Developing a medical language model based on reasoning in Persian language

Mehrdad Ghassabi

Bahman 04

Summary

	Gaokerena-V	Gaokerena-R
Github repository	Mehrdadghassabi/Gaokerena-V	Mehrdadghassabi/Gaokerena-R
Model repository	gaokerena/gaokerena-v1.0	gaokerena/gaokerena-r1.0
Arxiv id	2505.16000	2510.20059
Cost	$300	$70
Reasoning capability	No	Yes
Published in	ICBME2025	ICSPIS2025

Motivation

- AI and language models are transforming healthcare
- English models already provide expert-level medical Q&A
- But Persian still lacks open, reliable medical language models
- Building such models is crucial for privacy-friendly local deployment

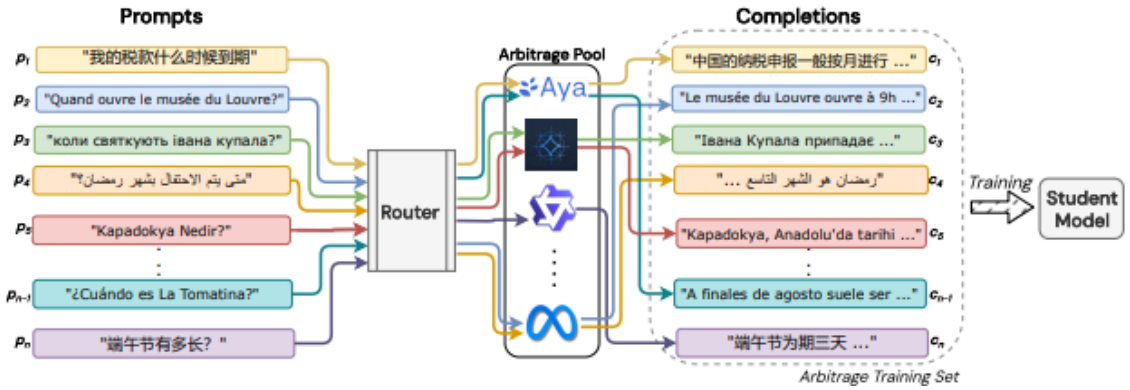
Problem Statement

- No Open Persian medical language models existed before
- No Open Persian medical corpus, QA dataset and benchmark existed before
- Facing almost a white paper on this topic

Phase 1

Collecting Data

- Crawled 100k articles from medical magazines
- Crawled 180k QA pairs from medical forums
- Crawling Codes were running about two weeks
- Cleaning Process has been done after crawling
- Created benchmark by translating from english



Baseline Model

- Choosing a general purpose language model that support Persian
- Almost all baseline models had problems understanding Persian
- Aya-expanse was an exception since it had been developed to be

Training

- Training on collected data lasted for 20 hours using a A100 GPU
- About 3200 question should be asked for testing using a L4 gpu
- Each question answered in about 10s
- In colab plans, A100 costs about $1 per hour and L4 costs about $0.5 per hour
- Therefore Each train and test iteration costed about $25 for us
- We found the best hyperparameters in 6 iterations

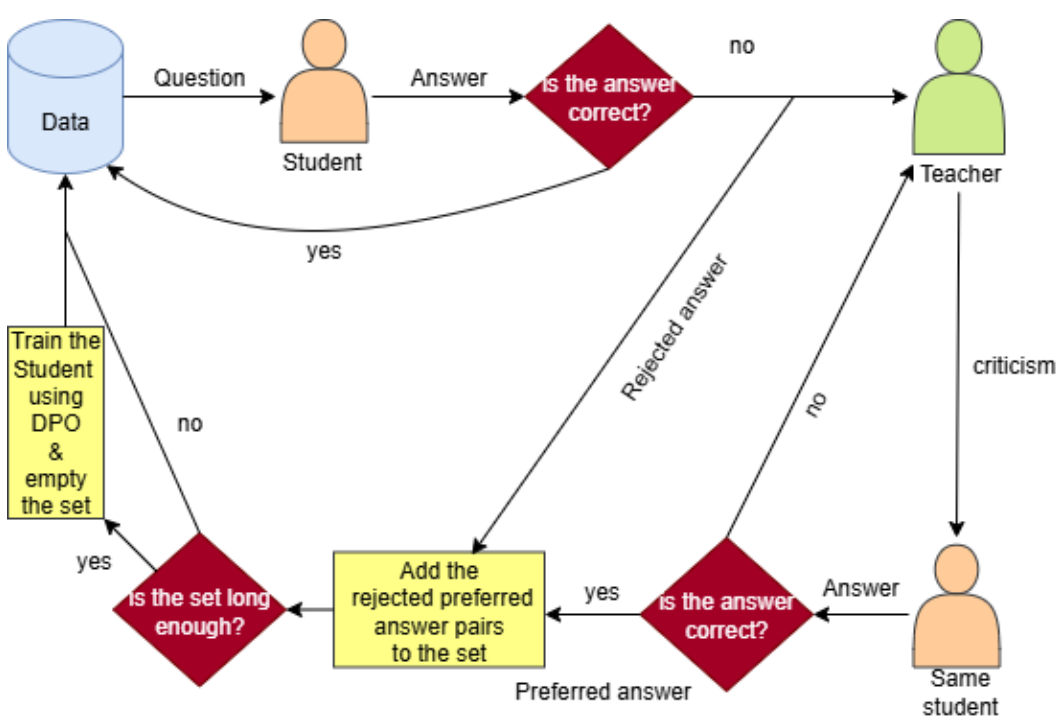
Results

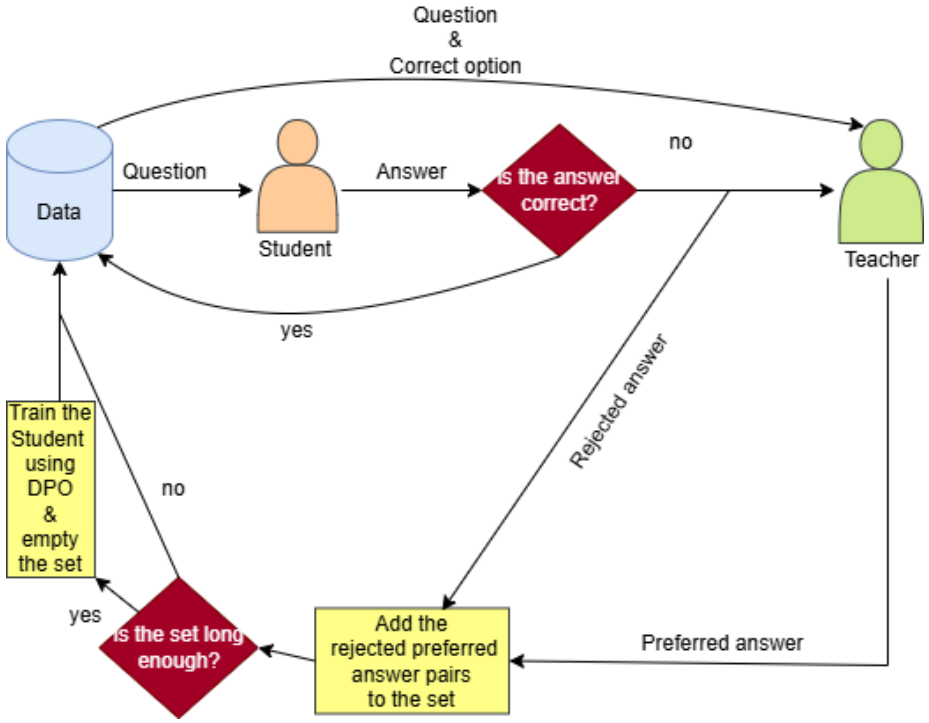
	gaokerena-V	aya-expanse	Qwen	PersianMind
Anatomy	48.14	40.74	41.48	25.18
Genetics	53.0	49.0	52.0	34.0
College-Medicine	43.93	44.51	43.35	20.23
Clinical-Knowledge	55.47	52.07	47.92	25.28
Professional-Medicine	47.05	45.58	43.01	23.89
College-Biology	47.22	45.14	42.36	32.63
Avg	49.31	46.64	45.17	25.89
IBMSEE Sept 2025	38.69	34.52	33.33	19.64
No. parameters	8b	8b	7.6b	6.8b

Phase 2

Reasoning in language models

- In NeurIPS 2019 Bengio asserted a key deficiency in current deep learning systems, they can't reason
- That's why Sutton bitter lesson repeats
- That deficiency exists in transformers too, a new artchitecture should be developed
- But as Feb 2026 we are bounded to use transformers and try to enhace its reasoning skills



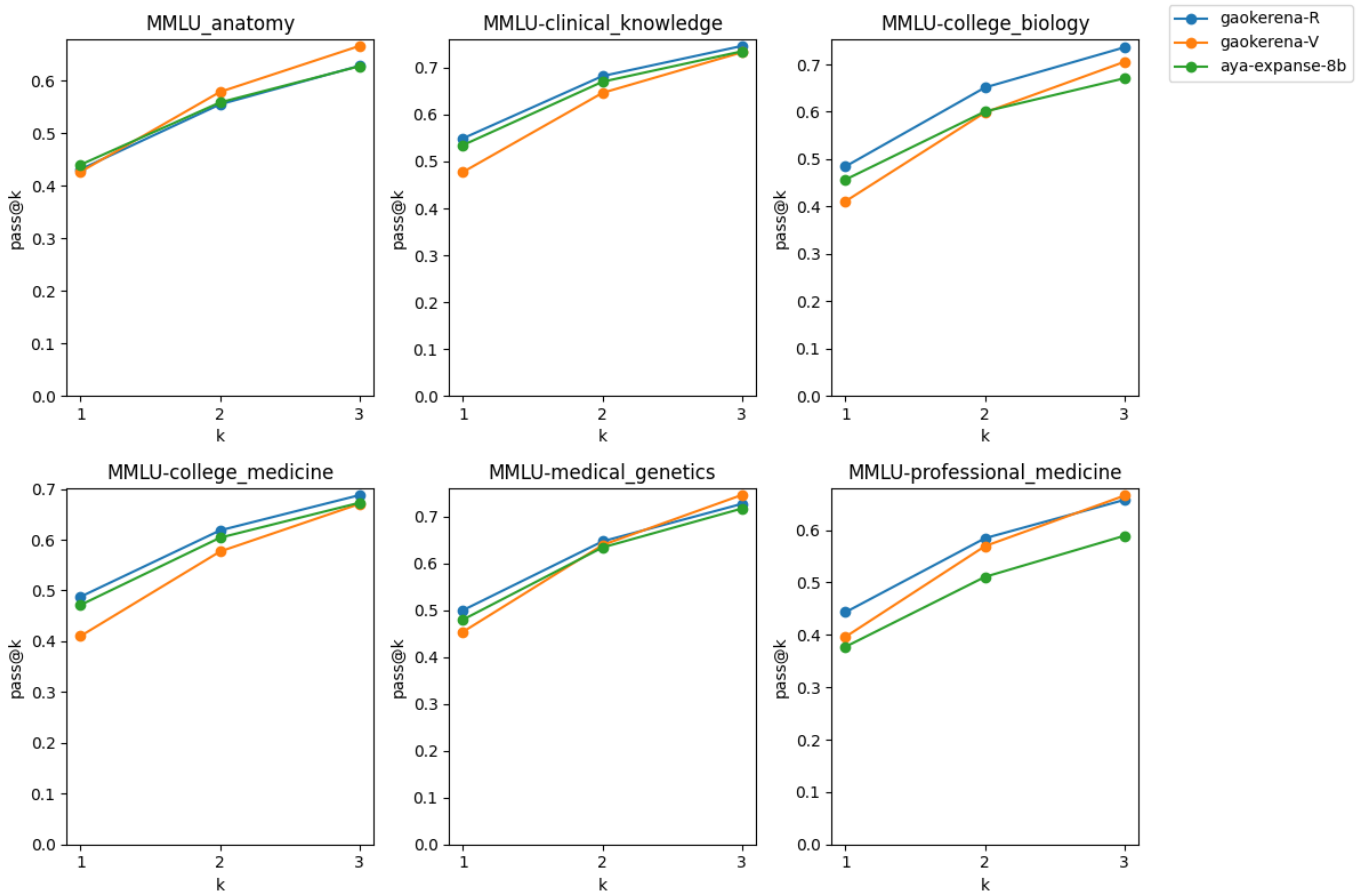


Proposed methods

Advantageous

- Without collecting data and just using RLAIF we got better Result
- Only a Single hour has been used for training
- Our experiments shows the importance of reasoning in medical domain

Results


MMLU_anatomy
MMLU-clinical_knowledge
MMLU-college_biology
MMLU-college_medicine
MMLU-medical_genetics
MMLU-professional_medicine
pass@k
k
gaokerena-R
gaokerena-V
aya-expanse-8b

Acknowledgement

- My father
- Dr. Baradaran
- Sadra Hakim
- Pedram Rostami
- Zahra Kazemi
- Milad Tavakoli
- Amirhossein Poursina