Med-CoDE: Medical Critique based Disagreement Evaluation Framework

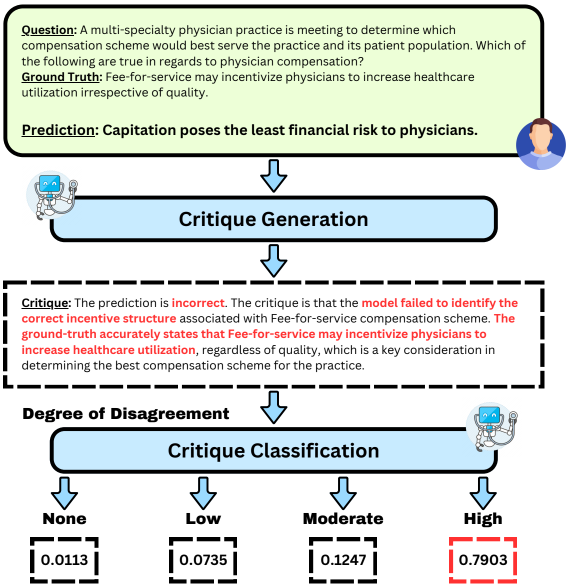
Mohit Gupta[*], Akiko Aizawa[+], Rajiv Ratn Shah[*]
[*]Indraprastha Institute of Information Technology Delhi, India,
[+]National Institute of Informatics, Tokyo, Japan
{mohit22112, rajivratn}@iiitd.ac.in[*]
aizawa@nii.ac.jp[+]

Abstract

The emergence of large language models (LLMs) has significantly influenced numerous fields, including healthcare, by enhancing the capabilities of automated systems to process and generate human-like text. However, despite their advancements, the reliability and accuracy of LLMs in medical contexts remain critical concerns. Current evaluation methods often lack robustness and fail to provide a comprehensive assessment of LLM performance, leading to potential risks in clinical settings. In this work, we propose Med-CoDE, a specifically designed evaluation framework for medical LLMs to address these challenges. The framework leverages a critique-based approach to quantitatively measure the degree of disagreement between model-generated responses and established medical ground truths. This framework captures both accuracy and reliability in medical settings. The proposed evaluation framework aims to fill the existing gap in LLM assessment by offering a systematic method to evaluate the quality and trustworthiness of medical LLMs. Through extensive experiments and case studies, we illustrate the practicality of our framework in providing a comprehensive and reliable evaluation of medical LLMs.

1 Introduction

Medical Question Answering systems based on Large Language Models represent a significant leap in leveraging artificial intelligence for healthcare. These systems are designed to process and respond to medical queries. The primary aim of Medical QA LLMs is to provide accurate, reliable, and timely information to support clinicians, researchers, and patients. Evaluating the performance of these LLMs is crucial to ensure their reliability and effectiveness in real-world medical applications. Performance evaluation typically involves assessing the accuracy, relevance, and coherence of the generated responses compared to established medical standards or expert opinions.


Question: A multi-specialty physician practice is meeting to determine which compensation scheme would best serve the practice and its patient population. Which of the following are true in regards to physician compensation?
Ground Truth: Fee-for-service may incentivize physicians to increase healthcare utilization irrespective of quality.

Prediction: Capitation poses the least financial risk to physicians.

Critique: The prediction is **incorrect**. The critique is that the **model failed to identify the correct incentive structure** associated with Fee-for-service compensation scheme. **The ground-truth accurately states that Fee-for-service may incentivize physicians to increase healthcare utilization**, regardless of quality, which is a key consideration in determining the best compensation scheme for the practice.

Figure 1: Med-Code Framework

Traditional methods for evaluating text generation, such as string similarity metrics (e.g., METEOR, BLEU, ROUGE), have been used widely across various domains. These metrics compare the overlap between generated and reference text-based on the n-gram matching, synonymy, and paraphrasing. While effective in general text generation tasks, these metrics pose significant limitations in the medical QA domain. Medical texts often require precise and contextually accurate responses where minor discrepancies can lead to substantial misunderstandings or clinical errors. Traditional metrics fail to capture the nuanced medical context, thereby providing an inadequate measure of LLM performance in this sensitive field.

To address the shortcomings of traditional evaluation methods, researchers have started exploring the use of LLMs themselves for evaluating other

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 112–119
April 30 - May 1, 2025 ©2025 Association for Computational Linguistics

LLMs. Frameworks such as Harness (Gao et al., 2023), DeepEval[1], MLFlow[2] represent this shift towards LLM-assisted evaluation. These frameworks aim to provide more contextual and comprehensive evaluations by leveraging the advanced capabilities of LLMs to understand the generated responses. Despite these advancements, the current LLM-assisted evaluation methods still lack a structured approach to quantifying disagreement and assessing reliability.

This research paper presents an reliable evaluation framework tailored for Medical QA LLMs. Drawing inspiration from the work of (Wang et al., 2023), our framework introduces a critique-based methodology that quantitatively assesses the discrepancies between model-generated responses and established medical ground truths. By employing a critique model, we analyzed the differences in LLM outputs and provide a comprehensive evaluation of their accuracy and reliability. The visual representation of Med-Code framework is shown in Fig. 1.

The contributions of this work are as follows.

- We curated a specialized medical critique dataset, incorporating medical Q&A pairs from benchmark datasets such as Medqa (Zhang et al., 2018), Medmcqa (Pal et al., 2022). etc. The dataset includes responses from various medical language models (LLMs) and a degree of disagreement label between the ground-truth answers and the models' responses.

- We developed an advanced evaluation pipeline based on the Shepherd model (Wang et al., 2023), where we fine-tuned the Phi-3 model for generating critiques and employed a BERT model for classifying them.

- To demonstrate the effectiveness of our evaluation framework, we conducted comprehensive experiments across four medical benchmark datasets, utilizing diverse evaluation techniques to ensure robust validation.

2 Related Work

This section discusses related work in the field of evaluation, highlighting previous contributions. Our motivation stems from the Shepherd Model (Wang et al., 2023), which introduces a large language model designed to generate critiques of model responses to given prompts. We extend this work by using critiques to evaluate discrepancies between model responses and ground truth.

Recent studies have shown that traditional metrics such as METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004) and BLEUScore (Zhang et al., 2020) are inadequate for accurately evaluating open-ended generation tasks due to their reliance on reference text (Chiang and Lee, 2023; Gu et al., 2021; Guan et al., 2021; Polišenská et al., 2020; Wu et al., 2021). Advances have led to new research using LLMs as evaluators, demonstrating their potential to overcome these limitations (Kim et al., 2024; Kocmi and Federmann, 2023; Liu et al., 2024b,c). Notably, approaches employing powerful LLMs like GPT-4 have achieved remarkable performance (Fu et al., 2024; Liu et al., 2023). However, current LLM-based evaluators exhibit shortcomings in robustness, as their performance is highly sensitive to prompts, leading to instability in the evaluation process. Recent studies have sought to address these challenges by generating explanations for evaluation outputs (Chiang and Lee, 2023), but this approach does not inherently improve robustness or reliability due to issues such as hallucinations (Xu et al., 2023).

In the context of medical AI, where accuracy and reliability are crucial, several research efforts propose strategies to evaluate LLM responses. An automatic evaluation metric and algorithm for LLMs' clinical capabilities is proposed in (Liu et al., 2024a), featuring a multi-agent framework with Retrieval-Augmented Evaluation (RAE) to assess the behaviors of a doctor agent. (Awasthi et al., 2023) propose a structured method for comprehensive human evaluation of LLM outputs, introducing the HumanELY guidance and tool. (Liao et al., 2024) introduce the Automated Interactive Evaluation (AIE) framework, which provides a dynamic, realistic platform for assessing LLMs through multi-turn doctor-patient simulations.

3 Methodology

In this section, we discuss the process of creating a fine-tuning dataset for the medical domain critique model, the approach we used for fine-tuning the LLM, and the development of classification model.

[1] https://docs.confident-ai.com/

[2] https://mlflow.org/

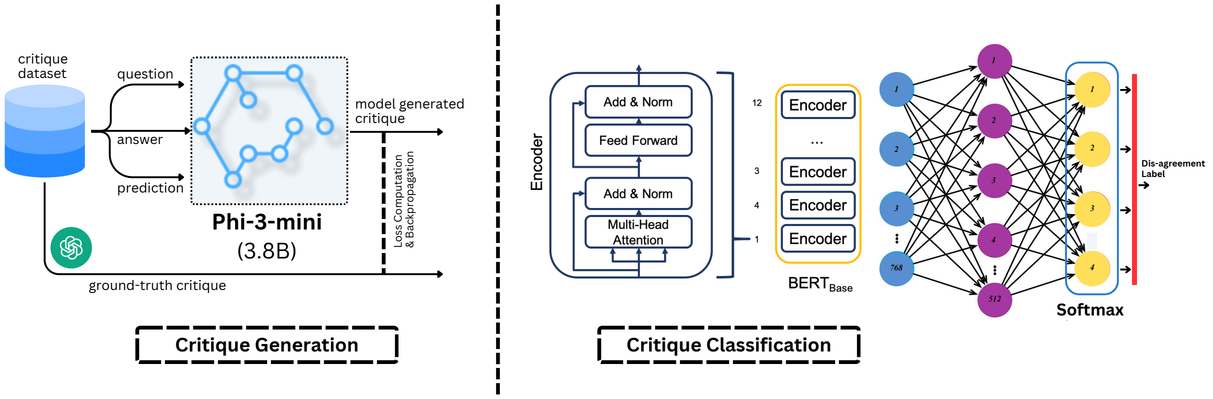


Figure 2: The overall Fine-tuning pipeline for Critique Generator & Classifier.

3.1 Dataset

In this research, we curated a specialized dataset using the OpenAI GPT-4 model to build a fine-tuning dataset for our critique generation model. Our final critique dataset comprises *38,819* samples, with an average critique length of *58.95* words. This dataset enables us to assess how well LLM responses align with ground-truth answers and to measure the degree of disagreement, providing a robust foundation for evaluating the performance of medical QA LLMs.

For medical domain data, we selected and combined small random subsets from standard medical QA datasets including Medqa (Zhang et al., 2018), Medmcqa (Pal et al., 2022), MMLU (Hendrycks et al., 2021), and Pubmedqa (Jin et al., 2019). These datasets encompass medical question-answer pairs from various medical fields, covering different levels of difficulty and types of questions. This comprehensive combination ensures that our critique model can effectively evaluate both objective and subjective questions.

```
<|user|>
You are a expert ai assistant. You are given a question, its ground-truth answer and the prediction from a model. Your task is to generate critique for the given prediction with respect to the given question, and ground-truth. This is very important and crucial task. While generating the critique, please keep the critique precise, clear and short.

### Question: {sample['question']}
### Ground-Truth: {sample['ground-truth']}
### Prediction: {sample['prediction']}

Only return the helpful answer below and nothing else.<|end|>
<|assistant|>
```

Figure 3: Critique Generation Prompt Template

After merging the random subsets, we employed SOTA medical domain LLMs, such as Meditron-7B (Chen et al., 2023), SelfBioRAG-7B (Jeong et al., 2024), to generate answers for each question. Each response was then critically evaluated using OpenAI GPT-4, which assigned a disagreement label from one of four categories: None, Low, Medium, and High. A High disagreement label indicates that the model-generated response is entirely incorrect and does not align with the ground truth in any aspect, whereas a None disagreement label signifies that the response is accurate and fully aligns with the ground truth without any extraneous information. In Low disagreement label the response is mostly accurate with minor additional details or slight deviations from the ground truth, lastly, the Moderate disagreement label, the response contains a mix of correct and incorrect information, with significant deviations from the ground truth, meaning the model is hallucinating.

3.2 Models

To build this lightweight evaluation framework, we employed two small models: ***Phi-3 3.8B*** (Abdin et al., 2024) for generating critiques & ***BERT*** (Devlin et al., 2019) for classifying the critiques. Although larger models with superior text generation capabilities and understanding are available, our objective was to create a domain-specific model tailored for a single task. Hence, these models were chosen. The visual representation of fine-tuning model architectures is shown in Fig. 2. This integrated pipeline proved efficient across all aspects, including computation, speed, and accuracy.

4 Experiments

In this section, we will delve into the experiment setup we have used for building this framework. It is divided into two subsections, first is for the critique generation model, and second is for the critique classification model.

4.1 Critique Generation Model

The objective of this model is to generate critiques based on a given question, its ground-truth answer, and the model's response. For this purpose, we employed the phi-3-mini model, which contains 3.8 billion parameters.

The hyperparameters configured for fine-tuning include *5* epochs, a batch size of *128*, a learning rate of *1.41e-5*, and the AdamW 8-bit optimizer. We utilized the LORA technique for efficient fine-tuning, with a rank parameter $r = 16$. The training process consumed an average of *20* GBs VRAM and required approximately *4-5* hours of GPU time. The data set was split into *30,000* samples for training, *4,409* for testing, and *4,410* for validation. The prompt template used in the fine-tuning and inference is given in Fig. 3.

Examples for each class of disagreement are provided in Fig. 4. These examples illustrate that the critiques generated by the model are highly precise and clear in identifying discrepancies between the ground-truth answers and the model's predictions, thereby supporting the efficacy of the fine-tuning process. To evaluate the quality of the dataset, we conducted a quality assessment on a small subset, as detailed in Section 5.1.

4.2 Critique Classification Model

For the critique classification model, we utilized the BERT base model, which contains *110M* parameters. This model is lightweight yet offers a deep bidirectional understanding of context, effectively capturing nuanced language patterns. The architecture of the entire classification network is depicted in Fig. 2.

Framework	**Accuracy**
GPT-3.5	78.12
Med-Code	71.72

Table 1: Human Evaluation Results of Disagreement Classification

The hyperparameters configured for fine-tuning are *25* epochs, a learning rate of *1e-3*, a dropout rate of *0.3*, a batch size of *16*, and a maximum sequence length of *208* tokens. The fine-tuning process employed a weighted average of all classes, with class weights specified as [*5.96, 1.34, 0.83, 0.52*]. The divergence function used is the Negative Log Likelihood. The total GPU utilization for fine-tuning this network is *2,771* MiB with *1* hour of GPU time. The data split used in this model training is *27,173* samples for training, *5,823* samples for validation, and *5,823* samples for testing.

We conducted a performance analysis of OpenAI's GPT-3.5 and our proposed framework, Med-Code, on a human labeled subset of 265 randomly selected samples. Each model received a question, a ground-truth answer, and the model's prediction, and we evaluated their accuracy in disagreement classification based on the critiques they generated. As shown in Table. 1, GPT-3.5 correctly classified approximately 207 out of 265 samples, and our proposed Med-Code framework produced results comparable to those of GPT-3.5 which is around 190 samples.

5 Results & Analysis

To assess the effectiveness of evaluating responses from large language models, we conducted experiments on four medical benchmark datasets: Medqa (Zhang et al., 2018), Medmcqa (Pal et al., 2022), Pubmedqa (Jin et al., 2019), and Mmlu (Hendrycks et al., 2021). These datasets are widely used in medical benchmarking and consist of objective-type questions. Our analysis focused on the test sets of these datasets using three LLMs: LLaMA-3 (AI@Meta, 2024), Mistral (Jiang et al., 2023), and BioMistral (Labrak et al., 2024). We selected these LLMs due to their demonstrated superior performance on general tasks and medical benchmarks.

We utilized Meteor and Rouge-L scores for automatic evaluation, the LLaMA-3 model for LLM-assisted evaluation, and our Med-Code framework to analyze LLM performance comprehensively. Med-Code categorizes responses into four degrees of disagreement, where an ideal model would show the highest average probability for "None" disagreement and the lowest for "High" disagreement. Detailed descriptions of each disagreement label are provided in the Section 3.1.

In Table 2, LLaMA-3, BioMistral, and Mistral models were used for inference. LLaMA-3 performed best on the MMLU dataset, achieving high scores in both automatic and LLM-assisted evaluations. Med-Code results showed that the "None" disagreement probability was the highest, indicating strong alignment between the model's responses and the ground-truth answers. Conversely, the "High" disagreement probability was the lowest, supporting the model's accuracy.

Dataset	**Automatic Evaluation**		**LLM-Accuracy**	**Dis-agreement Evaluation**			
	Meteor	**Rouge-L**		**None ↑↑**	**Low ↑**	**Moderate ↓**	**High ↓↓**
Results for LLaMA-3							
MEDQA USMLE	0.51	0.52	0.69	0.53	0.22	0.13	0.12
MEDMCQA	0.12	0.26	0.53	0.47	0.32	0.13	0.07
PUBMEDQA	0.11	0.12	0.39	0.55	0.30	0.10	0.05
MMLU	0.71	0.71	0.70	0.57	0.31	0.09	0.04
Results for BioMistral 7B							
MEDQA USMLE	0.14	0.07	0.74	0.44	0.29	0.16	0.11
MEDMCQA	0.16	0.08	0.61	0.35	0.39	0.18	0.08
PUBMEDQA	0.21	0.16	0.73	0.54	0.30	0.11	0.05
MMLU	0.33	0.19	0.70	0.32	0.41	0.19	0.07
Results for Mistral 7B v2.0							
MEDQA USMLE	0.16	0.12	0.68	0.47	0.28	0.15	0.01
MEDMCQA	0.56	0.11	0.56	0.33	0.38	0.20	0.08
PUBMEDQA	0.21	0.19	0.68	0.60	0.26	0.09	0.05
MMLU	0.37	0.25	0.65	0.36	0.37	0.19	0.07

Table 2: Evaluation Results for LLaMA-3, BioMistral 7B and Mistral 7B v2.0

The automatic evaluation results for BioMistral, a medical domain-specific LLM, did not convey significant information due to its poor string/semantic matching. However, BioMistral outperformed Mistral in LLM-assisted evaluation accuracy across all datasets, which was expected.

There was a strong positive correlation between accuracy and “None” disagreement probability, demonstrating that Med-Code effectively identified correct responses. Additionally, there is a positive correlation between METEOR scores and a ’Low’ disagreement probability, suggesting that the low semantic relation between ground truth and model predictions. The low positive correlation between LLM-assisted accuracy and both ’Moderate’ and ’High’ disagreement probabilities confirmed instances where the models hallucinated or produced incorrect results.

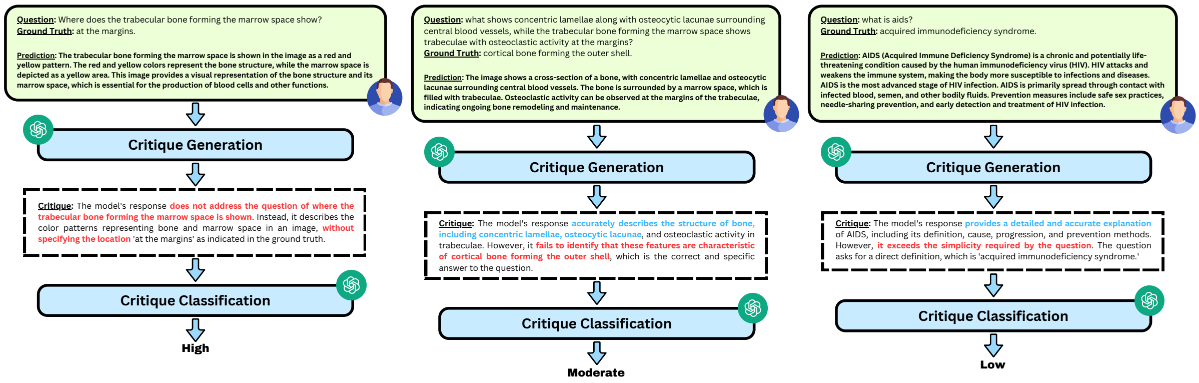
When examining the correlation between automatic evaluation scores like METEOR and ROUGE-L scores and LLM accuracy, the correlation is inconsistent across different LLMs. This inconsistency may be due to the fact that automatic metrics are based on string matching, while LLM-assisted accuracy relies on the model’s knowledge and logic. For example,

> *"If the model generates medicine X for disease D, but the ground truth answer lists medicine Y for the same disease, the automatic evaluation scores might be low. However, the LLM-assisted accuracy could still be correct because the model knows that X is equivalent to Y for disease D."*

5.1 Human Evaluation

To assess the quality of the critique data generated by the OpenAI model for fine-tuning purposes, we conducted a thorough evaluation on a randomly selected subset of 265 samples. Each sample was manually reviewed to determine how effectively the model understood the relationship between the ground-truth answer and the model’s prediction, and whether it could accurately identify minute discrepancies and details within the predictions.

Upon analysis, we found that approximately 240 out of the 265 samples (about *91%*) were accurately critiqued. The generated critiques successfully highlighted the flaws and discrepancies between the ground-truth and the predictions, demonstrating the model’s capability to provide precise and detailed feedback. This quality assessment validates the reliability of the generated data for fine-tuning the critique generation model. The ground-truth critiques are noted for their clarity and precision,



Question: Where does the trabecular bone forming the marrow space show?
Ground Truth: at the margins.

Prediction: **The trabecular bone forming the marrow space is shown in the image as a red and yellow pattern. The red and yellow colors represent the bone structure, while the marrow space is depicted as a yellow area. This image provides a visual representation of the bone structure and its marrow space, which is essential for the production of blood cells and other functions.**

Critique: The model's response **does not address the question of where the trabecular bone forming the marrow space is shown**. Instead, it describes the color patterns representing bone and marrow space in an image, **without specifying the location** 'at the margins' as indicated in the ground truth.

Question: what shows concentric lamellae along with osteocytic lacunae surrounding central blood vessels, while the trabecular bone forming the marrow space shows trabeculae with osteoclastic activity at the margins?
Ground Truth: cortical bone forming the outer shell.

Prediction: **The image shows a cross-section of a bone, with concentric lamellae and osteocytic lacunae surrounding central blood vessels. The bone is surrounded by a marrow space, which is filled with trabeculae. Osteoclastic activity can be observed at the margins of the trabeculae, indicating ongoing bone remodeling and maintenance.**

Critique: The model's response **accurately describes the structure of bone, including concentric lamellae, osteocytic lacunae**, and osteoclastic activity in trabeculae. However, it **fails to identify that these features are characteristic of cortical bone forming the outer shell**, which is the correct and specific answer to the question.

Question: what is aids?
Ground Truth: acquired immunodeficiency syndrome.

Prediction: **AIDS (Acquired Immune Deficiency Syndrome) is a chronic and potentially life-threatening condition caused by the human immunodeficiency virus (HIV). HIV attacks and weakens the immune system, making the body more susceptible to infections and diseases. AIDS is the most advanced stage of HIV infection. AIDS is primarily spread through contact with infected blood, semen, and other bodily fluids. Prevention measures include safe sex practices, needle-sharing prevention, and early detection and treatment of HIV infection.**

Critique: The model's response **provides a detailed and accurate explanation** of AIDS, including its definition, cause, progression, and prevention methods. However, **it exceeds the simplicity required by the question**. The question asks for a direct definition, which is 'acquired immunodeficiency syndrome.'

Figure 4: Critique data samples with different dis-agreement Labels

effectively pinpointing subtle differences between the ground-truth answers and the model's predictions. This ensures that the data can be effectively used for fine-tuning the critique generation model, allowing it to learn and adapt with high accuracy and precision.

6 Conclusion

In this work, we introduce Med-CoDE, an evaluation framework designed to assess the performance of Medical LLMs using critiques and degrees of disagreement. Med-CoDE excels in identifying subtle discrepancies between ground-truth answers and model predictions, offering a nuanced evaluation with four levels of disagreement. These levels provide insights into the model's behavior, such as hallucinations, accuracy, and adherence to the question. Our framework aids researchers in pinpointing areas where LLMs fall short, enabling targeted improvements. Extensive experiments on standard medical benchmark datasets demonstrate Med-CoDE's effectiveness in thoroughly and efficiently analyzing model behavior. This robust evaluation method is crucial for advancing the reliability and safety of AI-driven healthcare solutions. This evaluation framework is adaptable for assessing large language models across various domain-specific tasks as well as general tasks, simply by modifying the critique dataset.

7 Limitations

In this paper, we assess both automatic and human evaluation. Despite experimenting with a substantial number of data examples and utilizing human annotators to the best of our financial capabilities, there is room for further enhancement. Limited access to the costly OpenAI APIs meant that we used these resources judiciously, focusing on crucial areas. Additionally, computational constraints restricted the scope of our experiments. Nonetheless, these limitations highlight opportunities for future work to expand and refine the proposed framework with more extensive experimental analysis and resource allocation.

8 Ethical Considerations

The Med-CoDE framework, designed to assess the reliability and accuracy of medical LLMs, operates within a domain where the potential consequences of errors are particularly significant, given the direct impact on patient care and treatment outcomes.

In this work, only the publicly available standard benchmark medical QA datasets are used for training and evaluations. The Med-CoDE framework aims to enhance the evaluation of LLMs to ensure they meet rigorous standards of accuracy and reliability. However, it is essential to recognize that even well-evaluated models are not infallible and should not replace human judgment. Instead, they should be used as tools to support healthcare professionals, who must remain the final arbiters in clinical decision-making.

By addressing these ethical considerations, the Med-CoDE framework can contribute to the responsible development and deployment of medical LLMs, ultimately supporting safer and more effective healthcare solutions.

9 Acknowledgments

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahmoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Xihui (Eric) Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Olatunji Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

AI@Meta. 2024. Llama 3 model card.

Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra, Ashish Khanna, Kamal Maheshwari, Jacek Cywinski, Frank Papay, and Piyush Mathur. 2023. Humanely: Human evaluation of llm yield, using a novel web-based evaluation tool.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Jing Gu, Qingyang Wu, and Zhou Yu. 2021. Perception score: A learned metric for open-ended text generation evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12902–12910.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Preprint*, arXiv:2401.15269.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. *Preprint*, arXiv:2310.08491.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *Preprint*, arXiv:2403.08495.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu, Ming Zhang, Xiaowei Ma, Xiangguo Lv, Liya Ma, Zhiqiang Zhang, Wei Xue, Yiran Huang, and Jinjie Gu. 2024a. Towards automatic evaluation for llms' clinical capabilities: Metric, data, and algorithm. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5466–5475, New York, NY, USA. Association for Computing Machinery.

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2024b. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8560–8579, Mexico City, Mexico. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024c. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2638–2656, Torino, Italia. ELRA and ICCL.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Kamila Polišenská, Shula Chiat, and Jakub Szewczyk. 2020. Effects of semantic plausibility, syntactic complexity and n-gram frequency on children's sentence repetition.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Shepherd: A critic for language model generation. *Preprint*, arXiv:2308.04592.

Haiyan Wu, Zhiqiang Zhang, and Qingfeng Wu. 2021. Exploring syntactic and semantic features for authorship attribution.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.