ICSPIS

11th International Conference on Signal Processing & Intelligent Systems
Mazandaran University of Science and Technology, December 23-24, 2025

Mazandaran University of Science and Technology

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

5th Zahra Kazemi
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
zhrakazemi@mehr.ui.ac.ir

***Abstract*— Reasoning is a critical requirement for medical language models, where incorrect or poorly justified outputs can have serious consequences. Despite the importance of this capability, prior work has largely focused on high-resource languages, leaving Persian—a widely spoken and medically relevant language—significantly underexplored. In this work, we address this gap by introducing two complementary post-training frameworks designed to enhance medical reasoning in Persian language models. Our central hypothesis is that, while training on massive medical corpora can improve factual knowledge, targeted post-training using preference-based optimization methods can yield greater gains in medical reasoning efficiency and reliability, even with substantially less data. The newly introduced model shares the same baseline as our previous model, gaokerena-V, which was obtained by training the aya-expanse-8b on 57 million tokens of web-crawled Persian medical data, in this work we apply small-scale post-training using Direct Preference Optimization (DPO) and Reinforcement Learning from AI Feedback (RLAIF). This post-training relies on only 11,000 AI-generated preferred–rejected response pairs, comprising approximately 2 million tokens in preferred answers and 2.5 million tokens in rejected ones, without requiring costly human annotation. Experimental results on a Persian-translated Medical MMLU benchmark demonstrate that this lightweight post-training approach outperforms gaokerena-V by a margin of 3.67 percent, highlighting that structured preference-based reasoning supervision can be more effective than large-scale domain data training alone for improving medical reasoning in low-resource languages.**

Keywords— System-2 Reasoning, Small-Scale Language Models, Medical Language Models, Reinforcement Learning with AI Feedback, Direct Preference Optimization

I. Overview

Transformer-based language models[1] excel at fast, intuitive tasks—such as pattern matching, retrieval, and surface-level text generation—mirroring Kahneman's concept of "fast thinking"[2]. However, they struggle with deliberate, multi-step reasoning tasks that require "slow thinking," particularly in specialized domains such as medicine, where diagnostic accuracy depends on logical inference, evidence evaluation, and error correction. This limitation is even more severe in low-resource languages such as Persian, where both high-quality data and compute are scarce.

In 2019, Yoshua Bengio warned that deep learning systems lack true reasoning capacity and called for architectures that support out-of-distribution generalization[3]. While the transformer architecture was revolutionary, its success has largely come from scaling: larger models and larger datasets yield better performance. Yet, despite current large language models having trillions of parameters and being trained on tens of trillions of tokens, they still make surprisingly simple reasoning errors. They may even produce inconsistent answers when asked the same question directly or via chain-of-thought (CoT) prompting[4]. This deficiency becomes even more pronounced in small medical Persian language models, which have far fewer parameters and much less training data. Recent advances have attempted to strengthen the reasoning capacity of small language models—enhancing their performance on "slow thinking" tasks—but these methods often rely on large, well-curated datasets available only for languages with extensive medical resources. In contrast, Persian lacks sufficient high-quality medical datasets. To tackle this challenge, we propose two frameworks aimed at enhancing the reasoning capabilities

978-1-5386-2585-9/25/$31.00 ©2025 IEEE

of small Persian medical language models under limited data conditions.

In our proposed method, we machine-translated an English multiple-choice medical question answering dataset into Persian and applied reinforcement learning from AI feedback (RLAIF) [5]and direct preference optimization (DPO)[6]to enhance the reasoning ability of a baseline Persian medical model. The resulting model is named gaokerena-R. [1]

Our prior work, gaokerena-V[7], fine-tuned a Persian medical language model using approximately 57 million tokens (including a portion of a medical corpus and a dataset) via supervised fine-tuning (SFT). Although gaokerena-V demonstrated strong medical knowledge, gaokerena-R outperforms it when given chain-of-thought prompts—despite being trained on less medical data. Since both models share the same baseline model, aya-expanse-8b , we hypothesize that enhancing reasoning skills is more beneficial than scaling data for small, low-resource medical language models.

The contributions of this study are outlined below:

1. Two efficient RLAIF+DPO frameworks that generates high-signal CoT preference pairs using a teacher–student loop.

2. gaokerena-R [2] , an 8b-parameter medical model demonstrating that reasoning-focused training outperforms data scaling in low-resource medical NLP.

3. A machine-translated Persian medical multiple-choice question answering dataset designed for reasoning-focused model training.

II. Previous Studies

Research on Persian medical language models remains limited. Among existing efforts, our prior work, gaokerena-V, utilized web-crawled data to fine-tune aya-expanse-8b, improving its medical knowledge[8] . However, to our knowledge, no studies in Persian have explicitly focused on enhancing reasoning capabilities. By contrast, the English-language literature offers a rich body of work on both knowledge and reasoning in medical language models, which we review in this section.

A. Previous Studies In Medical Domain

The MedSSS [9] is a notable, self-evolving system designed to instill robust, long-chain reasoning capabilities into small, deployable medical language models, such as the Llama3.1-8B-Instruct base model[10] . Its core methodology involves leveraging Monte Carlo Tree Search (MCTS) [11] over diverse medical datasets to construct rule-verifiable reasoning trajectories. These generated paths are used for policy refinement via supervised fine-tuning (SFT) and direct preference optimization (DPO), and to train a unique Process Reward Model (PRM). This PRM employs a soft dual-sided labeling objective that provides crucial step-level supervision, penalizing reasoning steps that degrade node value to mitigate hallucination and enhance interpretability in complex clinical reasoning. The resulting MedSSS system demonstrated SOTA result across eleven clinical reasoning benchmarks, achieving an average performance gain of +14.12 in comparison to its base model and significantly surpassing 32B-scale general-purpose reasoning models.

The MedReason framework[12] is another significant contribution that addresses the scarcity of high-quality medical reasoning data by leveraging a structured medical knowledge graph (KG). This approach converts conventional question–answer (Q&A) pairs into a dataset of 32,682 factually grounded reasoning trajectories, serving as structural supervision. Fine-tuning models on these paths consistently boosts performance: the framework achieved average accuracy gains of +5.4% for LLaMA 3.1-Instruct-8B and up to 7.7% for DeepSeek-Distill-8B. The resulting MedReason-8B model established a new SOTA among 8B-parameter models, notably surpassing the Huatuo-o1-RL-8B by up to 4.2% on the MedBullets clinical benchmark, validating the importance of KG-based reasoning for interpretability and analytical depth.

The HuatuoGPT-o1[13] framework is a two-stage, verification-guided system designed to instill complex reasoning in medical LLMs using 40K verifiable medical problems. Initially, a verifier guides search strategies to generate verified reasoning trajectories for Supervised Fine-Tuning (SFT), teaching the model to refine its answers. Subsequently, Reinforcement Learning (RL) is applied using verifier-based rewards for further enhancement. This methodology resulted in an average 8.5-point performance gain for the 8B model on medical benchmarks, with the complex Chain-of-Thought (CoT) strategy alone providing an average 4.3-point boost.

B. Previous Studies In Other Domain

The DeepSeek-R1 model [14] demonstrated that applying a reinforcement learning framework, specifically the resource-efficient Group Relative Policy Optimization (GRPO) [15] , to the DeepSeek-V3-Base model can effectively enhance complex reasoning capabilities. While the pure RL training (DeepSeek-R1-Zero) yielded powerful reasoning behaviors, it simultaneously introduced significant side effects, most notably a deterioration in readability and an increase in language mixing. To counteract these issues, the authors implemented a multi-stage curriculum that incorporated a cold-start supervised fine-tuning (SFT) phase. This targeted intervention proved essential for restoring the model's linguistic quality and coherence. The resultant model, DeepSeek-R1, leveraged this carefully balanced training pipeline to achieve reasoning performance comparable to state-of-the-art models such as OpenAI-o1-1217, thereby confirming that the successful enhancement of reasoning through RL requires a subsequent, targeted fine-tuning stage to maintain linguistic integrity and coherence.

A key development in the area of reasoning improvement is Thought Preference Optimization (TPO), proposed by Wu et al .[16] as an iterative Reinforcement Learning from AI Feedback (RLAIF) approach. This preference-based framework compels the model to generate multiple candidate thought and response pairs, subsequently utilizing a judge model to evaluate only the resulting response quality. The collected rejected–preferred pairs are then used to train the backbone model via Direct Preference Optimization (DPO),

[1] All of our work is open-source and available at github.com/Mehrdadghassabi/Gaokerena-R

[2] Available at huggingface.co/gaokerena/gaokerena-r1.0

teaching it to generate high-quality internal thoughts without direct thought supervision. Utilizing an 8B parameter model as the base, this method demonstrated substantial performance gains in general instruction following, achieving an impressive 52.5% win rate on AlpacaEval (LC) and 37.3% on Arena-Hard. These results represent gains of over 4% compared to the direct response baseline, confirming that structured preference optimization effectively enhances reasoning capabilities, even across non-traditional domains such as marketing and health.

The study by N. Ho et al. [17] introduced Fine-tune-CoT, an effective knowledge distillation method that leverages prohibitively large language models (LLMs) to generate Chain-of-Thought (CoT) rationales for the fine-tuning of significantly smaller student models. This systematic transfer process enabled student models, which were approximately 25–100x smaller in parameter count, to acquire substantial reasoning capabilities. For instance, the 6.7B student model achieved an accuracy of 53.33% on MultiArith using diverse reasoning, demonstrating that complex, high-performance reasoning abilities can be efficiently transferred across a significant reduction in model scale.

III. Baseline Model

One of the most influential recent efforts in developing small language models for low-resource languages, such as Persian, is the aya-expanse model , which is available in both 8-billion- and 32-billion-parameter variants. Aya-expanse leverages a strong backbone architecture and relies heavily on synthetically generated data produced by larger language models to compensate for limited real-world training data.

Despite its strengths, a core limitation of aya-expanse is its uncertainty during reasoning-intensive tasks. In particular, under Chain-of-Thought (CoT) prompting, the model often produces different answers for the same question, even when the temperature is set close to zero. This behavior indicates instability in its reasoning process rather than randomness introduced by sampling.

Addressing this limitation by improving reasoning consistency and reliability is therefore essential, and it directly motivates the approach proposed in this work. Although our experiments focus on the medical domain, the proposed framework is model-agnostic and domain-independent, and can be readily extended to other fields that require robust multi-step reasoning.

IV. Proposed Methods

We propose two-stage frameworks to enhance the reasoning capabilities of small Persian medical language models. First, we employ Reinforcement Learning with AI Feedback (RLAIF) to construct a preference dataset $\mathcal{D} = \{(x, y_w, y_l)\}$, where x is a medical question, y_w is a preferred reasoning trajectory, and y_l is a rejected one. Second, we trained the student model π_θ using Direct Preference Optimization (DPO).

To generate high-quality preference pairs, we utilize a teacher-student architecture. The student (π_S) is the baseline Persian model, and the teacher (π_T) is a reasoning-capable model (DeepSeek-R). We employ two strategies to populate $\mathcal{D}$ based on the student's performance:

1. Teacher Correction for Reasoning Alignment: This strategy accounts for 95% of our training data. The student model π_S generates an initial response y_{init} for a given question x. If the response is incorrect, y_{init} is designated as the rejected response (y_l). The teacher model π_T is then prompted with the ground truth answer and instructed to generate a correct, detailed Chain-of-Thought (CoT) explanation. This teacher-generated trajectory serves as the preferred response (y_w). The resulting preference pair (x, y_w, y_l) explicitly contrasts flawed student reasoning with expert teacher reasoning. This approach is depicted in Figure 1.

2. Teacher-Guided Self-Correction: For the remaining 5% of the data, we employ an iterative feedback loop to encourage intrinsic error correction. This limited scale was primarily due to hardware constraints. If the student's initial response y_{init} is incorrect, the teacher generates a textual critique c that identifies the error without revealing the correct option. The student is then prompted with c to self-correct and generate a new response, y_{retry}. If y_{retry} is correct, it is designated as the preferred response (y_w), and the original incorrect response y_{init} remains the rejected response (y_l). This method ensures that the final y_w is generated by the student model's own distribution. The full process is detailed in Figure 2.

We optimize the student policy π_θ to satisfy the generated preferences using Direct Preference Optimization (DPO). DPO is an alignment method that avoids the need for a separate, expensive reward model. Instead, it optimizes the policy directly by minimizing the negative log-likelihood of the preferred response relative to the reference model π_{ref} (the frozen aya-expanse-8b). The objective function is defined as:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}$$

$$\left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

Where σ is the sigmoid function and β is a hyperparameter that controls the extent of the policy's deviation from the reference model. This single objective implicitly maximizes the probability of the valid reasoning trajectory y_w while suppressing the probability of the flawed reasoning path y_l. The final dataset comprised approximately 11,000 pairs, yielding about 2 million preferred tokens and 2.5 million rejected tokens.

V. Data

To implement the proposed methods, a high-quality Persian medical multiple-choice question-answering (MCQA) dataset was required. At the time of this study, no publicly available dataset satisfied these requirements. Consequently, we constructed a new dataset using machine translation to enable systematic evaluation and training in the Persian medical domain.

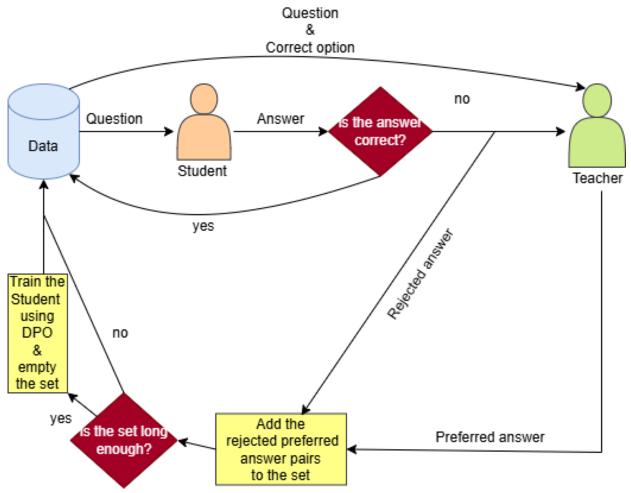

fig. 1.Method 1 Block Diagram

Specifically, we employed DeepSeek-V3[18] , a cost-effective large language model chosen due to strict budget constraints, to translate a randomly selected subset of the MedMCQA dataset [19] from English into Persian. MedMCQA is a large-scale medical MCQA benchmark covering a wide range of clinical and biomedical topics. Questions were randomly sampled to preserve topic diversity and ensure broad coverage across medical specialties.

To ensure the quality and reliability of the translated questions, we adopted a rigorous multi-model evaluation framework. An ensemble of language models was used as referees to assess the correctness, fluency, and semantic fidelity of each translated question and its corresponding answer options. To enforce strict quality control, a translated question was accepted only if all referees independently approved the translation. Due to budget limitations, we selected grok-3-mini [20] and gpt-4.1-mini[21] as referees, balancing evaluation reliability with computational efficiency.

Through this translation and verification pipeline, we obtained approximately 18,000 high-quality Persian medical multiple-choice questions. This dataset forms the foundation for all subsequent experiments and evaluations reported in this work.

VI. Carbon Footprint

The carbon footprint of our DPO fine-tuning process was estimated based on the hardware configuration and total runtime. The procedure ran for a combined total of 1 hour [3] on an NVIDIA H100 PCIe 80 GB GPU, with approximately 43 GB of VRAM utilized during training. The training loss curve is shown in Figure 3. Assuming an average power consumption of 350 watts per GPU, the total energy usage was approximately 0.35 kWh. Using the average carbon intensity of the Canadian electricity grid, where our server was located (0.086 kilograms of CO2 equivalent per kWh[22]), this corresponds to an estimated emission of 0.0301 kilograms of CO2 equivalent generated during the fine-tuning process. Compared to our previous model, gaokerena-V, which emitted 2.66 kilograms of CO2, this represents a substantial reduction in environmental impact[23].

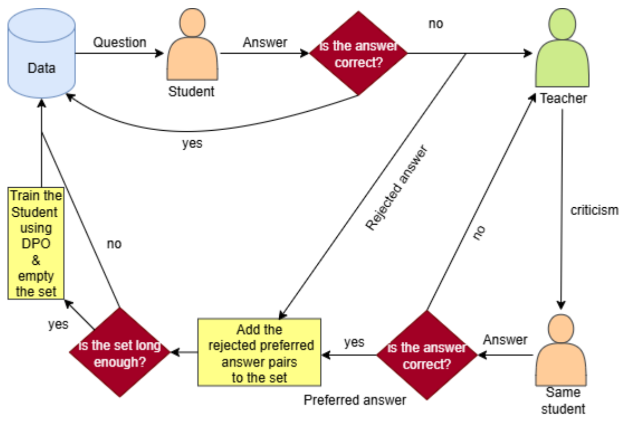

fig. 2.Method 2 Block Diagram

VII. Results

In this section, we compare gaokerena-R, introduced in this work, with gaokerena-V, presented in our previous study, as well as their shared baseline model, aya-expanse-8b. The comparison focuses on two complementary aspects: medical knowledge and medical reasoning capabilities. We define medical knowledge as the extent to which a language model possesses factual and domain-specific information related to medicine. On the other hand, medical reasoning capabilities refer to the model's ability to logically analyze medical problems, integrate multiple pieces of information, and perform multi-step inference to arrive at a correct conclusion. While medical knowledge can often be enhanced through large-scale training on extensive medical corpora, medical reasoning is a more sophisticated capability and does not necessarily improve with increased data scale alone.

To assess these two aspects separately, we evaluate medical knowledge using direct (straight) prompting, which primarily measures factual recall. Medical reasoning capabilities are evaluated using Chain-of-Thought prompting, which encourages explicit step-by-step reasoning.

Our results indicate that, in terms of medical knowledge, gaokerena-V outperforms gaokerena-R and aya-expanse-8b, with gaokerena-R and aya-expanse-8b exhibiting comparable performance. This outcome is expected, as gaokerena-V was trained on approximately 57 million tokens of medical data, whereas gaokerena-R was only post-trained on 11,000 preferred–rejected reasoning pairs.

In contrast, for medical reasoning capabilities, gaokerena-R consistently outperforms both aya-expanse-8b and gaokerena-V, followed by aya-expanse-8b, with gaokerena-V performing the weakest. These results demonstrate that the proposed reasoning-oriented training method substantially enhances the reasoning ability of gaokerena-R. Interestingly, we also observed that large-scale training on raw medical corpora appears to reduce the medical reasoning capabilities

[3] The reported 1 hour refers only to the training time and does not include the time spent on data generation.

of gaokerena-V, which warrants further investigation to understand the underlying causes.

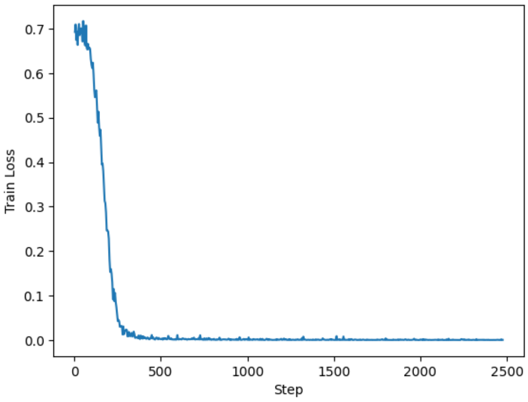

fig. 3.Training loss curve

A. *Medical Reasoning Capabillities*

To evaluate the medical reasoning capabilities of the models, we prompted them to produce COT reasoning paths at a decoding temperature of 1.0.[4] For a comprehensive evaluation, we assessed their performance using two metrics accuracy and pass@k on two datasets: the Persian-translated medical subset of MMLU[5] and the Iranian Basic Medical Sciences Entrance Exam conducted in September 2023[6].

1) Accuracy*:* For each question, the model was run five times, and each run produced a selected answer option. If the same option was chosen in three or more of the five runs, the model was considered confident, and that option was taken as the final prediction. Otherwise, the model was deemed uncertain, and the question was left unanswered. This procedure constitutes the self-consistency strategy used in our evaluation.

The self-consistency strategy provides a robust estimate of the models' reasoning capabilities by leveraging multiple reasoning trajectories. We evaluate model performance under two scoring schemes: with and without negative marking. In the negative marking setting, each incorrectly answered question is assigned a score of -0.33.

Evaluating performance under both schemes is important because different applications impose different requirements: in some cases, providing correct information is paramount, while in others, avoiding incorrect or misleading information is more critical. This trade-off is analogous to the balance between recall and precision. The results under both scoring schemes are reported in Table 1 and Table 2.

Notice that the medical categories shown in these tables correspond to those defined in the MMLU dataset, and we additionally report the average performance across all MMLU categories as a summary measure of overall model performance.

Pass@K: B. Brown et al. *[24]* investigated the variability of responses produced by a language model when given the same prompt multiple times. They argued that, given enough independent attempts, even a random or unskilled agent like a monkey behind a keyboard could occasionally produce the correct answer.like a monkey behind a keyboard could occasionally produce the correct answer. Motivated by this observation, they introduced the pass@k metric, which evaluates a model based on the probability of producing at least one correct answer among k independent samples for the same prompt. This metric provides a robust assessment of models that generate diverse outputs, capturing both their reasoning reliability and the benefits of multiple-sample generation.

Table I. Chain-of-Thought Prompted Performance Without Negative Marking

	gaokerena-R	**gaokerena-V**	**aya expanse-8b** (backbone)
Anatomy	**42.22**	39.25	40.74
Genetics	**50.0**	41.0	45.0
College-medicine	47.97	37.57	**48.55**
Clinical-knowledge	**55.84**	46.79	54.71
Professional-medicine	**44.85**	37.13	43.75
College-biology	**48.61**	36.80	43.75
MMLU(avg)	**48.76**	40.40	47.10
IBMSEE Sept2023	**38.69**	29.76	35.71
Generation time	≈5×35s	≈5×35s	≈5×35s

Table II. Chain-of-Thought Prompted Performance With Negative Marking

	gaokerena-R	**gaokerena-V**	**aya-expanse-8b** (backbone)
Anatomy	**29.13**	27.65	24.93
Genetics	**40.0**	32.0	33.0
College-medicine	**34.68**	25.24	34.48
Clinical-knowledge	**44.65**	35.59	42.51
Professional-medicine	**30.39**	25.0	**30.39**
College-biology	**36.80**	25.0	30.09
MMLU(avg)	**36.14**	28.57	33.55
IBMSEE Sept2023	**24.60**	15.87	19.84
Generation time	≈5×35s	≈5×35s	≈5×35s

The formal definition of this metric is given in Formula 1 , where M denotes the total number of generated samples per question, which is five in our experiments, and C represents the number of times the model selects the correct answer. After computing pass@k for each question in the dataset, we report the average value across all questions.

$$\text{pass@k=1-}\frac{\binom{\text{M-C}}{\text{k}}}{\binom{\text{M}}{\text{k}}} \qquad (1)$$

Due to budget and computational limitations, we ran the model only five times and therefore computed pass@k for $k \leq 3$. Results for the Persian-translated medical subset of the MMLU dataset are presented in Figure 4, while results for the Iranian Basic Medical Sciences Entrance Exam conducted in September 2023 are shown in Figure 5.

[4] Evaluation prompts are available at the GitHub repository

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

[6] Available at huggingface.co/datasets/gaokerena/KOPP

As illustrated in Figures 4 and 5, the gaokerena-R pass@k curve consistently and markedly exceeds those of gaokerena-V and the baseline aya-expanse-8b across nearly all evaluated categories, strongly underscoring the superior reasoning capability of gaokerena-R. This consistent dominance across k values indicates that gaokerena-R is not merely benefiting from chance or diversity in sampling, but is reliably arriving at correct answers through stronger and more stable reasoning processes. The sole exception appears in the Anatomy category, where gaokerena-V performs better, likely due to its training on a rich, anatomy-focused corpus that enhances domain-specific factual recall rather than general reasoning ability. In contrast, gaokerena-V exhibits substantial uncertainty, frequently producing different answers across multiple runs for the same question and, in many cases, selecting all four options over five runs—an instability that is even more pronounced than that of its baseline, aya-expanse. This high uncertainty artificially inflates gaokerena-V's pass@k scores at larger k (e.g., k = 3), especially given the four-option multiple-choice setting, where even a random strategy can achieve pass@4 = 1 by covering all options. Unlike gaokerena-V, gaokerena-R achieves strong performance at low k values, demonstrating that its superiority stems from genuine reasoning strength, consistency, and confidence, rather than from stochastic exploration or chance coverage of answer choices.

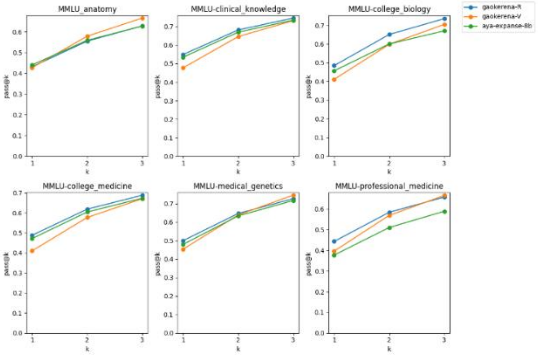

Fig.4.Pass@k results on the FA_MED_MMLU dataset using Chain-of-Thought prompting

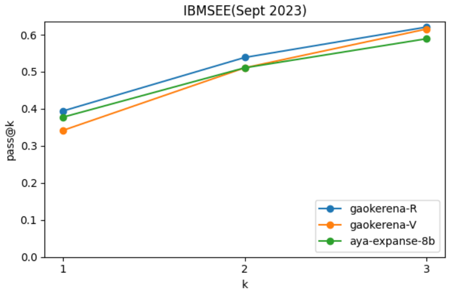

fig .5.Pass@k results on the IBMSEE (September 2023) dataset using Chain-of-Thought prompting

B. *Medical Knowledge*

As noted previously, to evaluate medical knowledge, we used direct (straight) prompting. This type of prompting primarily assesses the model's internal knowledge recall. We applied it to gaokerena-R, gaokerena-V, and aya-expanse-8b on the Persian-translated medical MMLU and IBMSEE datasets. As shown in Table 3, gaokerena-V achieves the best performance, which is expected given that it was trained on 57 million medical tokens, while gaokerena-R and aya-expanse-8b perform comparably. The comparable performance of gaokerena-R and aya-expanse-8b further highlights that the superior performance of gaokerena-R under Chain-of-Thought prompting, as discussed above, is due to its enhanced reasoning capability rather than increased medical knowledge.

Table III. Straight Prompted Performance

	gaokerena-R	**gaokerena-V**	**aya-expanse-8b** (backbone)
Anatomy	41.48	**48.14**	40.74
Genetics	49.0	**53.0**	49.0
College-medicine	**46.24**	43.93	44.51
Clinical-knowledge	52.45	**55.47**	52.07
Professional-medicine	41.91	**47.05**	45.58
College-biology	44.44	**47.22**	45.14
MMLU(avg)	46.28	**49.31**	46.64
IBMSEE Sept2023	35.11	**38.69**	34.52
Generation time	≈10s	≈10s	≈10s

VIII. ADDING A VERIFIER

When prompting aya-expanse-8b (and models that use it as a baseline, including the newly introduced gaokerena-R) with Chain-of-Thought (CoT) prompting, the model often produced multiple distinct answers for the same question, reflecting variability in its reasoning paths. This behavior highlights the inherent stochasticity in language model reasoning when guided to produce step-by-step explanations. To capitalize on this variability and improve prediction reliability, we employed a self-consistency strategy. Specifically, multiple independent CoT outputs were generated for each question, and a majority-voting mechanism was applied to select the answer that appeared most frequently as the final prediction. This approach leverages the diversity of reasoning trajectories to reduce the impact of occasional errors in individual reasoning paths.

In certain instances, however, the model's reasoning paths were extremely diverse; for example, within just five runs, the model sometimes generated all four possible options for a single four-choice multiple-choice question. Such behavior indicates a high level of uncertainty regarding the correct answer[25] . In these cases, the standard self-consistency strategy alone is insufficient, as there is no clear consensus among the generated outputs. To address this limitation, we incorporated a verifier to evaluate the conflicting responses and select the option containing the most accurate or least inconsistent information. This hybrid approach combines the reasoning strengths of the CoT-prompted model with an additional layer of reliability provided by the verifier.

As summarized in Table 4, we conducted experiments on both the Persian-translated medical MMLU and the IBMSEE (September 2023) datasets, generating five independent Chain-of-Thought (CoT) outputs per question. A question was considered uncertain if fewer than three of the five responses agreed on the same answer. For verification, we selected the baseline model, aya-expanse-8b, as it can be accessed by simply unmounting the updated parameters of the trained model, requiring minimal additional storage. This design enables efficient resolution of ambiguous cases without introducing an additional large model, while ensuring that the final predictions reflect both the enhanced reasoning capabilities of the trained model and the broader medical knowledge coverage of the baseline model.

Table IV. *Comparison of gaokerena-R (with verifier) and gaokerena-V*

	gaokerena-R + aya- expanse-8b (verifier)	**gaokerena-V**
Anatomy	47.40	**48.14**
Genetics	**56.0**	53.0
College-medicine	**50.28**	43.93
Clinical-knowledge	**58.86**	55.47
Professional-medicine	**48.89**	47.05
College-biology	**54.86**	47.22
MMLU(avg)	**52.98**	49.31
IBMSEE Sept2023	**46.42**	38.69
Generation time	≈5×35+10+8s	≈10s

IX. Future Research

The results show that gaokerena-V performs better with direct prompts, while gaokerena-R excels with chain-of-thought (CoT) prompts. This indicates that both models are prompt-dependent, and future work must address this critical limitation.

Specifically, future research should aim to develop prompt-invariant medical language models that integrate strong reasoning skills and medical knowledge, achieving superior performance regardless of the prompt format. This involves resolving the inherent trade-off between the superior factual recall of a data-scaled model (like gaokerena-V) and the enhanced logical consistency of a reasoning-focused model (like gaokerena-R). Achieving prompt invariance would represent an important step toward more reliable and generalizable small Persian medical language models.

References

[1] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in neural information processing systems*, 2017.

[2] D. Kahneman, *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011.

[3] Y. Bengio, “From system 1 deep learning to system 2 deep learning,” in *Neural Information Processing Systems*, 2019.

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, and others, “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems*, 2022, pp. 24824–24837.

[5] H. Lee and others, “Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with AI feedback,” *arXiv preprint arXiv:2309.00267*, 2023, [Online]. Available: https://arxiv.org/abs/2309.00267

[6] R. Rafailov and others, “Direct preference optimization: Your language model is secretly a reward model,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 53728–53741, 2023.

[7] M. Ghassabi and others, “Leveraging Online Data to Enhance Medical Knowledge in a Small Persian Language Model,” *arXiv preprint arXiv:2505.16000*, 2025.

[8] J. Dang and others, “Aya expanse: Combining research breakthroughs for a new multilingual frontier,” *arXiv preprint arXiv:2412.04261*, 2024.

[9] “Towards Medical Slow Thinking with Self-Evolved Soft Dual-sided Process Supervision,” *arXiv preprint arXiv:2501.12051*, 2025.

[10] A. Dubey *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024, [Online]. Available: https://arxiv.org/abs/2407.21783

[11] R. Coulom, “Efficient selectivity and backup operators in Monte Carlo tree search,” in *International Conference on Computers and Games*, Springer, 2006, pp. 72–83.

[12] J. Wu and others, “Medreason: Eliciting factual medical reasoning steps in LLMs via knowledge graphs,” *arXiv preprint arXiv:2504.00993*, 2025.

[13] J. Chen and others, “Huatuogpt-o1, Towards medical complex reasoning with LLMs,” *arXiv preprint arXiv:2412.18925*, 2024.

[14] D. Guo and others, “Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025.

[15] Z. Shao and others, “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv preprint arXiv:2402.03300*, 2024.

[16] T. Wu and others, “Thinking LLMs: General instruction following with thought generation,” *arXiv preprint arXiv:2410.10630*, 2024.

[17] N. Ho, L. Schmid, and S.-Y. Yun, “Large language models are reasoning teachers,” *arXiv preprint arXiv:2212.10071*, 2022.

[18] X. Bi and others, “Deepseek LLM: Scaling open-source language models with longtermism,” *arXiv preprint arXiv:2401.02954*, 2024.

[19] A. Pal, L. K. Umapathi, and M. Sankarasubbu, “MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering,” in *Conference on Health, Inference, and Learning*, PMLR, 2022.

[20] xAI, “Grok 3 Beta — The Age of Reasoning Agents.” 2025. [Online]. Available: https://x.ai/news/grok-3

[21] J. Achiam and others, “GPT-4 Technical Report,” *arXiv preprint arXiv:2303.08774*, 2023.

[22] Canadian Climate Institute, “The Big Switch in Canada’s Electricity Emissions.” 2023. [Online]. Available: https://440megatonnes.ca/insight/the-big-switch-in-canada-s-electricity-emissions

[23] A. Lacoste and others, “Quantifying the carbon emissions of machine learning,” *arXiv preprint arXiv:1910.09700*, 2019.

[24] B. Brown and others, “Large language monkeys: Scaling inference compute with repeated sampling,” *arXiv preprint arXiv:2407.21787*, 2024.

[25] M. Zhang and others, “Calibrating the confidence of large language models by eliciting fidelity,” *arXiv preprint arXiv:2404.02655*, 2024.