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
sadrahakim@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

A. Related Work In Medical Domain

B. Related Work In Other Domain

III. Data

IV. Training

A. Direct preference Optimization

B. Carbon Footprint

V. Result

A. Result using straight prompt

B. Result using COT prompt

C. Comparing gaokerena-V & gaokerena-R

VI. Future research

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).

[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.