KU-DMIS at MEDIQA-CORR 2024: Exploring the Reasoning Capabilities of Small Language Models in Medical Error Correction

Hyeon Hwang[1] **Taewhoo Lee**[1] **Hyunjae Kim**[1] **Jaewoo Kang**[1,2]
[1]Korea University [2]AIGEN Sciences
{hyeon-hwang,taewhoo,hyunjae-kim,kangj}@korea.ac.kr

Abstract

Recent advancements in large language models (LM) like OpenAI's GPT-4 have shown promise in healthcare, particularly in medical question answering and clinical applications. However, their deployment raises privacy concerns and their size limits use in resource-constrained environments. Smaller open-source LMs have emerged as alternatives, but their reliability in medicine remains underexplored. This study evaluates small LMs in the medical field using the MEDIQA-CORR 2024 task, which assesses the ability of models to identify and correct errors in clinical notes. Initially, zero-shot inference and simple fine-tuning of small models resulted in poor performance. When fine-tuning with chain-of-thought (CoT) reasoning using synthetic data generated by GPT-4, their performance significantly improved. Meerkat-7B, a small LM trained with medical CoT reasoning, demonstrated notable performance gains. Our model outperforms other small non-commercial LMs and some larger models, achieving a 73.36 aggregate score on MEDIQA-CORR 2024.

1 Introduction

Large language models (LM) have recently made significant advancements, finding usefulness across diverse applications in healthcare and medicine (Thirunavukarasu et al., 2023; Tian et al., 2024). For instance, OpenAI's GPT-3.5 (Brown et al., 2020) and GPT-4 (Achiam et al., 2024) have demonstrated their capabilities by achieving remarkable accuracy on standardized tests like the United States Medical Licensing Examination (USMLE). They have also shown excellence in real-world clinical applications—from responding to queries to diagnosing complex cases (Kung et al., 2023; Nori et al., 2023a,b; Singhal et al., 2023a,b).

However, deploying proprietary LMs in this sensitive sector presents significant challenges, primarily due to privacy concerns and the need for secure data handling (Thirunavukarasu et al., 2023; Li and Zhang, 2017; Meskó and Topol, 2023; Bartoletti, 2019). Since these models rely on APIs, it can be hard to use them in hospitals where a significant amount of sensitive personal information is present. Moreover, their vast computational requirements make them impractical for deployment on local servers in hospitals or medical research centers.

For these reasons, smaller open-sourced LMs are emerging as alternatives. For instance, models such as Mistral (Jiang et al., 2023) and BioMistral (Labrak et al., 2024) come with manageable sizes that are more suitable for deployment on local servers, while mitigating the security issues. How-

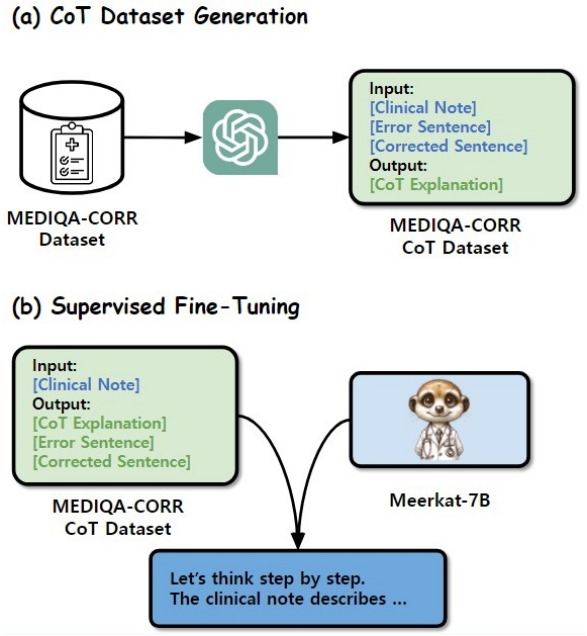

Figure 1: Overview of our proposed method. (a) In chain-of-though (CoT) dataset generation using GPT-4, we feed GPT-4 with clinical notes, error sentences, and correct sentences to generate CoT explanation that articulates error and correction. (b) In supervised fine-tuning, we fine-tune Meerkat-7B (Kim et al., 2024) with generated dataset to enhance its error detection and correction capabilities.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 526–536

ever, because these models have significantly fewer parameters (typically 7B) compared to large LMs (more than 100B), there are doubts about whether these models can provide factual responses based on their parametric knowledge. This necessitates rigorous verification before being deployed especially in the medical domain, where reliability is of utmost importance.

In this paper, we evaluate the reliability of small LMs in the medical domain. For this purpose, we utilize the MEDIQA-CORR 2024 shared task (Ben Abacha et al., 2024a), which tasks models with identifying potential errors in clinical notes and correcting them. This task assesses the ability of models to address common medical sense errors, enabling us to verify their reliability and identify hallucinations in small language models.

Our initial experiment found that when small LMs were evaluated in a zero-shot setting or trained using training data through simple supervised fine-tuning, their performance fell short of expectations. Notably, the scores were similar to random guessing in a binary classification task. This result suggests solving complex medical problems is challenging for small models lacking advanced reasoning capabilities.

Thus, we hypothesized that fine-tuning the model with chain-of-thought (CoT) reasoning (Wei et al., 2022) could effectively equip the model with these necessary reasoning capabilities. To implement this, we generated reasoning paths between the inputs and outputs of the training dataset using GPT-4 and then trained the model not only to generate correct answers but also to provide the underlying reasoning for each decision (Figure 1). This approach resulted in noticeable performance improvements, confirming the critical role of CoT reasoning in solving complex medical problems.

Furthermore, we observed that small LMs could benefit from reasoning capabilities aquired from other tasks. Specifically, Meerkat-7B (Kim et al., 2024), trained on an extensive medical CoT reasoning dataset for USMLE-style questions, showed greater performance improvements compared to other small LMs. This significant improvement highlights the importance of reasoning capabilities for small LMs to generate reliable responses.

Using this approach, we achieved an aggregate score of 73.36 for the natural language generation (NLG) evaluation, 63.46 for binary classification accuracy in detecting the presence of an error (error flag accuracy), and 61.51 for accuracy in identifying the specific sentence containing the error (error sentence accuracy) on the test set. Despite its much smaller size relative to proprietary Large LMs, Meerkat-7B demonstrated competitive performance in the MEDIQA-CORR 2024 shared task, achieving the best score among non-commercial/small LMs. This achievement is particularly significant considering the dominance of GPT-4-based frameworks among other teams.

2 Methods

2.1 Task Formulation

MEDIQA-CORR 2024 (Ben Abacha et al., 2024a) involves identifying medical errors in clinical notes and correcting them. This task is broken down into the following three sub-tasks: (1) binary classification, determining whether the clinical text contains a medical error or not, (2) span identification, detecting the specific text span associated with the error, and (3) natural language generation (NLG), creating a corrected version of the text in a free-form format. Sub-tasks 2 and 3 are performed only when an error exists in the given note.

In this study, we frame the task around generative models that produce free-form text as output. Let $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$ be a dataset, where N is the dataset size, $\mathbf{X} = \{x_n\}_{n=1}^{N}$, and $\mathbf{Y} = \{y_n\}_{n=1}^{N}$. The n-th clinical note, denoted as x_n, is composed of several sentences structured as follows:

$$x_n = \{s_1, s_2, \ldots, s_{T_n}\}, \tag{1}$$

where T_n is the number of sentences within the note, and s_i is the i-th sentence ($i \in \{1, \ldots, T_n\}$). The label set y_n consists of an error flag $e \in \{0, 1\}$, error sentence index k_n (if available), and corrected sentence s^* (if available).

We input the entire text x_n to the model, and in return, the model outputs its structured response as shown in Table 1. We parsed the model's structured response to construct the output format. If the model predicts 'the note does not contain an error.', we set the error flag to 0 and fill the fields for error sentence index and corrected sentence with '-1' and 'NA', respectively. Conversely, if the model indicates the presence of an error, we set the error flag to 1 and record both the error sentence index and the corrected sentence. The final submission format for the output is structured as follows:

$$\text{output} = \begin{cases} (e, k_n, s^*) & \text{if } e = 1 \\ (e, -1, \text{'NA'}) & \text{otherwise,} \end{cases} \tag{2}$$

<INPUT>
You are an expert tasked with providing a logical explanation as to whether there is an error in the given clinical note. Your job is to analyze the clinical note step-by-step and provide an explanation leading to the conclusion regarding the presence or absence of an error. You are strongly recommended to follow the output format: At the end of your response, without modifications, use the phrase "Therefore, the error sentence {ERROR_SENT} should be corrected to the corrected sentence {CORRECT_SENT}." or "Therefore, the note does not contain an error."

{NOTE}

ASSISTANT:

<OUTPUT>
{CoT Reasoning}

Therefore, the error sentence {ERROR_SENT} should be corrected to {CORRECT_SENT}.
or
Therefore, the note does not contain an error.

Table 1: Input and output format of our CoT dataset that was used to fine-tune small language models. The output format was specifically structured to simplify the parsing process.

2.2 Generating Reasoning Chains

We instructed GPT-4 to conduct a thorough analysis of clinical notes and provide explanations as to whether the given note potentially contains an error or not. Specifically, we use a clinical note, x, and an error flag, e, to prompt the model. When the note contains an error, we also provide both the error sentence $\hat{s}$ and the corrected sentence s^*; otherwise, we only provide the error flag as follows:

$$r = \begin{cases} \text{gpt}_{\text{err}}(x, \hat{s}, s^*) & \text{if } e = 1 \\ \text{gpt}_{\text{no}}(x) & \text{otherwise,} \end{cases} \quad (3)$$

where r is the generated reasoning chain, and gpt_{err} and gpt_{no} are the OpenAI API functions with the pre-defined input prompts. Figure 2 details the input prompts for error and non-error examples. In our initial experiments, we observed that when we did not provide label information to the model and instead asked it to determine the presence of errors and correct them, the model often gave incorrect predictions; therefore, we provided gold-standard labels to increase the recall rate of the reasoning data. An example of the reasoning chain generated by GPT-4 can be seen in Figure 3.

We generated five different reasoning paths for each example to supplement the limited amount of data. After filtering out samples that did not follow the specified output format, we obtained 9,712 and 3,207 examples from the training set and validation set, respectively. This generated dataset was pivotal in training our model, as it helped enhance the model's reasoning capabilities and as well as performance in correcting errors in clinical notes. The fine-tuning process enabled the model to generate explanations as coherent and contextually appropriate as those produced by GPT-4.

2.3 Model

As our backbone model, we used Meerkat-7B (Kim et al., 2024)[1] because it is specifically designed to handle complex medical queries through advanced multi-step reasoning. Built on Mistral-7B (Jiang et al., 2023), Meerkat-7B has been trained on a high-quality medical instruction-tuning dataset including extensive synthetic USMLE-style questions from 18 medical textbooks and corresponding CoT reasoning paths. The questions and CoT reasoning paths are generated by GPT-4, thereby endowing the model with distilled medical knowledge and reasoning capabilities from GPT-4. Leveraging these characteristics, Meerkat-7B has achieved state-of-the-art performance across various medical question-answering benchmarks that require complex reasoning.

2.4 Training and Inference

We adopted supervised fine-tuning to fine-tune a language model using our reasoning dataset. For a given clinical note, the model was trained to generate a reasoning path r first, and then structured output as shown in Table 1.

During inference, we employed a self-consistency method (Wang et al., 2023) to mitigate potential instability in the outputs generated by a single model. This method, often used as an ensemble technique, helps aggregate predictions

[1] https://huggingface.co/dmis-lab/meerkat-7b-v1.0

[Error Example]

```
...

We will let you know if there's an error and
pinpoint its location if found. Your job is
to analyze the clinical note step-by-step and
provide an explanation leading to the
conclusion regarding the presence or absence
of an error. Below is the input problem:

# Clinical Note
{NOTE}

# Conclusion
Error sentence: {ERROR_SENT}
Corrected sentence: {CORRECT_SENT}

...

# Your explanation:
```

[Non-error Example]

```
...

We will let you know if there's an error and
pinpoint its location if found. Your job is
to analyze the clinical note step-by-step and
provide an explanation leading to the
conclusion regarding the presence or absence
of an error. Below is the input problem:

# Clinical Note
{NOTE}

# Conclusion
No error in the given note.

...

# Your explanation:
```

Figure 2: The input prompts for generating CoT reasoning paths from error (left) and non-error (right) examples using GPT-4. These prompts guide GPT-4 through a detailed analysis of a clinical note to determine and explain the presence and the absence of errors within step-by-step reasoning.

from generative language models. The model generated 30 separate outputs for each input and then these outputs are aggregated to determine the most reliable result. If 'Therefore, the note does not contain an error.' is the predominant output, it is interpreted that the clinical note contains no errors. Conversely, if a specific corrected sentence emerges as the most consistent across the outputs, that sentence is selected as the final correction. This strategy reduces the impact of potentially erroneous outputs by leveraging the consensus from multiple outputs.

3 Experimental Settings

In all our experiments, we utilized eight 80GB NVIDIA A100 GPUs. When fine-tuning, we used a learning rate of 1e-6 and a batch size of 128.[2] For generating the CoT dataset, we used GPT-4 Turbo (`gpt-4-1106-preview`) through the OpenAI API.

3.1 Dataset

For our experiments, we utilized the official dataset (Ben Abacha et al., 2024b) provided by the MEDIQA-CORR 2024 shared task. Table 2 details the number of samples in each split. We used the training set for initial model tuning and selecting the best model and hyperparameters based on validation performance. For the final submission of the test set, the model was trained using a combination of the training and validation sets.

Dataset	Training	Validation	Test
MS	2,189	574	597
UW	-	160	328

Table 2: Statistics of the MEDIQA-CORR 2024 dataset. The training and validation sets were provided for model development, whereas the test split was specifically designated for the official evaluation during the challenge.

3.2 Metrics

For binary classification, we used error flag accuracy to evaluate whether the model accurately determines if a clinical text contains a medical error. We used error sentence detection accuracy for span identification to evaluate whether the model accurately outputs the index of the error sentence.

For NLG, we utilized the following evaluation metrics: ROUGE (Lin, 2004), which measures the overlap of ngrams between the generated text and the reference; BERTScore (Zhang et al., 2020), which evaluates semantic similarity using BERT embeddings; and BLEURT (Sellam et al., 2020), which assesses text generation quality based on a learned metric. Additionally, we used an AggregateScore, calculated as the arithmetic mean of ROUGE-1, BERTScore, and BLEURT. Note that these NLG evaluation metrics are computed when the model corrects an error sentence in the clinical

[2]We tested a range of learning rates, {1e-7, 5e-7, 1e-6, 5e-6, 1e-5}, and picked the best one based on performance on the MS validation set.

note that contains an error.

4 Results

4.1 Effect of Medical Reasoning on Clinical Note Correction

To verify the impact of fine-tuning with medical reasoning on clinical note correction, we evaluated three small LMs—Mistral-7B (Jiang et al., 2023), BioMistral-7B (Labrak et al., 2024), and Meerkat-7B (Kim et al., 2024)—using three methods: zero-shot CoT, fine-tuning with CoT reasoning, and fine-tuning without CoT reasoning.

Table 3 demonstrates that zero-shot CoT models exhibited poor accuracy and NLG evaluation results compared to models fine-tuned with CoT reasoning. Specifically, Mistral-7B performed worse than a random guess in the binary classification task, and BioMistral-7B largely failed to adhere to the output formats suggested in the prompts. Meerkat-7B demonstrated relatively strong performance, but there was considerable room for improvement. When fine-tuning Meerkat-7B with CoT reasoning, the performance improved by 33.51% in AggregateScore (AS), indicating that the model requires fine-tuning to adapt effectively to the target task.

In fine-tuning settings, models trained on the CoT dataset notably outperformed those trained without CoT reasoning in all metrics. Specifically, Meerkat-7B showed substantial improvements when trained with CoT reasoning: error flag accuracy increased by 9.23%, error sentence detection by 10.28%, AggregateScore by 3.36%. The result highlights the crucial role of medical reasoning in enhancing the reliability and performance of small LMs for medical domain problems.

Meerkat-7B, which was extensively trained on question-answering CoT data to enhance its complex reasoning capabilities, significantly outperformed other small language models in terms of accuracy metrics and NLG evaluation results when fine-tuned with CoT. Specifically, Meerkat-7B exceeded both Mistral-7B and BioMistral-7B in error flag accuracy, with improvements of 5.75% and 8.71% respectively. It also scored higher on NLG aggregate scores, outperforming Mistral-7B by 3.08% and BioMistral-7B by 6.79%. These results are attributed to the transfer of complex medical reasoning skills, acquired from other tasks, to the task of clinical note correction.

4.2 Official Evaluation

Based on the observations in the previous sections, we selected Meerkat-7B as our backbone model for the final submission, affirming its effectiveness for tasks requiring complex medical reasoning. Table 4 shows the official test results in the MEDIQA-CORR 2024.[3] Among the fourteen final submissions, seven teams employed large models, predominantly GPT-4, and five teams used smaller models. Large LMs demonstrated superior performance in both accuracy and NLG evaluation metrics. However, the results indicate that Meerkat-7B achieves competitive outcomes compared to them. Despite having significantly fewer parameters, our model secured fourth place overall and was the top performer among open-source and smaller LMs.

Based on the official results, our model shows substantial error flag accuracy and error sentence detection accuracy compared to other models. Still, a 63.46% accuracy rate in binary classification suggests room for improvement. To enhance our performance in binary classification, we could consider adopting an encoder model (such as BERT (Devlin et al., 2019)) focused specifically on this classification task, rather than relying solely on a general decoder model. A Two-step approach using an encoder model and decoder model in each step may help address both binary classification and correction of error sentences.

Conversely, our model achieved strong results in NLG evaluation, indicating a robust capability to generate accurate corrected sentences within the context of identified errors. This highlights its effectiveness in detailed text generation and correction tasks within the clinical domain. Given these strengths, we can expect more promising usability of our model in tasks where error existences are known, enhancing its practical application in error correction scenarios.

4.3 Case Study

We present a case study comparing reasoning from different approaches, using an example from the validation dataset. Figure 4 provides example outputs from three approaches: zero-shot CoT from each Mistral-7B and Meerkat-7B, and fine-tuned Meerkat-7B with CoT reasoning.

The clinical note describes a rash that started behind the ears and spread to the trunk and ex-

[3]We obtained basic information about the submissions from the official result sheet, but please note that we do not have precise details about the submissions.

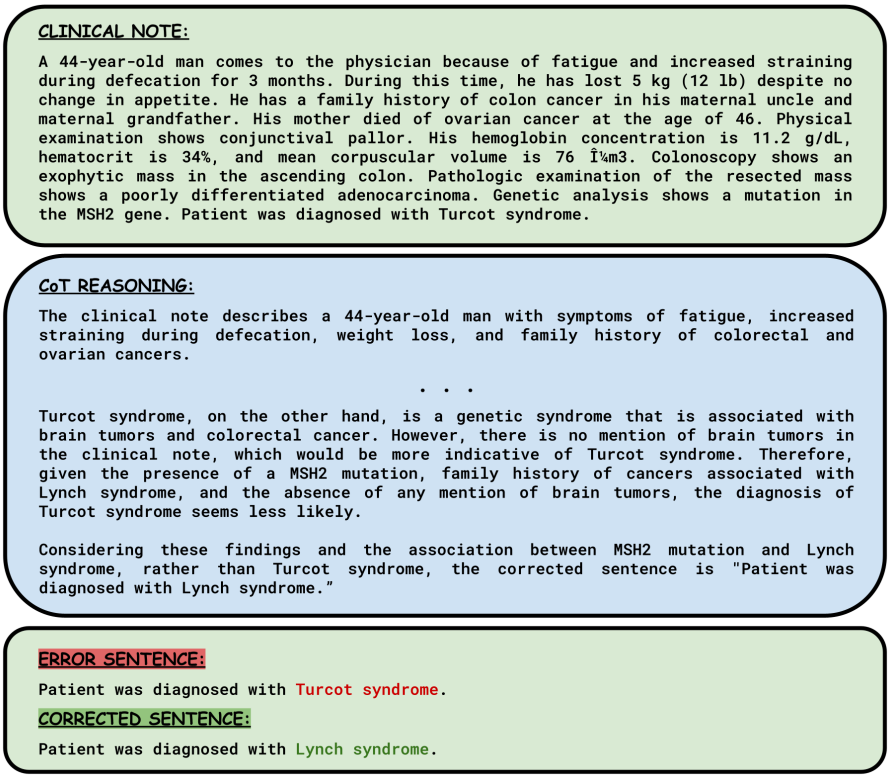

Figure 3: Example of CoT reasoning generated by GPT-4. The CoT reasoning shows a detailed explanation process in which GPT-4 uses the provided clinical note, error sentence, and corrected sentence to construct a logical reasoning path leading to the appropriate correction.

tremities, accompanied by mild sore throat, red itchy eyes, and headache. It concludes with a diagnosis of measles. However, the rash pattern and postauricular and suboccipital lymphadenopathy align more closely with rubella, which typically presents with milder symptoms and lymph node swelling. Measles would more likely involve a cough and more severe conjunctivitis, which are not mentioned.

In a zero-shot CoT setting, Mistral-7B did not detect any error in the note due to insufficient reasoning, while Meerkat-7B accurately identified the error sentence 'The patient has measles,' noting the lack of adequate evidence to conclude that the patient has measles, through step-by-step reasoning. However, the model failed to correct the sentence, indicating that it is not fully adapted to the task. In contrast, the fine-tuned Meerkat-7B with CoT reasoning successfully corrected the clinical note. It suggested that rubella is more consistent with the patient's symptoms by providing appropriate supporting reasoning. This case study demonstrates that although Meerkat-7B exhibits relatively decent medical reasoning in error detection within clinical notes compared to other baselines, fine-tuning is necessary to tailor the model for the target task.

5 Related Works

5.1 Commonsense Detection

Commonsense detection refers to the ability of an AI system, to use basic knowledge about the world that is typically obvious to humans, to understand and respond appropriately in various situations. It has traditionally been explored within general domains, such as SemEval-2020 Task 4 on Commonsense Validation and Explanation (Wang et al., 2020) and the CREAK dataset (Onoe et al., 2021). Unlike these general applications, MEDIQA-CORR 2024 Shared Task (Ben Abacha et al., 2024a) is specifically focused on the medical domain, where the implications of errors are particularly critical. Medical texts require a high degree of expertise and knowledge to not only detect errors but also correct them appropriately. This focus emphasizes the need for AI systems that perform reliably and accurately in healthcare, where factuality directly affects patient care.

	Accuracy Results		NLG Eval Results			
Model	**EF**	**ES**	**R1**	**BS**	**BL**	**AS**
Zero-shot CoT						
BioMistral-7B*	-	-	-	-	-	-
Mistral-7B	48.95	35.89	17.81	25.97	36.56	26.78
Meerkat-7B	**54.18**	**45.99**	**25.83**	**33.06**	**40.88**	**33.26**
Fine-tuning w/o CoT reasoning						
BioMistral-7B	**52.61**	47.74	49.09	57.27	50.77	52.38
Mistral-7B	48.78	46.86	61.01	66.63	58.83	62.16
Meerkat-7B	52.09	**50.17**	**61.60**	**68.26**	**60.38**	**63.41**
Fine-tuning w/ CoT reasoning						
BioMistral-7B	52.61	51.22	56.55	65.82	57.58	59.98
Mistral-7B	55.57	54.70	61.07	68.93	61.08	63.69
Meerkat-7B	**61.32**	**60.45**	**64.98**	**71.30**	**64.03**	**66.77**

Table 3: Performance of small language models on the MS validation set, evaluated through three methods: zero-shot CoT, fine-tuning without CoT reasoning, and fine-tuning with CoT reasoning. Metrics include error flag accuracy (EF), error sentence detection accuracy (ES), ROUGE-1 (R1), BERTScore (BS), BLUERT (BL), and AggregateScore (AS). We did not evaluate BioMistral-7B in the zero-shot CoT method (marked with an asterisk(*)) because this model does not generate responses in the required format, making parsing impossible. Due to superior performance compared to other models, we chose Meerkat-7B as our base model for the final submission.

5.2 Biomedical Language Models

With the success of transformer-based models on various NLP tasks, ongoing research has focused on applying them to the medical domain. Different transformer architectures have been trained with large amounts of biomedical text to encapsulate domain-specific context, including encoder-decoder-based (Yuan et al., 2022; Phan et al., 2021), encoder-based (Lee et al., 2020; Gu et al., 2021), and decoder-based (Luo et al., 2022) architectures. More recently, models equipped with billions of parameters have opened the era of Large LMs, showing superior performance and generalizability compared to smaller models. In line with this trend, recent works (Singhal et al., 2023a) have deployed various training strategies that enable Large LMs to excel at highly complex biomedical tasks, such as MedQA (Jin et al., 2021).

5.3 Reasoning Distillation

LMs have shown to generate CoT reasoning steps that can benefit end task performance, but only when equipped with at least 100 billion parameters (Wei et al., 2022). To this end, recent works have focused on distilling reasoning chains derived from larger models to smaller models (Li et al., 2022; Magister et al., 2023). SOCRATIC CoT (Shridhar et al., 2023) suggests a two-step approach, where a *problem decomposer* model interacts with a *sub-problem solver* model to reach the final solution.

6 Conclusion

In this study, we explored the capabilities of small open-sourced language models in medical error correction and the effect of CoT reasoning on this problem. Our findings confirm that CoT reasoning capabilities are highly encouraged for the task of clinical note correction, especially for small LMs. Particularly, Meerkat-7B, initially trained to solve complex medical questions using an extensive CoT dataset, demonstrates superior performance compared to other open-sourced small LMs. Despite having far fewer parameters than proprietary large LMs, Meerkat-7B achieves competitive performance in clinical note correction. This underscores the potential of well-designed smaller models to handle demanding medical AI tasks effectively. In future research, there should be ongoing efforts to continuously improve small LMs to enhance the reliability and safety of automated systems in healthcare, paving the way for more

Rank	Base Model	Model Size	Accuracy Results		NLG Eval Results
			EF	ES	AS
1	GPT-4	Large	**86.49**	**83.57**	**78.91**
2	GPT-4 & Claude Opus	Large	62.16	60.86	78.66
3	GPT-4	Large	52.22	52.00	78.06
4	Meerkat-7B (**Ours**)	Small	63.46	<u>61.56</u>	<u>73.36</u>
5	Palmyra	Small	56.00	52.00	73.30
6	OpenAI (Not Specified)	Large	66.92	61.08	71.09
7	GPT-4	Large	69.41	61.95	65.81
8	OpenAI (Not Specified)	Large	68.00	64.00	58.75
9	GPT-4	Large	67.41	60.97	58.10
10	GPT-4	Large	67.78	59.03	55.87
11	GPT-4	Large	56.65	49.08	48.09
12	BioMistral-7B	Small	50.16	37.84	45.01
13	BioMistral-7B	Small	53.95	36.32	44.83
14	BART & SVM	Small	<u>73.73</u>	60.00	44.56

Table 4: Official evaluation on MEDIQA-CORR 2024, featuring metrics such as error flag accuracy (EF), error sentence detection accuracy (ES), and aggregate score (AS). The table lists each base model used and roughly categorizes them into 'Large' or 'Small' based on their parameter size. Ranks are determined based on the aggregate score (AS). The best performance in each metric is highlighted in **bold**, while the best performance among small models is <u>underlined</u>. Our Meerkat-7B-based model achieved an aggregate score of 73.37, outperforming all small models and several large model-based systems.

accurate and trustworthy medical assistants.

Limitation

One limitation of our current approach is that the model's integration of external knowledge sources is not fully developed (e.g., knowledge base- or retrieval-augmented generation). While Meerkat-7B exhibits high-quality reasoning capabilities, it has not yet been optimized to incorporate external knowledge. Integrating this model with a retriever and utilizing biomedical knowledge sources could significantly improve its ability to solve complex cases while reducing the likelihood of generating hallucinations. Future work will explore adapting Meerkat-7B to harness external biomedical corpora, potentially increasing its accuracy and reliability.

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF2023R1A2C3004176, RS-2023-00262002), and the Ministry of Health & Welfare, Republic of Korea (HR20C0021(3)).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ivana Bartoletti. 2019. Ai in healthcare: Ethical and privacy challenges. In *Artificial Intelligence in Medicine*, pages 7–10, Cham. Springer International Publishing.

Asma Ben Abacha, Wen wai Yim, Velvin Fu, Zhaoyi Sun, Fei Xia, and Meliha Yetisgen. 2024a. Overview of the mediqa-corr 2024 shared task on medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Asma Ben Abacha, Wen wai Yim, Velvin Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2024b. Medec: A benchmark for medical error detection and correction in clinical notes. *CoRR*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec

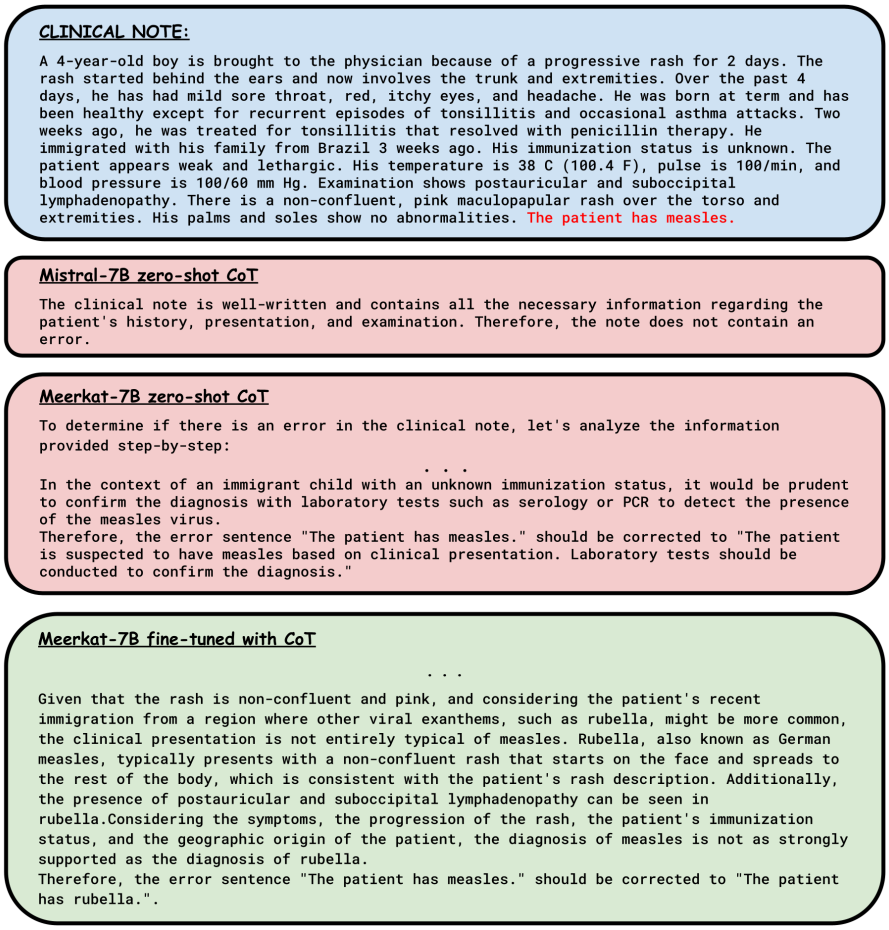

Figure 4: Example outputs from three distinct approaches to clinical note correction. The clinical note contains an error sentence highlighted in red text. The approaches include zero-shot CoT from both Mistral-7B and Meerkat-7B, and fine-tuned Meerkat-7B with CoT reasoning. A green rounded rectangle indicates an accurate correction of the error, while a red rounded rectangle signifies an incorrect response. Meerkat-7B zero-shot CoT detected the error sentence accurately but failed to correct the error sentence due to not fully adapting to the correction task.

Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

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.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Donghee Choi, and Jaewoo Kang. 2024. Small language models learn enhanced reasoning skills from medical textbooks. *Preprint*, arXiv:2404.00376.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance

of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

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. *Preprint*, arXiv:2402.10373.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. *Preprint*, arXiv:2210.06726.

Xiuquan Li and Tao Zhang. 2017. An exploration on artificial intelligence application: From security, privacy and ethic perspective. In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 416–420.

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.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Bertalan Meskó and Eric J Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *Preprint*, arXiv:2303.13375.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023b. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Preprint*, arXiv:2311.16452.

Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A dataset for commonsense reasoning over entity knowledge. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *Preprint*, arXiv:2106.03598.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

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, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023b. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2024. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop*

on Semantic Evaluation, pages 307–321, Barcelona (online). International Committee for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.