MediGANdist: Improving Smaller Model's Medical Reasoning via GAN-inspired Distillation

Stanford CS224N Custom Project

Allen Chau and Aryan Siddiqui
Department of Computer Science
Stanford University
achau774@stanford.edu and aryansid@stanford.edu

Abstract

The integration of artificial intelligence (AI) in healthcare, particularly through conversational agents, presents a transformative potential to augment medical education and patient care. Despite the remarkable capabilities of Large Language Models (LLMs) like GPT-4 in natural language processing, their application in healthcare is constrained by significant challenges, including data privacy concerns and the computational intensity required for local deployment. This study introduces a novel chatbot, designed to operate locally on personal devices, offering immediate access to medical information while ensuring enhanced data security. By fine-tuning a T-5 model using a Generative Adversarial Network (GAN)-inspired approach, we outperform state-of-the-art distillation techniques, addressing their limitations in capturing the complexities of medical reasoning. Our model demonstrates comparable medical reasoning capabilities to its larger counterparts, validated through its performance on the United States Medical Licensing Examination (USMLE). This work contributes to the growing field of AI in healthcare by demonstrating that it is feasible to achieve high-level medical reasoning with smaller models, enhancing accessibility without compromising on privacy or computational efficiency. Our findings underscore the potential of tailored AI systems to revolutionize medical education and patient support, paving the way for further research in AI-assisted healthcare solutions.

1 Key Information to include

- Mentor: David Lim
- External Collaborators (if you have any): N/A
- Sharing project: N/A
- Contributions: Aryan worked on GAN approach and training T-5. Allen worked on standard distillation and also RLHF, which we later did not implement. Both worked on writeup.

2 Introduction

The advent of Large Language Models (LLMs) such as GPT-4 [2] has catalyzed transformative changes across various domains, particularly in healthcare. Despite the capabilities of closed-source models like GPT, their application in medical contexts is constrained by privacy concerns and the inherent limitations in handling sensitive medical data. Consequently, there has been a growing interest in open-source alternatives like PMC-LLaMa [11], Med-Alpaca [4], and Chat-Doctor [7], designed to offer specialized solutions for medical question answering. Despite their advancements, the deployment of such LLMs in practical settings is limited by their computational requirements, hindering the development of local, on-device applications that are critical for real-time, private healthcare assistance.

In light of these considerations, our research aims to create a local heath chatbot. By fine-tuning a more compact T-5 model, we aim to replicate the medical reasoning capabilities of its larger counterparts without the large overheads. During our exploration, we found that traditional techniques of model fine-tuning and distillation inadequately capture the intricate nuances of medical reasoning. To bridge this gap, we introduce an innovative GAN-inspired method for model distillation, specifically designed to overcome the intricacies of medical reasoning tasks. [3]

This work is predicated on two foundational objectives: first, to substantiate the capability of scaled-down models to achieve medical reasoning performance on par with much larger LLMs. We will show this by evaluating models of different sizes on the United States Medical Licensing Examination (USMLE) [8]; and second, to establish the superiority of our GAN-inspired technique over existing state-of-the-art distillation methods. We will show this by evaluating different distillation techniques on the United States Medical Licensing Examination (USMLE)

Our contributions are twofold: (1) Enhancing the medical reasoning capabilities of smaller models, like T-5, to overcome the practical deployment limitations of LLMs in localized settings. (2) Introducing an innovative GAN-inspired approach to the distillation process, tailored to address the shortcomings of existing fine-tuning and distillation methods in complex medical reasoning tasks. Through these advancements, our work sets the stage for a new era of AI-driven healthcare support systems that are both powerful and privacy-preserving, demonstrating widespread accessibility to informed medical counsel irrespective of one's proximity to traditional healthcare facilities.

3 Related Work

The intersection of artificial intelligence (AI) and healthcare has emerged as a fertile ground for innovation, driven by the advent of large language models (LLMs) that have demonstrated remarkable capabilities in understanding and generating natural language. Among these, GPT-4 and its predecessors have set new benchmarks in natural language processing, sparking a wide array of applications in healthcare. However, the deployment of such models in healthcare settings is limited by substantial challenges, including data privacy concerns and the computational resources required for their operation.

In response to these challenges, recent research has focused on training smaller, more efficient models capable of operating within the constraints of healthcare applications. Techniques such as fine-tuning and model distillation have been pivotal in this endeavor. Standard distillation methods [5], such as those introduced by Hinton et al., involve training a smaller "student" model to replicate the behavior of a larger "teacher" model, thereby compressing knowledge into a more computationally efficient framework. Google's "Distilling Step-by-Step" paper [10] employs Chain-of-Thought (COT) prompting to create rationales for each question-answer pair. Subsequently, the student model is trained in two phases: first, to generate a rationale when provided with a question and its answer; second, to predict the answer when given both the question and the generated rationale. Moreover, Google's Med-PaLM 2 paper [9], though not specifically applied to distillation, introduces the concept of 'ensemble refinement.' This technique merges COT with self-consistency. Through temperature sampling, multiple rationales are generated for a given question. An LLM is then employed to synthesize a refined rationale, which incorporates the strengths and weaknesses of the initially generated rationales. The purpose is to improve the quality of rationales.

Projects such as MEDALPACA [4] and PMC-LLaMa [11] have played significant roles in advancing the field. By curating domain-specific datasets and models, these initiatives have provided crucial resources for the development of medical conversational AI, enabling more focused and effective model training.

Generative Adversarial Networks (GANs) [3], proposed by Goodfellow et al., represent another pivotal advancement in AI research. GANs operate on a principle of competition between two models: a generative model that produces data and a discriminative model that evaluates it. This dynamic fosters the generation of highly realistic data, offering a novel approach to model training and refinement.

Amidst these technological advancements, preserving privacy in AI-driven healthcare applications has emerged as a paramount concern [1]. Solutions leveraging federated learning, differential privacy, and encrypted computation have been explored to safeguard patient data. These methodologies aim to enable models to learn from sensitive data without directly accessing the information, thereby mitigating privacy risks.

In summary, our work builds upon a rich tapestry of research in model efficiency, domain-specific AI development, and privacy preservation. By harnessing GAN-inspired distillation techniques and incorporating privacy-preserving mechanisms, we contribute to the evolving landscape of AI in healthcare, aiming to bridge the gap between advanced computational models and practical, secure medical applications.

4 Approach

4.1 Standard finetuning

To establish a baseline, we applied traditional finetuning to our task. We had a dataset containing multiple choice questions. We trained a 770M T-5 model to predict the correct answer from the available choices (binary classification). Then, we evaluated the performance of the finetuned T-5 on the USMLE. This baseline setup provides a straightforward comparison for evaluating the effectiveness of our novel distillation approach shown later.

4.2 Google's COT approach

We also employed Google's COT approach as an alternative benchmark for comparison. For each question and answer pair in our dataset, we used GPT-3.5 Turbo (teacher) to generate a corresponding rationale explaining the answer. Then, we trained our 770M T-5 model (student) in a two step manner: (1) input combination of question and answer into T-5 model, directing it to generate rationale (2) input combination of question and rationale into T-5 model, instructing it to regenerate answer. We evaluated model performance through measuring (1) binary classification accuracy on answers (same as above) and (2) cosine similarity on rationales, where a threshold similarity of 0.3 or above between the teacher and student rationale means they can be considered equivalent in context. Again, we evaluated our model's performance on the USMLE.

4.3 GAN approach

The quality of rationales greatly influences the model's medical reasoning capabilities. Google's COT and Med-PaLM 2's combined COT and self consistency approach for generating rationales are a good first step. We propose a better alternative for generating rationales that are more medically grounded. Our GAN-approach method creates a series of guidelines that can serve as a few-shot prompt for both Google's and Med-PaLM's approaches.

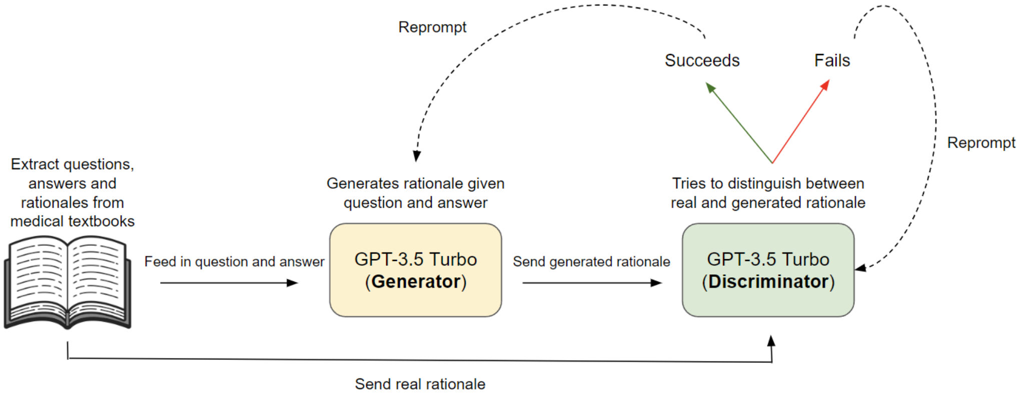
Reprompt
Succeeds
Fails
Reprompt
Extract questions, answers and rationales from medical textbooks
Generates rationale given question and answer
Tries to distinguish between real and generated rationale
Feed in question and answer
GPT-3.5 Turbo (Generator)
Send generated rationale
GPT-3.5 Turbo (Discriminator)
Send real rationale

Figure 1: GAN approach

We have a generator and discriminator. Both are GPT-3.5 Turbo models. The generator creates a rationale. The discriminator tries to identify which rational was generated by a medical expert. The core goal of this system is to improve the generator's capabilities to produce rationales that are not only medically accurate but also closely resemble the depth and precision found in medical field.

How? Both models have prompts. Overtime, the prompts are updated with guidelines based on how the discriminator performs. Below we break down the process.

To get started, we had to first construct a dataset containing medical questions, answers and rationales. We found medical exam preparation books to be an ideal source. We divided the pages into smaller chunks so that they can fit into a language model's context window. We used GPT-3.5 Turbo to extract medical questions, answers and explanations, forming a dataset.

For each entry in the dataset

- We feed in the question and answer to our generator.
- The generator using its guidelines (prompt) creates a rationale based on the question and answer.
- We feed in both the extracted rationale and generated rationale to the discriminator.
- The discriminator using its guidelines (prompt) tries to distinguish between both rationales.
- If the discriminator is successful, we probe how it made it's decision. This feedback becomes a new guideline (update prompt) for the generator.
- If the discriminator fails, it is instructed to analyze its mistake. This analysis informs a new guideline (update prompt) for discriminator.

At the end of the process, the generator's prompt contains a set of guidelines aimed at creating effective medical rationales. We aim for the model to discern and internalize the nuanced details that define a high quality medical rationale through a relentless cycle of generating, criticizing, and refining its outputs.

Returning to the paper's theme of distillation, these refined guidelines are then integrated into the prompts of the teacher model. This strategy enables the creation of medical rationales of a notably higher quality. After assembling our dataset with higher quality rationales, we proceed to train our T-5 model as usual using the above two-step approach and then assess its performance against the USMLE.

5 Experiments

5.1 Data

5.1.1 GAN

For the training of our GAN model, we had to source data that included medical questions, answers, and their corresponding rationales. We selected the Kaplan medical exam preparation book as our primary resource, due to its comprehensive coverage and structured presentation of medical knowledge.

5.1.2 T-5

For the training of the T-5 model, we had to find a dataset of multiple-choice medical questions and answers. We used MedAlpaca's Hugging Face repository, which offered a wide array of datasets.

Name	Count	Description
MedQA	10,554	General medical knowledge
MMLU clinical knowledge	265	Clinical knowledge MCQ
MMLU medical genetics	100	Medical genetics MCQ
MMLU anatomy	135	Anatomy MCQ
MMLU professional medicine	272	Professional medicine MCQ
MMLU college biology	144	College biology MCQ
MMLU college medicine	173	College medicine MCQ

Table 1: Multiple choice question datasets

5.2 Evaluation method

For the GAN model, our primary expectation is to observe convergence over the training period, a hallmark of effective GAN training. Additionally, we aim human evaluators will perceive our the generated rationales as high quality and specific. Regarding the T-5 model, our ultimate benchmark for success is its performance on the USMLE. Achieving a passing score on this exam would demonstrate the model's proficiency in medical reasoning.

5.3 Experimental details

We trained the T-5 770M model with a learning rate of 5×10^{-5}, batch size of 64, maximum input length of 1024, for a maximum of 10000 steps. The model f is trained to minimize the label prediction loss:

$$\mathcal{L}_{label} = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i), \hat{y}_i),$$

where ℓ is the cross-entropy loss between the predicted and target tokens.

5.4 Results

5.4.1 GAN

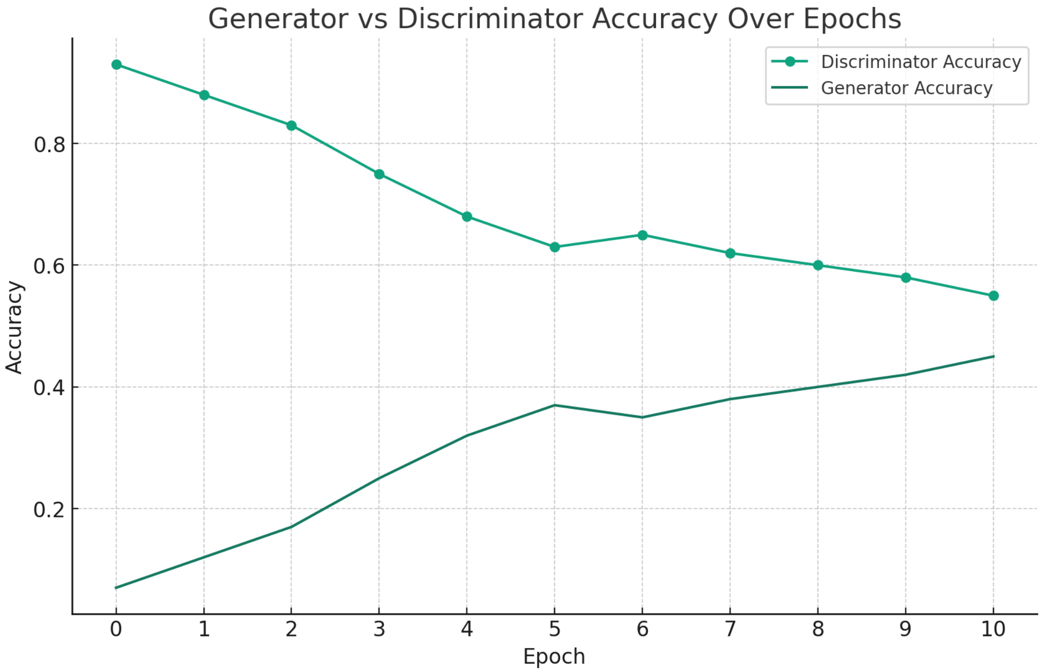

Figure 2: GAN training

The observed trend – fluctuating yet converging accuracies – aligns with expected dynamics of GAN training, where the generator and discriminator try to outcompete each other. The discriminator maintains a slight lead in accuracy over the generator by the end of the training process. This is an expected outcome, reflecting GPT's capabilities to slightly better distinguish between generated and real data, even as the other GPT improves its ability to create convincing outputs.

5.4.2 USMLE

Our GAN method outperforms all other distillation techniques in terms of accuracy. Our T-5 model comes close, but does not pass the USMLE ($< 60\%$). As expected, larger models tend to outperform T-5, especially Med-PaLM 2 which is finetuned for medical QA.

Name	Size	Accuracy on USMLE
Finetuned T-5	770M	31.4
COT T-5	770M	33.65
COT + Self Consistency T-5	770M	34.15
GAN T-5	770M	**36.71**
LLaMa (0-shot)	7B	44.54
ChatGPT	175B	57.0
Med-PaLM 2	340B	86.5

Table 2: Performance of different models on USMLE

6 Analysis

6.1 GAN

The final set of guidelines generated through our GAN methodology proved to be intriguing. The majority of these guidelines emphasized content and analytical depth. For instance, one guideline suggested, "It is imperative to offer mechanistic insights, detailing how and why a condition emerges, the molecular mechanisms behind a treatment's effectiveness, or the pathophysiological underpinnings of symptoms."

Question: What is the recommended first-line medication for managing hypertension in patients with diabetes?

Answer: ACE inhibitors are the recommended first-line medication for managing hypertension in patients with diabetes.

COT rationale: ACE inhibitors are preferred for diabetic patients with high blood pressure because they lower blood pressure and protect the kidneys, which is crucial since diabetes increases the risk of kidney problems.

GAN rationale: ACE inhibitors are prioritized for hypertension in diabetics due to their dual benefits: they reduce blood pressure by preventing angiotensin II formation, leading to vasodilation, and offer renal protection by lowering intraglomerular pressure, slowing diabetic nephropathy's progression. This dual action makes them especially valuable for diabetic patients, who face increased cardiovascular and renal risks.

However, a portion of the guidelines also appeared to prioritize the linguistic style found in medical textbooks over the actual analytical depth, marking a potential limitation of our method. An example of this is, "Prioritize the accuracy and specificity of medical terminology over the elaboration of contextual details."

In our scenario, the focus on textbook-style language was beneficial since the language used in USMLE preparation materials is similar, which probably contributed to an increase in our model's accuracy. Nonetheless, if the ultimate objective is to develop a chatbot, guidelines that concentrate on language rather than analytical depth may not prove to be advantageous.

6.2 T-5

Our finetuned T-5 model attained competitive accuracy on the USMLE, but still fell short against larger models, such as GPT. There a number of reasons for this: (1) larger models have more parameters, enabling superior capture and understanding of the nuanced, complex patterns inherent in medical data (2) LLMs such as GPT undergo extensive pre-training on diverse medical literature, embedding a wealth of implicit medical knowledge that enhances reasoning on medical QA tasks.

7 Conclusion

Our project has made significant strides in advancing the application of AI in healthcare through the development of a localized health chatbot. Our system has been carefully designed to operate with autonomy on user devices, providing immediate access to medical information without compromising on privacy and data security. By successfully fine-tuning a T-5 model to perform complex medical reasoning tasks, we have demonstrated that smaller models can indeed rival the capabilities of large language models, like GPT-4, in the realm of medical question answering to an extent.

Our GAN-inspired distillation approach outperforms state-of-the-art distillation techniques, representing a leap forward, showcasing not only an improvement in model performance on the USMLE—a rigorous benchmark for medical knowledge—but also an increase in efficiency, enabling deployment in low-resource settings where large models are impractical. Our work underscores the feasibility of delivering high-quality medical AI services directly to end-users, irrespective of their connectivity status or geographic location.

Despite these accomplishments, we acknowledge the limitations inherent in our study. The scale of data and the diversity of medical scenarios we could cover remain constrained by the model's capacity and the computational resources at our disposal. Furthermore, the ethical implications of deploying AI in sensitive areas such as healthcare necessitate ongoing vigilance and continuous improvement to ensure the safety and reliability of the advice provided.

Looking ahead, there are abundant opportunities for further research. Future work could explore the integration of larger datasets and the expansion of the model's understanding of rare medical conditions. Additionally, real-world user testing could yield invaluable insights into user experience and model performance in live scenarios. The exploration of advanced reinforcement learning techniques and more sophisticated feedback mechanisms promises to refine the model's decision-making processes even further. Lastly, a comprehensive evaluation of the ethical landscape, including user privacy and data handling, will be vital in ensuring that the deployment of such AI systems aligns with the highest standards of medical ethics and patient care.

8 References

References

[1] Mahed Abroshan, Michael Burkhart, Oscar Giles, Sam Greenbury, Zoe Kourtzi, Jack Roberts, Mihaela van der Schaar, Jannetta S Steyn, Alan Wilson, and May Yong. 2023. Safe ai for health and beyond – monitoring to transform a health service. `https://arxiv.org/abs/2303.01513`. Preprint submitted to arXiv.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, and et al. 2023. Gpt-4 technical report. `https://arxiv.org/abs/2303.08774`. Preprint submitted to arXiv on March 10, 2023.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. `https://arxiv.org/abs/1406.2661`. Preprint submitted to arXiv.

[4] Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca - an open-source collection of medical conversational ai models and training data. `https://arxiv.org/pdf/2304.08247.pdf`. Preprint submitted to arXiv on October 6, 2023.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. `https://arxiv.org/abs/1503.02531`. Preprint submitted to arXiv.

[6] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. `https://arxiv.org/pdf/2305.02301.pdf`. Preprint submitted to arXiv.

[7] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. `https://arxiv.org/abs/2303.14070`. Preprint submitted to arXiv.

[8] Prabin Sharma, Kisan Thapa, Dikshya Thapa, Prastab Dhakal, Mala Deep Upadhaya, Santosh Adhikari, and Salik Ram Khanal. 2023. Performance of chatgpt on usmle: Unlocking the

potential of large language models for ai-assisted medical education. https://arxiv.org/abs/2307.00112. Preprint submitted to arXiv.

[9] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, and et al. 2023. Towards expert-level medical question answering with large language models.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. https://arxiv.org/abs/2201.11903. Preprint submitted to arXiv.

[11] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. https://arxiv.org/pdf/2304.14454.pdf. Preprint submitted to arXiv.