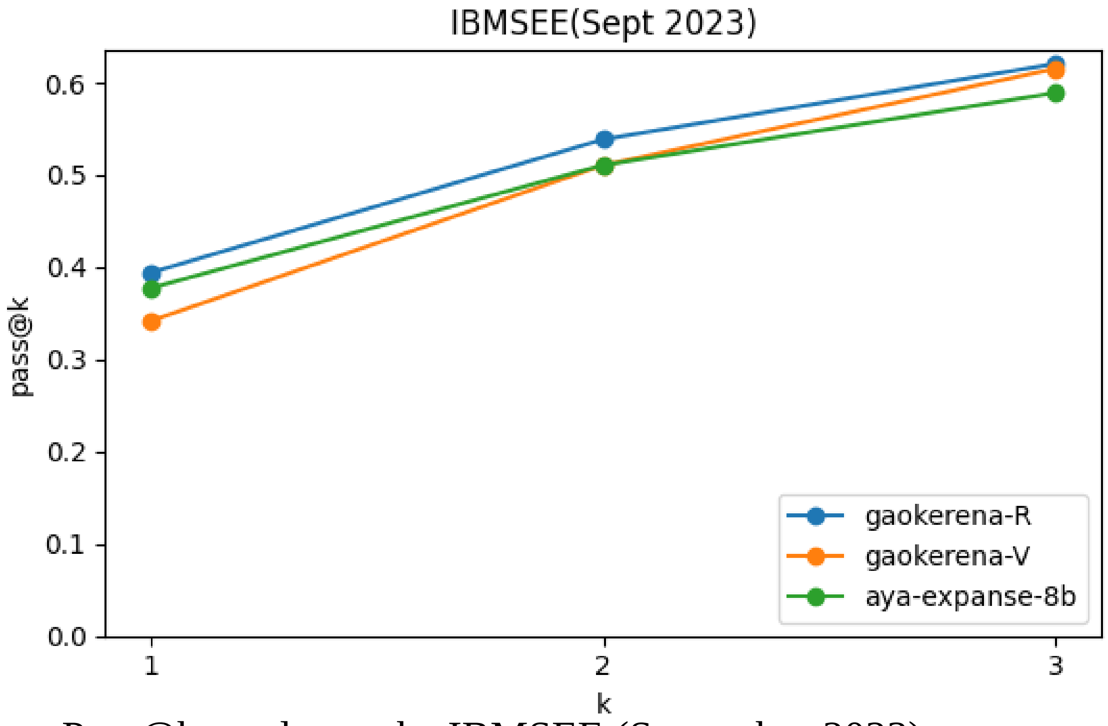
Direct Preference Optimization aligns the student model by increasing the likelihood of valid reasoning paths and suppressing flawed ones, without requiring an explicit reward model.


Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w
>
y_l
preference data
maximum likelihood
final LM

Evaluation Metric: Pass@K

Pass@K measures the probability that a model produces at least one correct answer within *K* independent attempts.

High Pass@K values, especially for small *K*, indicate stable and non-random reasoning behavior.



Pass@k results on the IBMSEE (September 2023) dataset using Chain-of-Thought prompting

Results: Reasoning Ability (Chain-of-Thought)

We evaluate reasoning performance using Chain-of-Thought prompting on medical benchmarks.

These results confirm that reasoning-focused optimization is more effective than data scaling for small, low-resource models

	gaokerena-R	**gaokerena-V**	**Aya-expanse-8b** (baseline)
MMLU- medical-genetics(fa)	**50.0**	41.0	45.0
MMLU(avg)	**48.76**	40.40	47.10
IBMSEE Sept2023	**38.69**	29.76	35.71

From table 1.Chain-of-Thought Prompted Performance Without Negative Marking

Results: Medical Knowledge & Prompt Dependenc

gaokerena-V excels with direct prompts, while gaokerena-R performs best with CoT prompts.

	gaokerena-R	**gaokerena-V**	**Aya-expanse-8b** (baseline)
MMLU- medical-genetics(fa)	49.0	**53.0**	49.0
MMLU(avg)	46.28	**49.31**	46.64
IBMSEE Sept2023	35.11	**38.69**	34.52

From table III. Straight Prompted Performance

Overall Performance: Hybrid Setup

We propose a hybrid configuration combining gaokerena-R’s reasoning ability with a knowledge-based verifier that verifies uncertain cases.

This hybrid approach achieves the best overall performance, outperforming gaokerena-V under direct prompting.

	gaokerena-R+ **Aya-expanse-8b** (verifier)	**gaokerena-V**
MMLU- medical-genetics(fa)	**56.0**	53.0
MMLU(avg)	**58.86**	55.47
IBMSEE Sept2023	**52.98**	49.31
prompt	COT for the main model Straight for the verifier	Straight

From table IV. Evaluation of two different configurations

Conclusion & Future Work

Our results show that targeted reasoning optimization can outperform data scaling in low-resource medical NLP.

The proposed RLAIF + DPO framework is efficient, cost-effective, and environmentally sustainable.

Future work will focus on building prompt-invariant medical language models that unify strong knowledge and stable reasoning.

Thanks for your Attention!

Our data will be available at:

- https://github.com/Mehrdadghassabi/Gaokerena-R