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

1st Mehrdad Ghassabi
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
m.ghassabi@eng.ui.ac.ir

2nd Sadra Hakim
School of Computer Science
University of Windsor
Windsor, Canada
hakim6@uwindsor.ca

3rd Hamidreza Baradaran Kashani
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
hrb.kashani@eng.ui.ac.ir

4th Pedram Rostami
School of Electrical and Computer Engineering
University of Tehran
Tehran, Iran
pedram.rostami@ut.ac.ir

***Abstract*—Enhancing reasoning capabilities in small language models is critical for specialized applications such as medical question answering, particularly in underrepresented languages like Persian. In this study, we employ Reinforcement Learning with AI Feedback (RLAIF) and Direct preference optimization (DPO) to improve the reasoning skills of a general-purpose Persian language model. To achieve this, we translated a multiple-choice medical question-answering dataset into Persian and used RLAIF to generate rejected-preferred answer pairs, which are essential for DPO training. By prompting both teacher and student models to produce Chain-of-Thought (CoT) reasoning responses, we compiled a dataset containing correct and incorrect reasoning trajectories. This dataset, comprising 2 million tokens in preferred answers and 2.5 million tokens in rejected ones, was used to train a baseline model, significantly enhancing its medical reasoning capabilities in Persian. Remarkably, the resulting model outperformed its predecessor, gaokerena-V, which was trained on approximately 57 million tokens, despite leveraging a much smaller dataset. These results highlight the efficiency and effectiveness of reasoning-focused training approaches in developing domain-specific language models with limited data availability.**

***Index Terms*—system2 deep learning,small language model,medical language models, RLAIF, direct policy optimization**

I. Introduction

In 2019, at the NeurIPS conference, Yoshua Bengio highlighted a critical limitation of current deep learning systems: their inability to perform tasks requiring robust reasoning skills [1]. While these systems excel at intuitive, perception-driven tasks, they struggle with reasoning-based challenges that demand deeper cognitive processing. This observation echoes Daniel Kahneman's distinction between "fast" and "slow" thinking in cognitive science [2]. Fast, intuitive thinking aligns with the strengths of current AI systems, whereas slow, deliberate reasoning remains a significant weakness.

One proposed solution to this deficiency is the revival of symbolic AI methods through neurosymbolic approaches. However, Bengio cautioned against such methods due to scalability issues. Instead, he urged the development of new architectures capable of out-of-distribution generalization—a core requirement for genuine reasoning.

The Transformer architecture [3], which underpins most modern language models, suffers from the same deficiency. Transformers often make "stupid mistakes" when faced with reasoning-intensive problems, prompting researchers to enlarge models and datasets in an attempt to achieve better results across tasks. This approach arguably simulates System 2 reasoning within a fundamentally System 1 framework.

This limitation becomes even more pronounced in smaller language models, where restricted capacity and limited data exacerbate reasoning challenges. In domains such as medicine—where reasoning, interpretation, and decision-making are essential—these weaknesses become especially consequential.

Recent efforts to enhance reasoning have largely focused on prompting or fine-tuning methods that enable models to imitate reasoning behavior rather than truly engage in it—a phenomenon we refer to as the "illusion of thinking" [4]. In the absence of architectures with intrinsic reasoning capabilities, such approaches provide only an approximation of genuine cognitive reasoning. Nevertheless, these efforts represent necessary steps toward achieving more reasoning-capable AI systems.

The present work aims to address this challenge by enhancing the reasoning abilities of a small Persian language model, aya-expanse-8b [5]. We first prompt the model to generate Chain-of-Thought (CoT) trajectories for multiple-choice medical questions. These responses are then reviewed and corrected using DeepSeek-R, a model adept at reasoning, to rectify

errors. This process constitutes Reinforcement Learning with AI Feedback (RLAIF) [6]. By pairing incorrect and corrected CoT trajectories, we subsequently apply Direct Preference Optimization (DPO) [7] to enhance the reasoning capabilities of the baseline model.

We call this improved model gaokerena-R. While its predecessor, gaokerena-V [8], demonstrated superior medical knowledge in straightforward prompting tasks, gaokerena-R outperformed it when evaluated using Chain-of-Thought (CoT) prompting [9]. Notably, gaokerena-R was trained on a much smaller dataset, focusing on reasoning enhancement rather than data scale, whereas gaokerena-V relied on extensive training on a large medical corpus and the MF3QA dataset [8].

Our findings highlight an important insight: in low-resource domains such as Persian medical AI, strengthening reasoning capabilities can be more effective than training on vast datasets. This underscores the importance of developing reasoning-oriented approaches, particularly in languages and fields with limited data availability. By improving reasoning abilities in smaller models, we can make meaningful progress toward practical, resource-efficient AI systems.

II. Related Work

To the best of our knowledge, no prior work has focused on developing Persian medical reasoning language models. Existing Persian medical models, including our previous Gaokerena-V, mainly address knowledge representation and general language understanding with limited attention to reasoning.

L. Pan et al. reviewed several valuable studies on enhancing reasoning capabilities in language models. [10] Accordingly, we focus here on English-language works that offer methodological insights into generating and integrating reasoning data for improving medical language models.

A. Related Work In Medical Domain

One of the notable efforts in developing small-scale medical reasoning language models is MedSSS [11]. This framework focuses on fine-graining the intermediate reasoning steps to enhance the reasoning ability of medical models. To accomplish this, the authors applied the Monte Carlo Tree Search (MCTS) [12] algorithm on medical multiple-choice question-answering datasets to synthesize structured reasoning trajectories using a policy model. Based on this approach, they constructed three datasets supporting different stages of training: a dataset for Supervised Fine-Tuning (SFT) of the policy model, a rejected–preferred answer dataset for Direct Preference Optimization (DPO) [13] ,and a soft dual-side label dataset for fine-tuning the Process Reward Model (PRM).Finally, the trained Policy Model was used as the primary reasoning policy, while the trained Process Reward Model acted as a decoding guide to refine and evaluate the reasoning process during generation.

Another significant contribution in this field is MedReason [14]. In this work, the authors leveraged a structured medical knowledge graph to transform conventional question–answer (Q&A) pairs into detailed medical reasoning trajectories. Each trajectory represented the step-by-step logical pathway connecting a clinical question to its correct answer, grounded in medical knowledge relationships such as symptoms, diagnoses, treatments, and physiological mechanisms. By constructing a reasoning-enriched dataset in this manner, the authors were able to fine-tune a baseline language model on these structured reasoning examples. This approach led to a marked improvement in the model's ability to perform reasoning-intensive tasks within the medical domain, demonstrating the effectiveness of incorporating knowledge graph–based reasoning supervision into the training process. The findings from MedReason further emphasize the importance of structured reasoning representations as a means of improving the interpretability and analytical depth of medical language models.

Another important advancement in the development of medical reasoning models is HuatuoGPT-o1 [15]. In this work, the authors introduced a verification-guided reasoning framework designed to improve the logical consistency and accuracy of generated reasoning trajectories. Specifically, they employed a verifier model to assess and guide the policy model during reasoning generation, ensuring that each reasoning path adhered closely to factual correctness and medical plausibility. By filtering and refining reasoning trajectories through this verification process, they created a high-quality dataset composed of verified reasoning sequences. The authors then leveraged both supervised fine-tuning and reinforcement learning techniques to train their baseline language model using these verified data. This dual training approach allowed the model to not only imitate correct reasoning behaviors but also internalize reasoning principles through reward-driven optimization. As a result, HuatuoGPT-o1 demonstrated significant improvements in reasoning accuracy and reliability across a variety of medical question-answering and diagnostic tasks, highlighting the potential of verifier-guided learning frameworks in advancing medical reasoning language models.

B. Related Work In Other Domain

Perhaps the most influential recent work in the broader field of AI reasoning is DeepSeek-R [16]. Building upon the DeepSeek-V3-Base model, the authors introduced a reinforcement learning framework based on Gradient Regularized Policy Optimization (GRPO) [17] to explicitly enhance the model's reasoning capabilities. In this setup, the reinforcement learning process was guided by a reward function specifically designed to evaluate and maximize reasoning performance. Through this training paradigm, DeepSeek-R demonstrated remarkable improvements in logical reasoning and problem-solving accuracy across various benchmarks. However, the application of reinforcement learning also introduced several side effects. While reasoning performance improved substantially, the model's readability and linguistic coherence deteriorated, and instances of language mixing became more frequent. To address these issues, the authors incorporated a small amount of cold-start supervised data and adopted a multi-stage training pipeline. This additional fine-tuning phase helped

restore natural language fluency and readability while retaining the strong reasoning skills acquired through reinforcement learning. The resulting model, DeepSeek-R, thus represents a critical step forward in reasoning-oriented AI, demonstrating that reinforcement learning can significantly enhance reasoning ability—provided it is balanced with targeted fine-tuning to maintain linguistic quality.

Another notable contribution in the area of reasoning enhancement is Thought Preference Optimization (TPO) [18]. In this work, the authors proposed a preference-based framework for improving reasoning quality in language models. Given a question, the model first generates multiple candidate reasoning trajectories. These responses are then evaluated by a judge model, which identifies the best and worst samples based on reasoning correctness and coherence. The collected best–worst pairs are subsequently used to train the baseline model through Direct Preference Optimization (DPO), encouraging it to prefer higher-quality reasoning paths. This approach effectively aligns the model's reasoning process with human-like evaluative feedback, demonstrating that reasoning quality can be substantially improved through structured preference optimization rather than relying solely on scale or supervised data.

Another relevant study was conducted by N. Ho et al. [19], who explored the transfer of reasoning capabilities from large language models to smaller ones. In their approach, a smaller model was fine-tuned using data generated by a larger model that exhibited stronger reasoning performance. The larger model produced reasoning trajectories and question–answer pairs that served as high-quality supervision signals for the smaller model. Through this distillation process, the smaller model effectively learned reasoning strategies and problem-solving patterns from its larger counterpart, achieving competitive reasoning performance with significantly reduced computational cost. This work demonstrates that reasoning ability can be efficiently transferred across models of different scales through targeted fine-tuning on reasoning-oriented synthetic data.

III. Proposed Methods

IV. Results

In this section, we compare the newly developed Gaokerena-R model with its predecessor, Gaokerena-V. While Gaokerena-V was trained on a large medical corpus and demonstrates strong factual knowledge and retrieval capabilities, Gaokerena-R was specifically designed to enhance medical reasoning. Owing to its reasoning-focused training pipeline, Gaokerena-R was trained on a substantially smaller dataset, resulting in slightly reduced coverage of general medical knowledge. However, this trade-off enabled it to develop deeper reasoning competence, allowing it to perform better on tasks requiring multi-step inference and logical consistency.

In the final evaluation, we compared the performance of Gaokerena-V under direct (straight) prompting with that of Gaokerena-R when provided with Chain-of-Thought (CoT) prompts. The results highlight that Gaokerena-R, despite its smaller scale and limited training data, achieves superior reasoning performance through structured reasoning guidance, demonstrating the effectiveness of reasoning-centered optimization over pure data scaling.

A. Medical Reasoning Capabillities

To assess the medical reasoning capabilities of the models, we performed evaluations using the FA_MED_MMLU dataset [1]. Both Gaokerena-V and Gaokerena-R were prompted [2] to generate Chain-of-Thought (CoT) reasoning traces with a temperature setting of 1.0. For each question, five independent samples were generated, and a majority-voting mechanism was applied to determine the final answer: if three or more of the five generations selected the same option, that option was chosen as the final prediction; otherwise, the question was left unanswered to reflect model uncertainty.

This evaluation framework provides a more robust estimation of reasoning consistency and agreement across multiple reasoning trajectories. The results are presented in Table I for the without negative marking setting, and in Table II for the with negative marking setting. These two scoring schemes allow for a fair comparison of the models' reasoning accuracy under different evaluation criteria.

TABLE I
CHAIN-OF-THOUGHT PROMPTED PERFORMANCE WITHOUT NEGATIVE MARKING

	gao kerena-R	**gao kerena-V**	**aya-expanse-8b** (baseline)
MMLU-anatomy(fa)	**42.22**	39.25	40.74
MMLU-medical-genetics(fa)	**50.0**	41.0	45.0
MMLU-college-medicine(fa)	47.97	37.57	**48.55**
MMLU-clinical-knowledge(fa)	**55.84**	46.79	54.71
MMLU-professional-medicine(fa)	**44.85**	37.13	43.75
MMLU-college-biology(fa)	**48.61**	36.80	43.75
MMLU(avg)	**48.76**	40.40	47.10
IBMSEE Sept2023	**38.69**	29.76	35.71
Number of parameters	8b	8b	8b
inference time	$\approx 5 \times 35s$	$\approx 5 \times 35s$	$\approx 5 \times 35s$

B. Medical Knowledge

C. Final Evaluation

References

[1] Bengio, Yoshua. "From system 1 deep learning to system 2 deep learning." Neural Information Processing Systems. 2019.

[2] Kahneman, Daniel. "Thinking, fast and slow." Farrar, Straus and Giroux (2011).

[1] Available at huggingface.co/datasets/gaokerena/FA_MED_MMLU

[2] Evaluation prompts are available at github.com/Mehrdadghassabi/Gaokerena-R/blob/main/evaluations/zeroshot-COT/kopp/gaokerena-r1.0/Untitled2.ipynb

TABLE II
CHAIN-OF-THOUGHT PROMPTED PERFORMANCE WITH NEGATIVE MARKING

	gao kerena-R	**gao kerena-V**	**aya-expanse-8b** (baseline)
MMLU-anatomy(fa)	**29.13**	27.65	24.93
MMLU-medical-genetics(fa)	**40.0**	32.0	33.0
MMLU-college-medicine(fa)	**34.68**	25.24	34.48
MMLU-clinical-knowledge(fa)	**44.65**	35.59	42.51
MMLU-professional-medicine(fa)	**30.39**	25.0	**30.39**
MMLU-college-biology(fa)	**36.80**	25.0	30.09
MMLU(avg)	**36.14**	28.57	33.55
IBMSEE Sept2023	**24.60**	15.87	19.84
Number of parameters	8b	8b	8b
inference time	$\approx 5 \times 35s$	$\approx 5 \times 35s$	$\approx 5 \times 35s$

[3] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[4] Shojaee, P., et al. "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. Apple." 2025,

[5] Dang, John, et al. "Aya expanse: Combining research breakthroughs for a new multilingual frontier." arXiv preprint arXiv:2412.04261 (2024).

[6] Lee, Harrison, et al. "Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024." URL https://arxiv. org/abs/2309.00267 2309 (2023).

[7] Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." Advances in neural information processing systems 36 (2023): 53728-53741.

[8] Ghassabi, Mehrdad, et al. "Leveraging Online Data to Enhance Medical Knowledge in a Small Persian Language Model." arXiv preprint arXiv:2505.16000 (2025).

[9] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in neural information processing systems 35 (2022): 24824-24837.

[10] Pan, Liangming, et al. "Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies." arXiv preprint arXiv:2308.03188 (2023).

[11] Jiang, Shuyang, et al. "MedS^3: Towards Medical Slow Thinking with Self-Evolved Soft Dual-sided Process Supervision." arXiv preprint arXiv:2501.12051 (2025).

[12] Coulom, Rémi. "Efficient selectivity and backup operators in Monte-Carlo tree search." International conference on computers and games. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[13] Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." Advances in neural information processing systems 36 (2023): 53728-53741.

[14] Wu, Juncheng, et al. "Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. ArXiv, abs/2504.00993, 2025."

[15] Chen, Junying, et al. "Huatuogpt-o1, towards medical complex reasoning with llms." arXiv preprint arXiv:2412.18925 (2024).

[16] Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." arXiv preprint arXiv:2501.12948 (2025).

[17] Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." arXiv preprint arXiv:2402.03300 (2024).

[18] Wu, Tianhao, et al. "Thinking llms: General instruction following with thought generation." arXiv preprint arXiv:2410.10630 (2024).

[19] Ho, Namgyu, Laura Schmid, and Se-Young Yun. "Large language models are reasoning teachers." arXiv preprint arXiv:2212.10071 (2022).