KnowLab_AIMed at MEDIQA-CORR 2024: Chain-of-Though (CoT) prompting strategies for medical error detection and correction

Zhaolong Wu[1*], **Abul Hasan**[2*],
Jinge Wu[2], **Yunsoo Kim**[2], **Jason P.Y. Cheung**[1†], **Teng Zhang**[1†], **Honghan Wu**[2†]
[1]Department of Orthopaedics and Traumatology, University of Hong Kong
[2]Institute of Health Informatics, University College London
{wuzl01}@connect.hku.hk,
{cheungjp, tgzhang}@hku.hk,
{a.kalam, jinge.wu.20, yunsoo.kim.23, honghan.wu}@ucl.ac.uk

Abstract

This paper describes our submission to the MEDIQA-CORR 2024 shared task for automatically detecting and correcting medical errors in clinical notes. We report results for three methods of few-shot In-Context Learning (ICL) augmented with Chain-of-Thought (CoT) and reason prompts using a large language model (LLM). In the first method, we manually analyse a subset of train and validation dataset to infer three CoT prompts by examining error types in the clinical notes. In the second method, we utilise the training dataset to prompt the LLM to deduce reasons about their correctness or incorrectness. The constructed CoTs and reasons are then augmented with ICL examples to solve the tasks of error detection, span identification, and error correction. Finally, we combine the two methods using a rule-based ensemble method. Across the three sub-tasks, our ensemble method achieves a ranking of 3rd for both sub-task 1 and 2, while securing 7th place in sub-task 3 among all submissions.

1 Introduction

The rise of Large Language Models (LLMs) such as GPT4 (Achiam et al., 2023), Med-PaLM (Singhal et al., 2023), and LLaMA (Touvron et al., 2023a,b) have inspired investigations into their potential use in automatically analysing Electronic Health Records (EHRs). However, the usefulness of LLMs in clinical settings remains challenging due to the fact that these models are trained on large-scale corpora which may contain inaccuracies, common mistakes, and misinformation (Thirunavukarasu et al., 2023; Ji et al., 2023). To motivate research on the problem of identifying and correcting common sense medical errors in clinical notes using LLMs, the MEDIQA-CORR (Medical Error Detection Correction) shared tasks are proposed. Herein, we describe our submissions to the shared tasks presenting two methodologies and an ensemble approach using GPT4, all utilising In-Context Learning (ICL) (Brown et al., 2020) in conjunction with Chain-of-thought (CoT) (Wei et al., 2022; Wang et al., 2022b) and reason prompts. The ensemble method achieves accuracies of 69.40% and 61.94% for sub-task 1 and sub-task 2, respectively, while obtaining a BLUERT score of 0.6541 for sub-task 3.

2 Shared Tasks and Dataset

2.1 Shared Tasks

The MEDIQA-CORR 2024(Ben Abacha et al., 2024a) proposes three sub-tasks:

1. **Binary Classification (sub-task 1)**: To detect whether a clinical note contains a medical error.
2. **Span Identification (sub-task 2)**: To identify the text span (i.e. Error Sentence ID) associated with the error, if a medical error exists in the clinical note.
3. **Natural Language Generation (sub-task 3)**: To generate a corrected text span, if a medical error exists in the clinical note.

2.2 Dataset

The training dataset is derived from a single source called as MS Training Set, where as the validation and test datasets are derived from two different sources termed as MS and UW Validation/Test set (Ben Abacha et al., 2024b). The MS Training Set is comprised of 2,189 clinical notes. The MS Validation Set includes 574 clinical notes, while the UW Validation Set includes 160 clinical notes. The

[*]These authors contributed equally to this work.
[†]Corresponding authors.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 353–359
June 21, 2024

Test dataset has in total 926 clinical notes derived from two sources.

3 Methods

3.1 ICL-RAG- augmented with CoT prompting(ICL-RAG-CoT)

The Chain-of-Thought (CoT) prompting method, which includes a sequence of reasoning steps, has demonstrated enhancements in the problem-solving capabilities of LLMs over standard prompting techniques, particularly in solving mathematical tasks (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2024). Recent studies, such as the one conducted by (Kim et al., 2023), have introduced datasets that incorporate CoT instructions aimed at addressing various Natural Language Processing (NLP) tasks. These tasks include question answering and natural language inference and have been tailored for smaller-scale language models like Flan-T5 (Longpre et al., 2023). Motivated by these developments, we conduct a manual analysis of a subset derived from both the MS Training set and UW Validation set to investigate the prevalent error types within clinical notes. Our examination reveals three broad categories of errors evident in the clinical notes; they are : (1) Diagnosis, (2) Intervention, and (3) Management. Using these categories we construct three separate prompts, shown in Figure 1, that are augmented with ICL examples.

To address the three sub-tasks, our initial approach, referred to as ICL-RAG augmented with CoT prompting (ICL-RAG-CoT), adopts a two-stage prompting methodology with GPT4. For the binary classification and span identification tasks (i.e. sub-task 1 and sub-task 2), we guide GPT4 systematically through a sequence of prompts, each tailored to detect and identify medical errors. The first prompt in the sequence is a standard prompting which tasks the model to detect errors in a clinical note, supplemented with in-context examples. If no medical error is detected, we proceed to prompt GPT4 iteratively by augmenting our CoTs in Figure 1 with ICL examples until an error is identified. Once all CoTs are exhausted, the clinical note is considered error-free. In the second stage, for the NLG task, we prompt GPT4 independently by specifying the predicted incorrect sentence number (i.e., Sentence ID) obtained from the first stage. A prompt template is provided in Appendix A; see Figure 4. In order to generate In-context examples for prompting LLMs, our methodology incorporates the Retrieval-Augmented Generation (RAG) approach, as proposed by Lewis et al. (2020); Jin et al. (2024). Utilising the e5-large-unsupervised model (Wang et al., 2022a), we transform the MS-Training dataset into a vectorized database. This process involves applying cosine similarity to find the k-most similar training instances for each validation and test input. In our experiments we select k=4.

3.2 ICL-RAG- augmented with reason (ICL-RAG-Reason)

In our second method, referred to as ICL-augmented with reason (ICL-RAG-Reason), we aim to address three sub-tasks simultaneously using a single prompt containing ICL examples and their corresponding reasons for correctness or incorrectness. However, this method requires to prompt the LLM to pre-process the training data separately. Consequently, the ICL-RAG-Reason method begins by prompting GPT4 to generate a brief reason for the correctness or incorrectness of a clinical note from the MS Training set; see Figure 2 for an example. If a note contains an error, we prompt the LLM by concatenating it with the corrected sentence to explain why the clinical note is deemed incorrect. In the case of a correct training example, we prompt the GPT4 to provide us with the clinical characteristics that validate the note's correctness. Thus, we automatically construct reasoning instructions for each MS Training notes. We employ a similar RAG method to ICL-RAG-CoT; however, we utilize OpenAI embeddings [1] to embed all clinical notes across the three datasets. For every input validation and test note, we sample 4 (4-shot) training notes from a pool of its semantically most similar k notes, comprising two correct and two incorrect notes. We augment selected training notes with their *Reasons* for being correct or incorrect and create the final prompt; ; see Figure 5 in Appendix A for an example of prompt template. The ICL-RAG-Reason method samples ICL examples three times to ensure that the model is shown different reasoning paths. This sampling strategy provides us with three different solutions which is resolve by majority voting to ensure consistency and then take the corrected sentence by randomly selecting one from two correct answers.

[1] https://platform.openai.com/docs/guides/embeddings

(1)

Focus on intervention

[Input . . .]
The text above is divided into sentences each with a number. Please check the diagnosis result in the text above. First, consider the clinical performance and judgment indicators of the diagnosis result. Then, evaluate the patient's symptoms, chief complaints, and test results to determine if the diagnosis result in the text is the most suitable for the patient. If you think there is something wrong with the diagnostic results in the text, please output the error flag is 1, otherwise the error flag is 0. If there is no clear diagnosis in the text, error flag is 0. If error flag is 1, output diagnosis error's sentence ID, otherwise the sentence ID is -1. Output must follow the format, don't output other words.
Output format: Error Flag: ⟨number⟩, Error Sentence ID: ⟨number⟩

(2)

Focus on diagnosis

[Input . . .]
The text above is divided into sentences each with a number. Please check the diagnosis result in the text above. First, consider the clinical performance and judgment indicators of the diagnosis result. Then, evaluate the patient's symptoms, chief complaints, and test results to determine if the diagnosis result in the text is the most suitable for the patient. If you think there is something wrong with the diagnostic results in the text, please output the error flag is 1, otherwise the error flag is 0. If there is no clear diagnosis in the text, error flag is 0. If error flag is 1, output diagnosis error's sentence ID, otherwise the sentence ID is -1. Output must follow the format, don't output other words.
Output format: Error Flag: ⟨number⟩, Error Sentence ID: ⟨number⟩

(3)

Focus on management

[Input . . .]
The text above is divided into sentences, and each sentence has a number. Please check management sentence. Please only consider the normal case. Then check management whether have error. I will give you three error type: 1. Please first review relevant medical guidelines. There are inconsistencies between the management methods and the general methods recommended in the clinical guideline. 2. This management maybe over-aggressive exam/therapy. 3. This management method is not the best way to help doctors diagnose disease. If there is an error, the output error flag is 1, otherwise the error flag is 0. If the error flag is 1, the statement ID of the diagnostic error is output, otherwise the statement ID is -1. The output must follow the format and no other text should be output.
Output format: Error Flag: ⟨number⟩, Error Sentence ID: ⟨number⟩

Figure 1: Three types of Chain-of-Thought (CoT) prompts utilised in the ICL-RAG-CoT method: (1), (2), and (3) direct the GPT4 model to focus on intervention, diagnostic, and management errors, respectively.

Reason Generation

The following clinical note contains an error in the sentence number 6.
A 53-year-old man comes to the physician because of a 1-day history of fever and chills, severe malaise, and cough with yellow-green sputum. He works as a commercial fisherman on Lake Superior. Current medications include metoprolol and warfarin. His temperature is 38.5 C (101.3 F), pulse is 96/min, respirations are 26/min, and blood pressure is 98/62 mm Hg. Examination shows increased fremitus and bronchial breath sounds over the right middle lung field. After reviewing imaging, the causal pathogen was determined to be Haemophilus influenzae. An x-ray of the chest showed consolidation of the right upper lobe.
The corrected sentence is:
After reviewing imaging, the causal pathogen was determined to be Streptococcus pneumoniae.
Question: What is(are) the clinical error(s) present in the clinical note? Write no more than 50 words.

Answer

The clinical error in the note is the identification of the causal pathogen. The note states it is Haemophilus influenzae, while it should be Streptococcus pneumoniae, as per the corrected sentence. The pathogen can't be determined solely from imaging, but requires laboratory testing.

Figure 2: Reason generation template utlised in the ICL-RAG-Reason method

3.3 Ensemble

We integrate the ICL-RAG-CoT and ICL-RAG-Reason methods using a rule-based approach, henceforth termed as the Ensemble method. This approach initially considers predictions generated by the ICL-RAG-CoT method for sub-task 1 and sub-task 2 as correct, while predictions for sub-task 3 from ICL-RAG-Reason are also deemed correct. It then resolves conflicts by identifying clinical notes from the MS and UW Validation and Test sets that are predicted as incorrect by both methods but have differing Error Sentence IDs. Finally, the Ensemble method prompts GPT4 (see see Figure 6 in Appendix A for an example), providing it with ICL examples, each containing an error, to generate a corrected sentence by specifying the Eorror Sentence ID predicted by the ICL-RAG-CoT.

3.4 Evaluation

We evaluate the performances of our methods with the official evaluation scripts on MS and UW Validation Set [2]. Sub-task 1 and 2 are evaluated by using Accuracy. The Natural Language Generation task (i.e. sub-task 3) is evaluated with with ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), and BLEURT (Sellam, Thibault and Das, Dipanjan and Parikh, Ankur, 2020). We report performances as

[2] https://github.com/abachaa/MEDIQA-CORR-2024

Table 1: Main results. Here Acc, AG, R1, and AGC denote Accuracy, Aggregate, ROUGE-1, and AggregateC scores, respectively.

	Sub-task 1	Sub-task 2	Sub-task 3				
Method	Acc	Acc	AG	R1	BERT	BLEURT	AGC
MS Validation							
ICL-RAG-CoT	**0.6620**	**0.6236**	**0.6350**	**0.6028**	**0.6658**	**0.6363**	**0.5067**
ICL-RAG-Reason	0.6010	0.5644	0.6165	0.5739	0.6577	0.6178	0.4298
Ensemble	**0.6620**	**0.6236**	0.6184	0.5777	0.6560	0.6215	0.5048
UW Validation							
ICL-RAG-CoT	**0.7437**	**0.6500**	0.6525	0.6701	0.6519	0.6355	0.6091
ICL-RAG-Reason	0.6875	0.5625	0.6340	0.6180	0.6343	0.6499	0.5350
Ensemble	**0.7437**	**0.6500**	**0.6740**	**0.6762**	**0.6729**	**0.6728**	**0.6174**
Test							
ICL-RAG-CoT	**0.6940**	**0.6194**	0.6255	0.6130	0.6399	0.6235	0.5346
ICL-RAG-Reason	0.6540	0.5837	0.6509	0.6343	0.6703	0.6482	0.5119
Ensemble	**0.6940**	**0.6194**	**0.6581**	**0.6434**	**0.6767**	**0.6541**	**0.5730**

the arithmetic mean of ROUGE-1 F1, BERTScore, BLEURT-20. Furthermore, Aggregate scores and AggregateComposite scores, the overall measures across the mentioned metrics, are provided.

4 Results

We attain accuracies of 66.20%, 74.37%, and 69.40% on the MS Validation, UW Validation, and Test datasets, respectively, for the binary classification task of error detection (i.e. sub-task 1) using the ICL-RAG-CoT method; see Table 1. For the span identification task, i.e. sub-task 2, the same method achieves accuracies of 62.36%, 65.00%, and 61.94%, respectively. It is noteworthy that the Ensemble method achieves similar accuracies. In the sub-task 3, which involves Natural Language Generation (NLG), the ICL-RAG-CoT method performs less effectively compared to the ICL-RAG-Reason method. It reaches a BLEURT score of 0.6363 on the MS Validation Set. However, our Ensemble approach surpasses the other two methods, achieving BLEURT scores of 0.6729 and 0.6541 for the UW Validation and Test sets, respectively. We observe similar perfomances across other NLG metrics; see Table 1. This is because the reasoning generation method. i.e. ICL-RAG-Reason achieves better performances than the ICL-RAG-CoT method particularly in the NLG task.

5 Discussion

Our CoT prompting strategy works well in conjunction with the RAG system. As depicted in Figure 3, across various few-shot settings (e.g., 2, 3, 4, and 5-shot settings), the ICL-RAG-CoT method consistently outperforms scenarios where CoT is not employed alongside RAG in the binary classification task. We observe that both the 3-shot and 5-shot settings yield lower performance compared to the 2-shot and 4-shot settings. This disparity suggests that class imbalance in few-shot settings could potentially deteriorate performance. This motivates our selection of 4-shot setting consistently across all our experiments. One of the limitations of our study is that we do not rigorously evaluate the NLG Task, i.e. sub-task 3. Consequently, our overall ranking falls towards the lower end of the top 10 (ranked 7 over-all). While our Ensemble prompting strategy demonstrates a good performance by leveraging reasoning gathered independently from GPT4, there remains scope for improvement. For

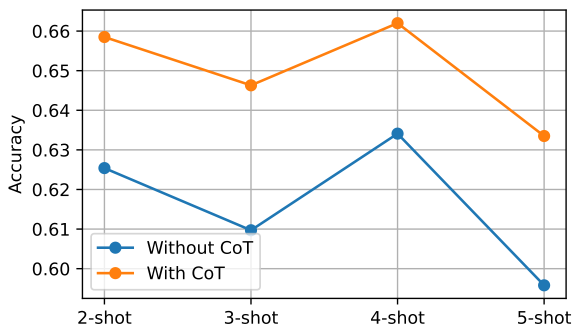

Figure 3: Comparison of few-shot examples with or without CoT using ICL-RAG-CoT method on the Binary Classification Task (i.e. sub-task 1) on the MS Validation Set

instance, further enhancement could be achieved by evaluating the generation of LLMs against clinical and/or biomedical knowledge bases to verify their output.

6 Conclusion

We present our submission to the MEDIQA-CORR shared task for medical error detection and correction. Our study evaluates the effectiveness of the GPT4 model through various prompting strategies employing CoT prompting and Reasoning methods. Specifically, our CoT prompting strategies achieve high accuracies in error detection and identification tasks. Additionally, our Ensemble method, which combines outputs from both methods, demonstrates a better performance on the NLG task than the CoT prompting alone. In the future, we aim to explore our approach for other downstream tasks in the clinical domain using open-source LLMs.

7 Ethical Statement

Our research employs large language model (LLM) to improve the accuracy of medical records. However, before deploying and utilising the methods proposed with LLM, it is necessary to adhere to ethical and moral principles. The storage and use of patient data must strictly comply with data protection and privacy laws, such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR), to ensure that data access is strictly controlled and process transparency is maintained.

8 Acknowledgement

Abul Hasan and Honghan Wu are funded by the National Institute for Health Research (NIHR) Artificial Intelligence and Multimorbidity: Clustering in Individuals, Space and Clinical Context (AIM-CISC) grant NIHR202639. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

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, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, Suiyuan Zhu, Mengnan Du, Yongfeng Zhang, and Yanda Meng. 2024. Health-llm: Personalized retrieval-augmented disease prediction model. *arXiv preprint arXiv:2402.00746*.

Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The CoT Collection: Improving Zero-shot and Few-shot Learning of Language Models via Chain-of-Thought Fine-Tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12685–12708.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Sellam, Thibault and Das, Dipanjan and Parikh, Ankur. 2020. BLEURT: Learning Robust Metrics for Text

Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. 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. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA:OpenandEfficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: OpenFoundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Prompt Templates

Few-shot Prompt

Detect whether the text below contains a medical error? If an error is found, set the Error Flag to 1 and output Error Sentence ID; otherwise, set Error Flag to 0 and Error Sentence ID to -1.
Output must follow output format!
The output should not exceed 50 words!
Output format:
Error Flag: ⟨number⟩
Error Sentence ID: ⟨number⟩
Input:
⟨Same as (a)⟩[. . .]
Please read the example below:
Example 1:
0 A 6-year-old girl is brought to the physician for intermittent fevers and painful swelling of the left ankle for 2 weeks.
1 She has no history of trauma to the ankle.
2 She has a history of sickle cell disease.
3 Current medications include hydroxyurea and acetaminophen for pain.
4 Her temperature is 38.4 C (101.2 F) and pulse is 112/min.
5 Examination shows a tender, swollen, and erythematous left ankle with point tenderness over the medial malleolus.
6 A bone biopsy culture confirms the diagnosis as it grew Streptococcus pneumoniae.
Error Flag: 1
Error Sentence ID: 6
Error Sentence: A bone biopsy culture confirms the diagnosis as it grew Streptococcus pneumoniae.
Corrected Sentence: A bone biopsy culture confirms the diagnosis as it grew Salmonella enterica.
Example 2:
⟨another example ⟩[. . .]
[CoT Part]

Answer

Error Flag: 1
Error Sentence ID: 5

Figure 4: A template used in ICL-RAG-CoT for the few-shot prompting to solve sub-task 1 and 2.

Few-shot Prompt

You are a helpful clinical assistant who has clinical knowledge. You will be provided with four Examples of patient records, each of which may or may not contain an error. I will specify reasons for each record being correct or incorrect under the “Reason” section. Then, you will be given a “Patient Record” to analyze whether it contains a medical error. You should examine the diagnosis, management, and intervention-related statements.[...]
Example: 1
0 A 53-year-old man comes to the physician because of a 3-month history of a nonpruritic rash.[...]
Output:
Error Flag: 0
Error Sentence ID: -1
Corrected Sentence: NA
Example: 2 [...]
[Reasons:]
Example 1 contains no error because:
This clinical note is correct as it provides comprehensive and relevant patient information, including symptoms, medical history, physical exam findings, and results of various laboratory tests.[...]
Patient Record:
0 A 6-year-old girl is brought to the physician for intermittent fevers and painful swelling of the left ankle for 2 weeks.[...]
Output format:
Error Flag:⟨number⟩
Error Sentence ID:⟨number⟩
Corrected Sentence:⟨text⟩

Answer

Output:
Error Flag: 1
Sentence ID: 6
Corrected Sentence: A bone biopsy culture confirms the diagnosis as it grew Salmonella, not Streptococcus pneumoniae.

Figure 5: A template used in ICL-RAG-Reason for the few-shot prompting to solve all sub-tasks simultaneously.

Few-shot Prompt

You are a helpful clinical assistant who has clinical knowledge. You will be provided with four Examples of patient records, each of which contains an error. I will specify reasons for each record being incorrect under the “Reason” section. Then, you will be given a “Patient Record” with a possible Error Sentence ID to analyze its error. You should examine the diagnosis, management, and intervention-related statements. If an error is found in the Patient Record then generate a Corrected Sentence.[...]
Example: 1
0 A 27-year-old man comes to the physician with throbbing right scrotal pain for 1 day. [...]
Output:
Error Flag: 1
Error Sentence ID: 5
Corrected Sentence: Further evaluation reveals chlamydia trachomatis as the causal pathogen.
Example: 2 [...]
[Reasons:]
Example 1 contains an error because:
The clinical error in the note is the incorrect identification of the causal pathogen. The symptoms and signs described, including the Prehn’s sign (relief of pain on lifting the testicle), are indicative of epididymitis, which is most commonly caused by Chlamydia trachomatis in sexually active young men, not Neisseria gonorrhoeae.[...]
Patient Record:
0 A 22-year-old man comes to the physician because of a progressive swelling and pain in his right ring finger for the past 2 days.[...]
Error Flag: 1
Possible Error Sentence ID: 6
Output format:
Corrected Sentence:⟨text⟩

Answer

Output:
Corrected Sentence: Patient was diagnosed with a rupture of the flexor digitorum profundus tendon.

Figure 6: A template used in Ensemble method for the few-shot prompting to solve the sub-task 3.