Enhancing Medical Data Prediction with Chain-of-Thought Reasoning

Vikrant Bhati, Ishani Kohli , Eeshan Umrani , Darshan Nere , Sarvesh Chakradeo
vikrant@vt.edu, ishanik02@vt.edu, eeshanumrani@vt.edu, darshannere@vt.edu, sarveshdc@vt.edu

1 Problem statement

Large Language Models, or LLMs, have shown high rate of success at tasks involving natural language processing, but they struggle with many aspects of precision and reasoning in crucial fields like medicine. This distinction is crucial in situations where accuracy and reasoning could mean the difference between life and death. LLMs find it difficult to understand the complexity of reasoning needed for patient data analysis and medical diagnoses, which involve numerical computations, logistical inference, and context-sensitive decision making where even small mistakes can have serious effects on patient safety.

In practice, LLMs tend to lose context for what inferences can go with particular information and how to link a series of logical steps over time, especially in the event of a multi-turn medical dialogue scenario, and more complex diagnostic challenges. Also, non-COT LLMs all work, in essence, as black-boxed computational systems that display several answers without explaining the reasoning behind them. Because of this lack of sequential reasoning, users become less confident in the model's decision-making process, and it becomes more challenging to confirm that its assumptions or logic are sound, particularly in challenging or high-stakes tasks, and it also prevents or complicates any verification of their conclusions by a healthcare professional.

Our project addresses these challenges by developing an reasoning LLM in size of 3 Billion parameters that is specifically targeted for medical applications, with an emphasis on improving its logical reasoning, organization in context, and transparency and easier to use.

We suggest using Chain-of-Thought (CoT) prompting in conjunction with Group Relative Policy Optimization (GRPO) to address these issues. In order to promote clear and understandable decision-making processes, CoT prompting specifically encourages LLMs to break down complicated issues into intermediate logical steps. Using structured rewards that assess logical coherence, keyword relevance, and stepwise efficiency, this reinforcement learning technique enables models to iteratively improve their reasoning processes and maximize performance. The main goals are the following:

- Implementing Chain-of-Thought (CoT) Reasoning - This enables the model to decompose complex medical tasks (e.g., differential diagnosis) into precise logical steps.

- Improving Context Management - Giving LLM the relevant background information on a patient's symptoms, history, and test results throughout multi-turn conversations so that they can make the right diagnosis

- Enhancing Transparency and Trust - Helping medical professionals to validate the model's logic and conclusion and giving them concise, understandable justification for each recommendation.

2 What you proposed vs. what you accomplished

2.1 Objectives Accomplished

- We have successfully implemented CoT prompting and the model is now able to deconstruct complicated medical tasks into logical, sequential steps.
- We allowed the model to learn and optimize its reasoning process through reinforcement learning by using GRPO, which consistently improved diagnostic performance.
- We got a feedback on our proposal to incorporate an LLM-as-Judge in our project for evaluation and we have been successful in doing that and also incorporating other evaluation methods like manual evaluation and perplexity score to compare results and measure the clarity and interpretability of our model.
- We have also successfully tested our model's scalability using different datasets such as medical-o1-reasoning-SFT, med-qa and big-bio-med-qa

2.2 Objectives Unaccomplished

- We wanted to perform a comparative study of our approach on different classes of models like chat models vs reasoning models or small models vs large models but we had to prioritize optimizing and thoroughly evaluating our primary model on different datasets because of time constraints and the computational resources needed for training and evaluating multiple large-scale models.

3 Related work

Developing medical reasoning models with a chain-of-thought process is an emerging area of research in NLP. Several prior works have explored related challenges, including medical question answering, reasoning-based dialogue systems, and reinforcement learning approaches for medical decision-making. In the following, we discuss some relevant studies and their connections to our project.

One notable work is (Wei et al., 2022), which explores the use of chain-of-thought prompting in large-language models to enhance reasoning capabilities. Their results demonstrate that structured reasoning paths improve model interpretability and decision-making accuracy in complex tasks. Our project extends this idea by leveraging GRPO (Guided Reinforcement Policy Optimization) to fine-tune a model specifically for medical reasoning.

(Gramopadhye et al., 2024) investigate the effectiveness of chain-of-thought-driven reasoning for open-ended medical question answering. They propose a modified version of the MedQA-USMLE dataset to mimic real-life clinical scenarios and explore language model-driven forward reasoning for correct responses to medical questions. However, their work primarily focuses on prompt engineering rather than reinforcement learning, which our project aims to address through GRPO training.

Another relevant study, (Wu et al., 2023), examines the application of chain-of-thought prompting to medical diagnostic reasoning. They introduce Diagnostic-Reasoning CoT (DR-CoT) to improve diagnostic accuracy by prompting large language models with exemplars that mimic doctors' reasoning processes. While their work focuses on enhancing diagnostic reasoning through prompting, our project aims to advance the reasoning capabilities of such models through reinforcement learning-based fine-tuning.

Additionally, (Wang et al., 2024b) propose JMLR (Joint Medical LLM and Retrieval Training) to enhance reasoning and professional question-answering capabilities in medical language models. Their approach integrates retrieval-augmented generation with joint training to reduce hallucinations and improve factual accuracy. We plan to build upon their insights by applying GRPO training to enhance reasoning consistency.

(Chang, 2023) explores the application of the Socratic method to AI, fostering crit-

ical reading and thinking. He introduces SocraSynth, a framework that convenes multiple large language models in a collaborative and adversarial dialogue to uncover knowledge and insights previously inaccessible to human understanding. Our project expands upon this by incorporating chain-of-thought generation within an RL-driven fine-tuning process to ensure both accuracy and interpretability.

(Wang et al., 2024a), which presents DiReCT, a diagnostic reasoning framework designed to improve LLMs' diagnostic capabilities by utilizing a sizable, annotated clinical note dataset. Their methodology draws attention to the discrepancy between human diagnostic abilities and LLM performance. However, the fact that their approach depends on static, annotated data for training limits it. In order to overcome this limitation, our project uses Guided Reinforcement Policy Optimization (GRPO), which allows the model to continuously learn and optimize its reasoning.

Another related study is (Sukhwal et al., 2024), which introduces a disease Question and Answer system based on joint reasoning. This method enhances disease question answering by combining knowledge graphs and language models. Its reliance on external knowledge graphs, however, restricts its scalability.

In (Jeong and Sohn, 2024), the authors evaluate LLM-generated diagnostic reasoning across multimodal medical datasets, including text and images. While their model demonstrates strong performance in multimodal scenarios, it lacks dynamic learning capabilities. Our project, which focuses on text-based diagnostic reasoning, enhances model accuracy through GRPO, enabling it to optimize its reasoning over time.

The study by the Nature Editorial Board (Board, 2024) investigates diagnostic reasoning exercises to improve the clinical decision-making abilities of LLMs. This method's reliance on manually created templates limits it even though structured prompts increase accuracy. By incorporating GRPO, our project expands on this by enabling the model to learn and optimize its reasoning dynamically without the need for human intervention. Finally, (Singhal et al., 2022) identifies gaps in LLMs' logical reasoning skills by examining their capacity to reason about complicated medical issues. By utilizing Chain-of-Thought (CoT) reasoning in conjunction with GRPO, our project directly fills these gaps and guarantees not only accurate reasoning but also adaptive learning and decision-making transparency.

By integrating insights from these prior works, our project aims to bridge the gap between struc- tured medical reasoning and reinforcement learn- ing optimization. Most of these works rely on CoT prompting but our approach leverages CoT with Reinforcement Learning.

Apart from this, we have achieved these results on a smaller model as compared to related work in this field which makes it suitable for a variety of healthcare settings because it guarantees quicker inference speeds, reduced computational costs, and simpler deployment on low-resource devices. The use of GRPO training for medical datasets represents a novel approach, promising improved model performance in clinical decision-making scenarios.

4 Datasets Used

In our study, we conducted comprehensive experiments using three different datasets to train and evaluate our model's performance. Specifically, we utilized:

- The medical-o1-reasoning-SFT - The dataset is taken from Hugging Face's FreedomIntelligence dataset and comprises 90,120 triples of clinical questions with long, multi-step reasoning chains and abridged answers. On disk, it takes about 247 MB in the form of raw JSON, or 144 MB when stored in Parquet, and in total contains over 15 million tokens for even just the English "train" split alone. Each item is composed of an open-text "Question" of a patient case or lab result, a

"Complex_CoT" field with step-by-step chain-of-thought reasoning (usually in the range of 500 to 1,200 tokens), and a short "Response" field—typically fewer than ten tokens—indicating the diagnosis or recommendation.

- BigBio-Med-QA - We used this dataset to compare our model performance, and this dataset was taken from Hugging Face with a vast question-answering dataset covering a wide range of medical topics, such as diagnosis, treatment, drug interactions, and medical definitions. In this dataset, each section includes a medical question, a correct answer, and optionally a directive prompt, which makes it apt for evaluating the model's capacity for comprehending and producing correct medical answers.

- PubMedQA—We used this dataset to compare our model performance with other performance matrices. This dataset was taken from Hugging Face, which has a vast question-answering dataset. Questions are drawn from real research papers, with an emphasis on factual, evidence-based responses. It offers a detailed examination of the capacity of the model to comprehend and produce correct answers from biomedical text, hence improving its factual coherence.

4.1 Data Preprocessing

During training, we split the dataset into training and testing data where we used 80 percent of the data for training purposes and the remaining 20 percent of the data for testing purposes. This ensured that we could test the model's performance on data it had not seen before.

We used the same 80-20 split for training and testing across all datasets in the project to have a consistent split across all datasets we were using. This held irrespective of which datasets were being used and ensured that the performances of the model could be compared among other medical question types and levels of complexity.

Through this standardized data split, we ensured that training was achieved in a standardized and comparable way which limited the potential for data leakage and overfitting balances and allows for an accurate measurement of the model's generalization capabilities.

We performed light data processing our medical reasoning dataset, as the material was already cleaned, well-structured, and free of obvious noise, so we preserved its original form to maintain fidelity to the authentic prompt–response format. We also confirmed that every example fit within our 512-token context window and contained only expected characters and symbols. By relying on the dataset's inherent quality and minimizing interventions, we ensured that our modeling results reflect the data as it was originally intended.

4.2 Data annotation

Since we employed an existing corpus of medical reasoning examples with gold-standard chains of thought and answers, no new annotation effort was undertaken. There was no pilot study or interannotator adjudication, and consequently no need to compute agreement metrics. This decision allowed us to dedicate our resources entirely to model development and evaluation rather than to the logistics of generating or validating annotations.

5 Baselines

In our experiments, we started with Qwen2.5-3B-Instruct as the backbone for all baselines, since it strikes a nice balance between size and capability and comes with an official walk-through on its strengths and limitations. The first thing we tried was zero-shot chain-of-thought prompting: we simply prefaced each medical question with "Please think step by step" and let the model loose. No examples, no weight updates just plain prompting. This gave us a clear view of what the model could do "out of the box" when encouraged to explain its reasoning. On our held-out test set, accuracy hovered around 35%, which told

us that while Qwen2.5-3B could sometimes piece together a correct answer, it often got stuck without concrete examples to guide it.

Next, we moved on to few-shot CoT prompting. Here, we added five hand-picked question-answer demonstrations, complete with their chain-of-thought traces, right in the prompt before each new query. The idea was to let Qwen see exactly how a good solution should unfold, and it did help: test accuracy climbed to about 42%. It wasn't a dramatic leap, but it confirmed that in-context examples can nudge the model toward clearer, more consistent reasoning without touching the model's weights.

Finally, we set up a supervised fine-tuning baseline. Using an SFT dataset tailored for medical reasoning, we re-trained Qwen2.5-3B to produce both the reasoning steps and final answer in one go. This gave us a solid benchmark for a pure weight-update approach, yielding around 56% on direct answer accuracy. Since our main contribution builds on reinforcement learning, having this SFT result helped us measure the incremental gains more fairly.

6 Approach

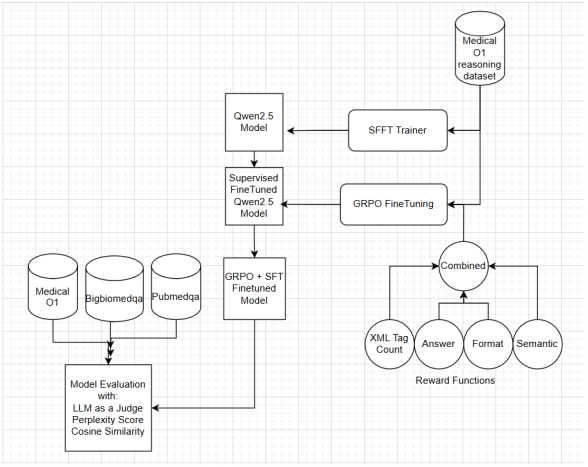

Figure 1: Approach

6.1 Base Model

In this research, we based our training on a small-scale language model called Qwen2.5-3B. This model has about 3 billion parameters and can be used for multiple contexts, including mathematical reasoning tasks. The reason for using this model is its small structure and fast implementation using LoRA adapters. We used LoRA 4-bit adapters, which allowed us to work on a single GPU. These adapters helped us fine-tune only small parts (about 5-10% of the chosen parameters) rather than changing the entire parameter set.

6.2 Dataset Formatting

To train and evaluate the Qwen model effectively, we ensured that the dataset, FreedomIntelligence/medical-o1-reasoning-SFT, was organized into a consistent and structured format. This structured approach was crucial for both Supervised Fine-Tuning (SFT) and Guided Reinforcement Policy Optimization (GRPO) as it allowed us to evaluate the model's outputs based on two primary aspects: structural coherence and accuracy of the answers. To achieve this, we required the model's outputs to adhere to a predefined format. This format was enforced by attaching a structural guide to the prompt, showing the model how to answer correctly. Specifically, we used the XML COT FORMAT function, which divides the output into two distinct sections:

Reasoning: This contains the chain-of-thought (CoT) reasoning, which demonstrates the logical steps leading to the answer.

Final Answer: This is the outcome derived from the reasoning process.

To ensure that these sections were extracted correctly during training and evaluation, we implemented the extract xml segment function. This function uses a pattern-matching approach to search for specific XML tags (e.g., ¡reasoning¿ and ¡answer¿) within the data. It isolates and retrieves the text enclosed within these tags, enabling the system to differentiate between reasoning and the final answer. After extracting the XML content, the extract reasoning and extract answer functions further process the data to extract the reasoning and answer sections explicitly.

6.3 SFT Training

To start our fine-tuning of the Qwen model, we initially began with the Supervised Fine Tuning (SFT) of the model. We used the SFTTrainer class from the trl (Transformers Reinforcement Learning) library. We utilized the unsloth library to optimize the precision training, and we applied LoRA adapters to fine-tune only a small part of the parameters. For the prompt, we formatted the dataset into three parts: the question, the complex CoT (i.e., reasoning), and the final answer. Then we merged all this to get a text value, along with an EOS token. We passed this dataset as our training dataset and set the dataset text field to the merged input for our training model. Finally, we used SFT Trainer on our model for about 80% of the data.

6.4 Group Relative Policy Optimization

We use Group Relative Policy Optimization (GRPO) which offers a substantial enhancement over conventional reinforcement learning algorithms like Proximal Policy Optimization (PPO) by removing the need for a resource-intensive critic network. This approach significantly lowers both memory requirements and training expenses. GRPO accomplishes this by utilizing a group-based advantage estimation for each input, dramatically cutting down the computational overhead.

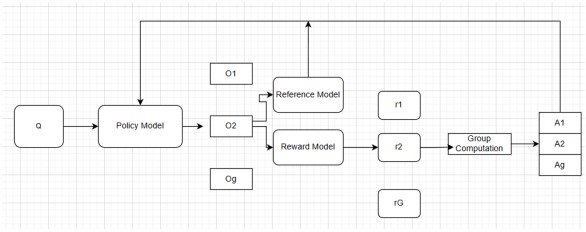

Figure 2: GRPO Architecture

To train the Baseline model further with Group Relative Policy Optimization (GRPO), we built upon the SFT-trained dataset. We utilized the FastLanguageModel from the unsloth library to efficiently manage precision training and optimize GPU memory. The most important aspect here is the use of reward functions in our GRPO model. While SFT already provides sufficient context about the dataset to the model, the rewards in GRPO encourage the model to think in terms of a CoT (Chain of Thought) and improve its quality. These rewards are what make changes to the RL algorithm's policy, enabling better predictions in terms of CoT and the final answer. The rewards we used were:

Reward Functions	Total Weightage
Sematic Similarity	42%
Format Compliance	15%
Answer Matching	29%
XML Count	15%

Figure 3: Reward Functions

Semantic Similarity: We used the Sentence Transformer *all-MiniLM-L6-v2* model. It is an efficient and small pre-trained model. We used it to calculate the cosine similarity between the Chain-of-Thought (CoT) reasoning in the dataset and the CoT produced by our model. The *all-MiniLM-L6-v2* model is quick and efficient at a number of tasks that calculate how similar texts are. The CoT text of the dataset and the generated CoT text were converted into dense vector representations (embeddings) by the Sentence Transformer. Cosine similarity was then computed between the embeddings to provide a quantifiable means to determine how close in meaning the two texts are. A larger cosine similarity score indicates that the generated CoT is closer in meaning to the reference CoT, with improved reasoning quality and content. This reward mechanism explicitly encourages the model to generate CoT explanations that are not only factually accurate but also logically consistent with the input question. In this way, the model is trained to provide correct answers and explain the reasoning process clearly.

Format Compliance: In this reward, we assigned a score of 0.5 to the model if it produced outputs in the correct format using ¡reasoning¿...¡/reasoning¿ and ¡an-

swer¿...¡/answer¿ tags. This reward ensures structural correctness and aids in evaluation.

Answer Matching: This metric is solely based on comparing the final answer with the original answer. A reward of 1 is assigned if the generated answer is correct.

XML Count: Since the GRPO output is expected in XML format with reasoning and answer tags, this is an important reward for structural coherence. It measures a weighted score based on the structure and tag count in the XML.

Combined Reward: Finally, all these rewards are combined using specific weights to produce a single combined reward. The combined reward is calculated as:

Combined Reward = 0.6 × Semantic Similarity + 0.2 × Format Compliance + 0.4 × Answer Matching + 0.2 × XML Count

6.5 Model and GPU Details

We use the Qwen-2.5-3B model(3 billion parameters) as our model. It's part of the Qwen2.5 series developed by Alibaba, which includes models ranging from 0.5 to 72 billion parameters. This particular model has approximately 3.09 billion parameters and is designed for tasks that require following instructions, such as chat-based interactions.

We used collab powered T4 GPU's and a RTX A5000 GPU from on Demand GPU provider to train and test our models.

6.6 Results and Findings

In this section, we present a comprehensive comparison between the baseline LLM (non-CoT, non-GRPO) and our GRPO-enhanced CoT model across three biomedical QA datasets. We evaluate on the Base Test Dataset, BioMedQA, and PubMedQA using three complementary metrics: (1) automated accuracy as judged by an external LLM ("LLM-as-Judge," Gemini 2.0 Flash), (2) perplexity, and (3) human (manual) evaluation of answer correctness and reasoning clarity.

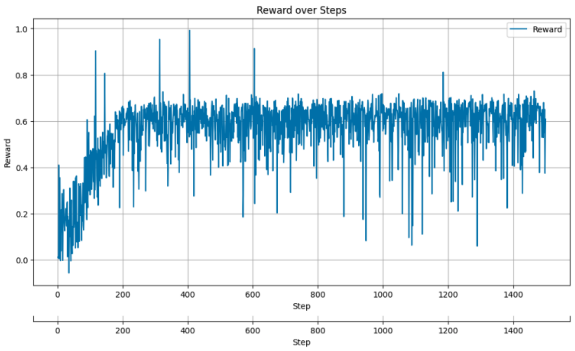

Figure 4: GRPO Rewards Training

6.7 Quantitative Evaluation

Table I summarizes the automated accuracy scores and perplexity for both models on each dataset. Accuracy is measured as the percentage of questions for which the model's final answer matches the reference answer, as judged by Gemini 2.0 Flash. Perplexity is computed over the test set to assess model fluency and confidence in its generated tokens.

Dataset	Model	Accuracy (%)
Base Test Dataset	Baseline	56
	Baseline + GRPO	70
BioMedQA	Baseline	52
	Baseline + GRPO	56.4
PubMedQA	Baseline	47
	Baseline + GRPO	56.2

Figure 5: Human Evaluation Results

Across all three datasets, the GRPO-CoT model outperforms the baseline by an average of +8.9 percentage points in accuracy, while reducing perplexity by an average of 4.6. This demonstrates both more correct answers and greater confidence/coherence in its language modeling.

6.8 LLM-as-Judge Evaluation

We have used Gemini 2.0 Flash as an independent evaluator to help impartially evaluate the correctness of the model responses to medical inquiries and clinical scenarios across our three different datasets, to eliminate set biases in evaluating the correctness of answers with no human involvement. Gemini 2.0 Flash is an excellent Large Language Model (LLM) with superior natural language understanding

and reasoning and was suitable option for this role.

Dataset	Model	Accuracy (%)
Base Test Dataset	Baseline	52
	Baseline + GRPO	61
BioMedQA	Baseline	46.4
	Baseline + GRPO	58.3
PubMedQA	Baseline	43
	Baseline + GRPO	57.2

Figure 6: Evaluation Results using LLM as Judge

The evaluation with the use of Gemini 2.0 Flash was conducted to enable an analysis that consisted of more than simply matching the syntactic format of answers, but instead was meant to evaluate the responses not simply as for a perfect match, but for the logical thought, medical correctness, and the reasoning for reaching that solution.

7 Error analysis

In many questions, the SFT baseline model considers only a certain number of initial input points to give out response, and sometimes misses or even wrongfully guesses symptoms for results. This obviously turns out to be erronous as the model needs to understand the complete problem, and build context around the problem using CoT to derive an answer. As the GRPO rewards and methods encourage the model to focus on providing a reasoning, and also makes it focus on the quality of CoT, we get better results due to context building.

In addition to this, as our model of a relatively small (3B parameters), struggles low contextual memory and missed critical intermediate steps of findings the response. This limitation is particularly pronounced when processing multi-step reasoning tasks, where the depth and coherence of the generated chain of thought are essential for precise outputs. By integrating GRPO-based reinforcement and CoT prompting, these smaller models can be significantly enhanced, leveraging structured reasoning pathways to overcome inherent capacity constraints and produce more contextually aligned and accurate responses.

8 Contributions of group members

We all made equal contributions to the project's writing, background research. Apart from that, each member's implementation responsibilities are detailed below

- Ishani: Collected and processed the relevant medical data, making sure that it was formatted, cleaned, and ready for model training. She then created the GRPO reward function, carefully crafting it to efficiently direct the model's learning process by allocating suitable rewards according to the precision and logical consistency of the diagnostic reasoning of the model. This reward function was essential to maximizing the model's capacity for making decisions.

- Darshan: Implemented, and trained the Supervised Fine-Tuning (SFT) model, which served as the baseline model for our project. Darshan undertook identifying a proper model architecture and optimizing the model via supervised learning, thus providing a starting point for further optimization.

- Vikrant: the development and training of the Group Relative Policy Optimization (GRPO) model, which included the implementation of the GRPO training framework, setting the reward function, and making sure that the model was learning through its interactions. By utilizing the GRPO, Vikrant allowed the model to constantly optimize its diagnostic reasoning skills making the model adaptive and always improving as it received feedback.

- Eeshan: Worked on running a evaluation of the model's performance across a number of medical datasets: bigbio/med-qa, pubmed-qa, and medical-o1-reasoning-SFT. Eeshan was responsible for establishing the evaluation, determining performance

metrics, and conducting statistical analysis on the performance variables to identify trends and deficiencies in the model's performance. Eeshan ensured that the evaluation was thorough, fair and consistent across all datasets.

- Sarvesh: Conducted a thorough evaluation process using numerous measurements to assess the performance of the models. This included an LLM-as-Judge (Gemini 2.0 Flash) for unbiased semantic evaluation and human-level understanding via manual review. Sarvesh tabulated the evaluation with thoughtful clarity and organization to demonstrate the diagnostic accuracy of the models, reasoning quality, and recommendations for improvement. This extensive assessment was the basis for the final analysis and conclusions of our project.

9 Conclusion

In conclusion, our experiments demonstrate that incorporating a chain of thought significantly enhances model performance, with GRPO proving to be an effective method for structuring a reward system that prioritizes better reasoning. By optimizing for thoughtful and structured responses, the model was able to achieve improved accuracy across multiple datasets within the same domain. Notably, we observed a significant accuracy improvement of 11% compared to the SFT baseline model, as evaluated using an LLM judge. These findings highlight the potential of GRPO in advancing the reasoning capabilities of language models.

Exploring adaptive reward weighting and deploying the model in real-world decision-support settings will be essential next steps. Overall, this work lays a foundation for transparent, stepwise reasoning in medical LLMs and points toward scalable strategies for deploying trustworthy AI in healthcare.

10 AI Disclosure (this section does not count toward the minimum 8-page requirement)

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - No

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - No
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - No

References

Nature Editorial Board. 2024. Diagnostic reasoning prompts reveal the potential for large language models in clinical reasoning. *Nature Digital Medicine*, 7(10):1234–1245.

Angel Chang. 2023. Socrasynth: Socratic method for ai critical thinking. In *NeurIPS Workshop on AI and Critical Thinking*.

Saurabh Gramopadhye et al. 2024. Chain-of-thought-driven reasoning for open-ended medical question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

H. Jeong and Y. Sohn. 2024. Evaluating llm-generated multimodal diagnosis from medical mcqs. *arXiv preprint arXiv:2402.01730*.

S. Singhal, A. Patel, and M. Zhuang. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.

R. Sukhwal, K. Kumar, and A. Singh. 2024. A joint-reasoning based disease q&a system. *arXiv preprint arXiv:2401.03181*.

V. Wang, J. Lee, and S. Chen. 2024a. Direct: Diagnostic reasoning for clinical notes via large language models. *arXiv preprint arXiv:2408.01933*.

Zhiheng Wang et al. 2024b. Jmlr: Joint medical llm and retrieval training for enhanced reasoning. *Journal of Medical AI*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Yuhao Wu et al. 2023. Diagnostic-reasoning chain-of-thought for medical question answering. *Nature Medicine*.