

CLD-MEC at MEDIQA- CORR 2024 Task: GPT-4 Multi-Stage Clinical Chain of Thought Prompting for Medical Errors Detection and Correction

Renad M. Alzghoul¹, Abdulrahman Tabaza¹, Aya Abdelhaq¹, Ahmad Altamimi¹

¹Princess Sumaya University for Technology, Amman, Jordan

{ren20228156, abd20200209, aya20228163, a.altamimi}@std.psut.edu.jo

Abstract

This paper demonstrates CLD-MEC team submission to the MEDIQA-CORR 2024 shared task for identifying and correcting medical errors from clinical notes. We developed a framework to track two main types of medical errors: diagnostics and medical management-related errors. The tracking framework is implied utilizing a GPT-4 multi-stage prompting-based pipeline that ends with the three downstream tasks: classification of medical error existence (Task 1), identification of error location (Task 2), and correction error (Task 3). Throughout the pipeline, we employed clinical Chain of Thought (CoT) and Chain-of-Verification (CoVe) techniques to mitigate the hallucination and enforce the clinical context learning. The model performance is acceptable, given it is based on zero-shot learning. In addition, we developed a RAG system injected with clinical practice guidelines as an external knowledge datastore. Our RAG is based on the Bio_ClinicalBERT as a vector embedding model. However, our RAG system failed to get the desired results. We proposed recommendations to be investigated in future research work to overcome the limitations of our approach.

1 Introduction

Medical errors identification and handling in clinical practice is paramount to ensuring optimized patient safety and efficient healthcare delivery. Yet, tracking medical errors is complex to achieve. The umbrella of medical error instances is wide, along with the different phases of the patient journey from assessment to diagnosis, followed by medical management. In addition to the various medical error types that occur throughout these phases. The most common types are diagnostics errors and clinical management errors. Understanding the patient journey-related phases and types of medical errors is crucial in modeling a tracking system framework. Clinical notes of Electronic Health Records (EHR)

are considered as documented references of the entire patient's medical journey, from the first point of care to post-medical care plan follow-up. Developing technologies that work on processing clinical notes and notifying healthcare providers with real-time medical error signaling and correction will move healthcare to its next level with a new leveraged paradigm in patient safety. MEDIQA-CORR 2024 shared tasks covers three tasks related to medical error detection and correction from clinical notes (Ben Abacha et al., 2024a). This competition is to establish state-of-the-art techniques (SOTA) to formulate a reliable clinical task of this kind. In this paper, we demonstrate our participation in addressing these three challenges. The complexity of designing a tool that addresses all incidents of medical errors from clinical notes comes from the variety of clinical notes' architecture/context. This variation occurs across:

- Different patient care stages (assessment, diagnosis, medical management plan, follow-up) The type of medical error incidents varies based on the phase of a patient's journey. For instance, diagnostic errors are most likely during the assessment or the diagnosis stage. While at the medical management and plan phase, clinical management errors are the most common.
- The level of documented details related to the history of Present Illness (HPI), Past Medical History (PMH), medication history, clinical findings, diagnosis, medical management plans, and follow-up.

Considering the stage of patient care and the level of clinical note details when mapping case scenarios of medical error incidents within the solution framework is a functional step in building a sustainable medical error tracking system. For instance,

detailed documentation of clinical findings and diagnostic tests may facilitate the identification of diagnostic errors during the assessment or diagnosis stage. On the other hand, comprehensive documentation of treatment plans and medication history may aid in identifying medical management errors during the medical management and planning phase.

Our approach to handling this challenge involves designing a framework that consistently addresses the most common two types of medical errors: diagnostics and medical management errors. To detect these types, we propose a tracking algorithm based on classifying the context of clinical notes to map them with their related medical error case scenarios, ending with detecting medical error incidents of one of these two types. To formulate this in the framework, each clinical note should be screened for three case scenarios of medical error incidents. We categorize clinical notes into two levels of contextual architecture. Level 1 (L1) addresses the first case scenario of medical error. While level two (L2) helps us track the second and third case scenarios of medical error instances. The details of these case scenarios and clinical note levels are demonstrated in section 5.

We implied our tracking approach in a GPT-4 multi-stage prompting-based pipeline that ends with the three downstream tasks: classification of medical error existence (Task 1), identification of error location (Task 2), and correction of error (Task 3). The pipeline is composed of four main stages, as illustrated in Figure 1. Throughout the pipeline, we applied clinical Chain of Thought (CoT) and Chain-of-Verification (CoVe) to mitigate the hallucination and enforce the model to reference its reasoning rational response according to its Evidence-Based Medicine (EBM) clinical practice guidelines attributed knowledge of GPT-4 acquired during training.

As a side work, we developed a RAG system injected with clinical practice guidelines as an external knowledge datastore.

2 Background and Related Work

Large language models (LLMs) have proven their potential in various domains, including finance, marketing, and education. Healthcare is a wide area with many horizons (medical education, translational medicine, clinical practice, domain-specific clinical specialty), and the efficiency of LLMs

varies within each horizon area. In some instances, pretrained language models (PLMs) show an efficient performance on specific basic NLP clinical tasks such as Named Entity Recognition (NER), classification, and relationship extraction (RE). However, efficient performance is yet to be reliable and implemented on generative advanced clinical NLP tasks, including clinical text generation, medical question answering, and clinical text summarization. Thus, there is a considerable area for optimizing and leveraging the state-of-the-art in the area of applied generative clinical NLP.

Hallucination and out-of-source generation are some of the main limitations that LLMs and PLMs face, especially with up-to-date and niche-focused domain-related tasks.

RAG and CoT are leading techniques that have been shown to mitigate the limitations mentioned above (Towhidul Islam Tonmoy et al., 2024). RAG framework works on optimizing the output of LLMs by appending LLMs with an external up-to-date knowledge/ data store to be attributed/injected in the generative process through a retrieval and query process (Shuster et al., 2021). RAG contextualizes the model to be more aligned with domain-specific downstream tasks, ending with a more accurate, customized, and specific evidence-grounded response with its data source to be more valid. It encompasses 3 main components: the retrieval, the generation, and the augmentation techniques. Pretrained LLMs performance is comparable with LLMs with RAG. RAG can overcome the need to retrain/finetune LLMs on up-to-date or domain-specific information. Instead, it augments the knowledge with LLMs without retraining the model and results with applicable performance (Gupta et al., 2024).

RAG with LLMs has shown its potential to drastically advance LLMs' usability and reliability. In the healthcare domain, integrating RAG with LLMs has been applied with notable enhancements in the generated responses of LLMs to make them more accurate, informative, and reliable. LLMs output aligned remarkably with the augmented RAG case-specific medical knowledge. (Zakka et al., 2024; Ge et al., 2023) incorporated RAG into LLMs with a medical knowledge database for medical guidelines and treatment recommendations. These LLMs with RAG outperform standard LLMs significantly on the level of accuracy, user satisfaction and consistency. Another study illuminates the impact of appending clinical trials related

to medical knowledge to LLM with RAG on an exceeding performance of this framework compared to experts in clinical trial screening (Unlu et al., 2024).

From the perspective of our shared task, medical error correction is one of its downstream tasks. Which needs for techniques to support and enhance formulating this task. Factual Error Detection and Correction with Evidence Retrieved (FLEEK) (Bayat et al., 2023), is an innovative solution that overcomes hallucinations. It performs two tasks: fact verification and fact revision. It splits an input passage into sentences and uses a sequential pipeline to verify each sentence and correct it so it reduces hallucinations with unstructured knowledge, such as web-based and structured knowledge graphs. Facts are defined as units of information that describe entities, relations, or events and are represented using a semi-structured triple format. FLEEK's performance is evaluated using benchmarks and preliminary experiments using manually created evaluation data. (Dhuliawala et al., 2023) introduced Chain-of-Verification (CoVe), a method to reduce hallucinations in large language models by breaking down verifications into more straightforward questions and self-correcting them. Factored CoVe helps alleviate copying hallucinations and provides performance gains over original responses.

For accurate clinical diagnosis, (Savage et al., 2024) explored LLMs in medicine to imitate the Clinical Reasoning Rationale (CRR) as a COT approach to perform differential diagnosis steps during the medical diagnosis process. They created a diagnostic prompting method that allows LLMs to construct diagnosis while accurately mimicking clinical reasoning using CoT prompts. This led to GPT-4 being prompted to imitate the thought processes of clinicians, giving doctors a comprehensible justification for assessing the precision of LLM replies. These techniques are utilised in order to enhance our model's ability to identify medical errors in the clinical context and consider potential corrections.

RRED (Min et al., 2022) is a deep learning framework designed to detect errors in radiology reports. The system creates artificial existing errors using an error generator and supervised learning techniques.

The method addresses error detection in radiology reports using a deep learning framework with a rich contextual and medical understanding. The error

generator generates realistic errors from existing radiology reports, creating synthesized datasets for training the error detector. The error detector employs a BERT-based architecture to detect errors based on a semantic understanding of radiology reports.

3 Dataset

3.1 MEDIQA-CORR 2024

MEDIQA-CORR 2024 proposed three shared tasks related to medical error detection and correction from clinical notes. Table 1 illustrates the characteristics of the data concerning both the input and output parameters for each task.

Three datasets were provided in this challenge: training, validation, and testing (Ben Abacha et al., 2024b).

3.1.1 Training Dataset

The initial dataset is derived from the University of Washington (MS) Training Set. It comprises 2,189 clinical texts, all of which either have one error or none at all (1 denoting that the text has an error and 0 denoting that there are no errors). This data set includes the original clinical note, the error sentence, the corrected sentence, and the corrected text as a whole. These clinical notes document the patient's related conditions throughout different patient care phases.

3.1.2 Validation Dataset

Two validation datasets: MS validation set contains 574 clinical texts, and the University of Washington (UW) validation set includes 160 clinical texts. Clinical notes are unlabeled in this dataset for validation purposes. However, The labeled notes of the dataset were accessible.

3.1.3 Testing Dataset

This dataset was provided by MEDIQA-CORR 2024 and contains 574 clinical texts, which include only the sentences without flagging the error or correcting it. It will serve as a means of testing our model to determine its performance for tasks 1, 2, and 3.

3.2 Clinical Guidelines Dataset

For clinical knowledge enhancement, we utilized the Clinical Guidelines corpus dataset (Chen et al., 2023), comprising 47,000 clinical practice guidelines sourced from 17 reputable online medical references. We utilized this dataset as a data store to

Table 1: Example of the input and the structured output format for tasks 1, 2, and 3.

Input	<p>0 A 9-year-old girl is brought to the pediatrician by her mother who reports that the girl has been complaining of genital itching over the past few days.</p> <p>1 She states she has noticed her daughter scratching her buttocks and anus for the past week; however, now she is scratching her groin quite profusely as well.</p> <p>2 The mother notices that symptoms seem to be worse at night.</p> <p>3 The girl is otherwise healthy, is up to date on her vaccinations, and feels well.</p> <p>4 She was recently treated with amoxicillin for a middle ear infection.</p> <p>5 The child also had a recent bout of diarrhea that was profuse and watery that seems to be improving.</p> <p>6 Her temperature is 98.5 F (36.9 C), blood pressure is 111/70</p> <p>7 mmHg, pulse is</p> <p>8 83/min, respirations are 16/min, and oxygen saturation is 98% on room air.</p> <p>9 Physical exam is notable for excoriations over the girl’s anus and near her vagina.</p> <p>10 Suspected of infection with Giardia lamblia.</p>
--------------	---

Task	Output	
1: Error Flag	1	
2: Error Sentence	10 Suspected of infection with Giardia lamblia.	
3: Corrected Sentence, Corrected Text	<p>Corrected Sentence: Suspected of infection with Enterobius vermicularis.</p>	<p>Corrected Text: A 9-year-old girl is brought to the pediatrician by her mother, who reports that the girl has been complaining of genital itching over the past few days.</p> <p>....</p> <p>Suspected of infection with Enterobius vermicularis.</p>

be augmented with LLM by the RAG framework to enhance the clinical practice domain of knowledge.

4 Description of the Three Tasks

In this research, we worked on three tasks; the description of each task is as follows:

Binary Classification (Detecting Medical Errors): In this task, we had to determine whether the text contained a medical error or not. This task involved binary classification (0/1) regarding the existence of the medical error in the text.

Span Identification (Locating Errors within Text): In the second task, if there is a medical error in the given clinical text, the model should identify the precise text span linked to it. The exact location or the section where the error is found in

the medical text.

Natural Language Generation (Correction of Errors): In the last task, the model must provide a free text correction where the medical error is detected in the clinical text. This task aims to improve the quality and accuracy of mistake resolution in clinical situations by producing a human-like context to augment the automated correction process.

5 Methods

We build a GPT-4 prompting-based pipeline that processes the text of clinical notes and endeavors to detect medical errors and correct them if they exist. The model pipeline addresses medical error incidents using three types of clinical note context case scenarios. These case scenarios are classi-

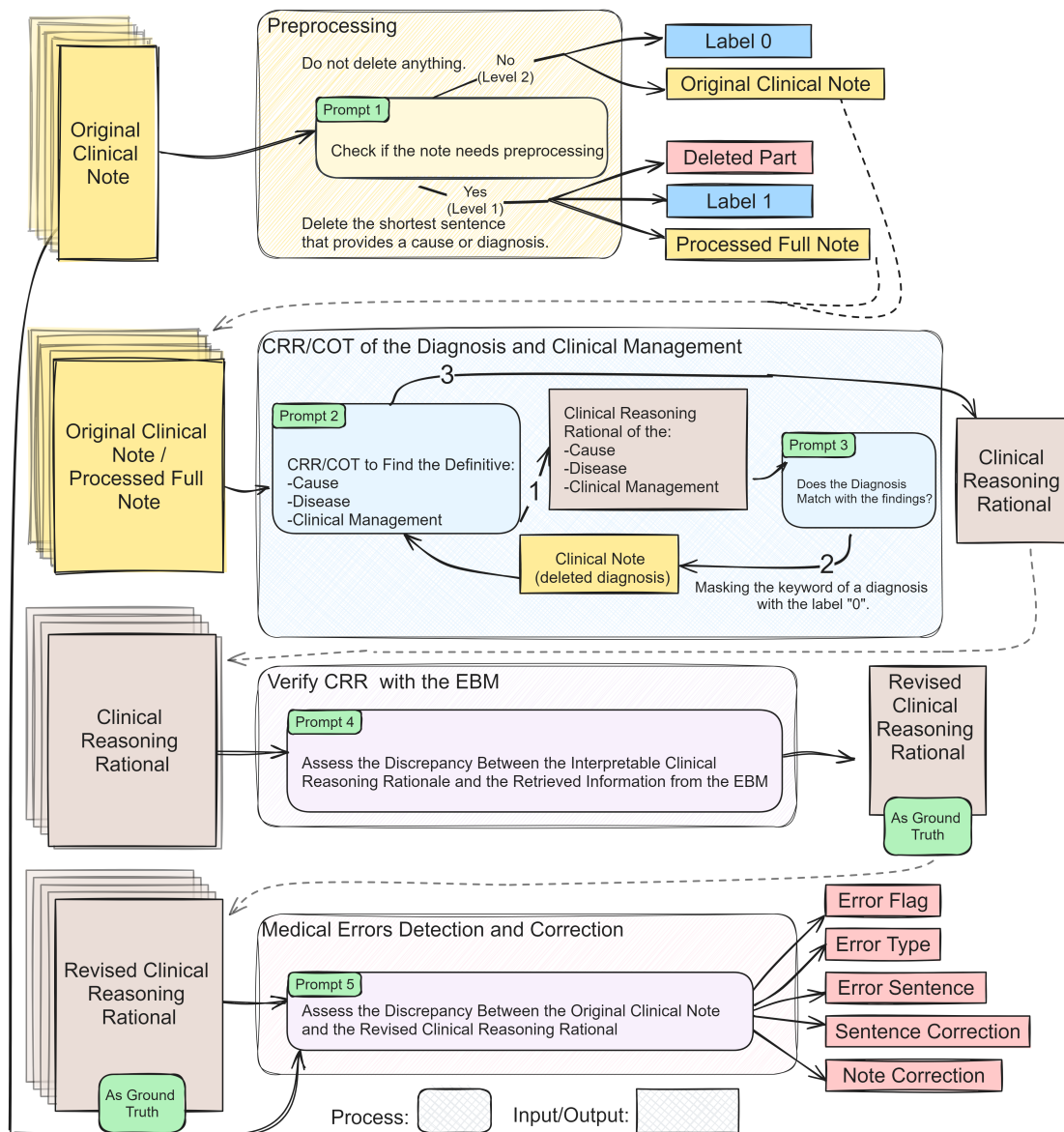


Figure 1: GPT-4 Multi-stage prompting pipeline. The pipeline comprises four phases, each with certain prompts and processes. The data flow process throughout the pipeline with detailed inputs and outputs is depicted in the figure. CRR: Clinical Reasoning Rationale; COT: Chain of Thought; EBM: Evidence-Based Medicine.

fied according to the types of medical errors we aim to track, including diagnostic and clinical management errors. **Diagnostic errors** occur when a patient’s medical condition is attributed to an incorrect cause (pathogen, poison, etc.) or misdiagnosis. While **medical management types of errors** involve an incorrect medication, lab test, procedure, or medical image for a patient’s medical condition. These errors occur due to an incorrect interpretation of findings to the correct cause and diagnosis or an incorrect interpretation of the cause and diagnosis to the most appropriate medical management. These consequences of incorrect interpretations that lead to medical errors are addressed

through our solving approach, which is systematically employed throughout the pipeline to track incidents of these types of medical errors. In a brief overview, the pipeline involves removing the cause or diagnosis from the clinical note and subsequently generating a CoT process. This process is tailored to find the most probable cause, findings, and clinical management of a patient’s clinical case. Then, assess for discrepancy between the generated CoT and original clinical note, thus indicating if an error exists and correcting it. The output of each prompt is formulated to be unified across all dataset. The pipeline is composed of four main processing phases. In the following subsec-

tions 5.1, 5.2, 5.3, 5.4, we elucidate these phases with the details of the techniques used within each phase.

Clinical notes can be articulated in various architectures, depending on the timeline and stages of the patient’s medical journey. We categorized clinical notes into two levels of contextual architecture. Level one (L1) addresses the first case scenario of medical error, while level two (L2) helps us track the second and third case scenarios of medical error instances. This facilitates formulating a clear link between the context of clinical note and the most suspected medial error case scenario to be signaled as the following:

L1: Documented HPI or PMH without clinical management actions related to the cause or the diagnosis. At this level, the clinical management would be directly moved to detect the suspected cause and diagnosis of the medical condition upon the existing clinical findings without thinking of any further needed medical management actions to confirm the diagnosis or to manage the current patient’s medical status.

L2: Documented HPI or PMH with clinical management actions related to the cause or the diagnosis. At this level, the patient is already known with the diagnosis and related cause. A wider margin of settings. This level will track the second and third case scenarios of medical error instances. Medical errors at this level mostly will be related to signal the second case scenario. The third case scenario is related to the previous step of the patient medical journey, at the diagnosis step. Specifically when the diagnosis and clinical findings are not directly connected to each other in most common clinical contexts, yet the note contains a clinical management action related to the incorrect patient’s diagnosis. Refer to Figure 2 of Appendix B that demonstrates an example of each case scenario.

5.1 Preprocessing

In order to orient the model to the downstream tasks of detecting medical errors and correcting them, the model is firstly promoted to detect the shortest part of a sentence that declares the cause or the diagnosis of a medical condition in the clinical note in order to be masked. The rationale behind this masking is to force the generated CoT (in the next phase) without being biased by the already declared cause or diagnosis in the clinical note. This process is applied to L1 clinical notes to help track the first case scenario of medical error incidents. With L1

clinical notes, the model will return the “Deleted Part” and the “Processed Full Note”. While L2 type clinical notes should be returned as they are. As shown in Figure 1. “Processed Full Note” and L2 type of clinical notes are designed to be passed to the next stage. The prompt, “Prompt1” of this stage is shown in Table 3 of Appendix A.

5.2 Clinical Reasoning Rationale/CoT of the Diagnosis and Clinical Management

This stage is tailored to return the cause, diagnosis, and medical management of each clinical note returned from the previous step. These returns are based on CRR. CRR is a CoT technique applied in the clinical context. Two prompts are used at this stage. The first one, “Prompt 2”, as shown in Figure 1. We engineered the prompt to do step by step deduction to create a differential diagnosis from which to find the most likely cause and diagnosis of medical condition in a clinical note. The answer is constrained to the documented clinical findings of the clinical note, directing the model to be more definitive to the most probable correct cause and diagnosis without expanding the probability of other differential diagnoses based on further clinical investigations actions beyond what the note handles. Subsequently, upon the most likely cause and diagnosis of a medical condition, the model deduces the most correct clinical management (treatment, clinical care plan, intervention, procedure...etc.) using a step-by-step process. “Prompt 2” is demonstrated in Table 3 of Appendix A. The “Clinical Reasoning Rationale” output from “Prompt 2” is then employed to be used as a reference for the next prompt, “Prompt 3”. “Prompt 3” serves as a checkpoint of the third case scenario of medical error instances. If the documented diagnosis is based on clinical findings that are not directly related to each other in the common clinical context, it indicates a diagnostic type of medical error that needs to be processed. The model at this prompt is designed to process any discrepancy between the clinical finding and the diagnosis by marking the keyword of the incorrect diagnosis with the label "0". Then, the processed note will be passed again to “Prompt 2” to find the correct cause, diagnosis, and clinical management. The final “Clinical Reasoning Rationale” output at this phase will be passed to the next phase, “Verify CRR with the EBM”. “Prompt 3” is demonstrated in Table 3 of Appendix A.

5.3 Verify CRR with the EBM

The functionality of this stage is to verify the baseline interpretable “Clinical Reasoning Rationale” output by instructing the model to generate questions that target and retrieve each information in the CRR note, then correct any discrepancy, following (Bayat et al., 2023; Dhuliawala et al., 2023). The prompt of this phase, “Prompt 4”, demonstrated in Table 4 of Appendix A, helps in forcing the model to reference EBM clinical practice guidelines attributed knowledge of GPT-4 acquired during training. The output, verified (CRR), will be taken as ground truth knowledge for the next stage, “Medical Errors Detection and Correction”.

5.4 Medical Errors Detection and Correction

This final stage, “Prompt 5”, as shown in Table 5 of Appendix A, is designed to be the cut-point step for the three tasks. The verified CRR is taken as ground truth knowledge for clinical notes. It should include the correct cause, diagnosis, and clinical management for the note. The model is instructed to compare the verified CRR with the original clinical note for cause, diagnosis, and clinical management discrepancies, as shown in Figure 1. Clinical notes with the contextual architecture of L1, discrepancies related to the cause or diagnosis should be cached. For L2, discrepancies related to the cause or diagnosis should be cached as well, along with clinical management discrepancies. Clinical management is a wide aspect, including interventions related to treatment, ordering certain lab tests and images, transfer, and procedure. The CRR includes all the necessary clinical management actions related to the clinical note case, while the original note might include one of them. This case scenario might drive the model to detect it as a discrepancy, correcting it with the appropriate completed clinical management plan. Additionally, if a diagnostic error exists, it should be corrected, ending with two medical errors identified. For L2, since the clinical note contains only one error. The model is instructed to prioritize correction for diagnostics errors (cause and disease). Then, to clinical management-related errors. From this phase, we should have the "Error Flag" for task 1, the "Error Sentence" for task 2, the "Sentence Correction" for task 3, and the "Note Correction" as a full note.

5.5 RAG

To enhance the accuracy and relevance of the generated clinical response while throughout the processes related to CRR/CoT and CoVe, our approach conducted an experiment using a RAG framework. Our RAG system is based on the parameters outlined in Table 2.

Clinical Guidelines corpus dataset is utilized as an external knowledge database. Our RAG system should integrate this knowledge into the prompt output through the query and retrieval process. The process of generating the query is based on the instructions stated in “Prompt 4”. Where the RAG system is utilized as the ground truth for this stage.

6 Experiments and Results

Here, we report the experimental findings demonstrating our model’s effectiveness on the shared tasks. Our model was performed on the three tasks utilizing a zero-shot learning approach and a GPT-4 prompting-based pipeline with CoT and CoVe methods and structured output.

Since our approach mainly focuses on zero-shot learning. Thus, we only used the training and validation for prompt optimization until we reached a reasonable output in tracking the three case scenarios. For a comprehensive show of our approach’s functionality in tracking the three case scenarios, please refer to Appendix B. It provides an example experiment of tracking each one of the case scenarios, illustrating its input/output at each prompt through the entire process.

The Function Calling feature of OpenAI API and JSON mode is utilized to get the aimed structured format and ensure consistent output throughout the dataset.

The results of our approach performance on the testing dataset show that the accuracy of the first and second tasks is 0.566 and 0.49, respectively, without using external knowledge sources, fine-tuning methods, or group learning. While for task 3 and the main results, performance metrics yielded the following scores: ROUGE-1-F of 0.427, BERTScore of 0.48, BLEURT of 0.53, their Aggregate-Score (Mean of ROUGE-1-F, BERTScore, BLEURT-20) of 0.48, and their Composite Scores of 0.34. These metrics assess the model’s ability to produce contextually appropriate corrections for clinical errors identified in clinical text.

Table 2: RAG parameters.

Parameter	Value
Chunk_Size	500
Chunk_Overlap	32
Embedding Model	Bio_ClinicalBERT
Embedding Dimension	768
Model Pooling Strategy	Mean
Vector Index	Faiss Hierarchical Navigable Small Worlds Index with Neighboring Vectors of 32
Chunking Strategy	Recursive Character Text Splitter from langchain

7 Discussion and Future Scope

The performance of our model is somehow acceptable but unreliable when applied to the testing dataset. The algorithmic approach of following three case scenarios of medical events based on clinical note contextual architecture might undertrack other medical error incidents case scenarios. For the LLM we have used, GPT-4, we have aimed to be built based on a RAG framework incorporated with clinical practice guidelines. Our hypothesis was to optimize the output of an pre-trained generative model in the arena of clinical practice (niche-focused) to get reliable, inferential and ground truth knowledge without hallucinations. RAG framework is the best to be employed with a massive LLM such as GPT, BART, or T5. For limited hardware resources, we took GPT-4, as an open LLM model. Our RAG system failed in retrieving relevant queries. It was supposed to be connected to our pipeline at phase 4, "Verify CRR-CoT with the EBM", but for irrelevant retrieved chunks, we continued the work without it. This shortcoming performance could be one of the following:

- The vector embedding model we used, Bio_ClinicalBER, is not one of the vector embedding models that are already designed for the RAG frameworks (trained with a retrieval objective). The choice of Bio_ClinicalBERT was to test a clinical embedding model rather than general used ones. In addition to a limited time, we could not test the SOTA models with our RAG such as ColBERT (Khattab and Zaharia, 2020).
- The chunking strategy we used is a naive technique, which might be the cause of the poor informative chunks.
- The complicated structure of the utilized Clinical Guidelines dataset as an external datastore

for our RAG system.

Along the pipeline, we used CRR-CoT and CoVe as prompt optimization techniques. Future work should investigate the performance of public LLMs, to unlock their known capabilities for these downstream tasks within hardware accessible facilities. In addition to exploring advanced chunking strategies such as semantic chunking and finetuning a domain-specific model such as Bio_ClinicalBERT for retrieval.

8 Limitations

Our work is limited by two points. The first one is our inability to produce a reliable RAG system due to time constraints. We could not explore how our approach would perform with a successful RAG system implementation; specifically, we utilized rich, niche-focused external knowledge to boost the reliability and applicability of the generated output. Secondly, with our limited computing and financial capacity, we would not be able to experiment with other massive LLMS, whether they are general, clinically fine-tuned, or pre-trained ones.

Acknowledgments

The research was financially supported by Princess Sumaya University for Technology (PSUT). We sincerely thank our university, PSUT, for their assistance in supporting this study.

References

- Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyi, Samira Khorshidi, Fei Wu, Ihab F Ilyas, and Yunyao Li. 2023. [Fleek : Factual error detection and correction with evidence retrieved](#). *arXiv preprint arXiv:2310.17119*.
- Asma Ben Abacha, Wen wai Yim, Velvin Fu, Zhaoyi Sun, Fei Xia, and Meliha Yetisgen. 2024a. Overview

- of the medqa-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*.
- Zeming Chen, Alejandro Hernández, Cano Angelika, Romanou Antoine, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, and Vinitra Swamy. 2023. [MED-ITRON -70B: Scaling medical pretraining for large language models](#). *arXiv preprint arXiv:2311.16079*.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. [Chain-of-Verification reduces hallucination in large language models](#). *arXiv preprint arXiv:2309.11495*.
- Jin Ge, Steve Sun, Joseph Owens, Victor Galvez, Ok-sana Gologorskaya, Jennifer C Lai, Mark J Pletcher, and Ki Lai. 2023. Development of a liver disease-specific large language model chat interface using retrieval augmented generation. *medRxiv*.
- Aman Gupta, Anup Shirgaonkar, Angels de Luis Balaguer, Bruno Silva, Daniel Holstein, Dawei Li, Jennifer Marsman, Leonardo O Nunes, Mahsa Rouzbahman, and Morris Sharp. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*.
- Omar Khattab and Matei Zaharia. 2020. [Colbert : Efficient and effective passage search via contextualized late interaction over bert](#). In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.
- Dabin Min, Kaeun Kim, Jong Hyuk Lee, Yisak Kim, and Chang Min Park. 2022. Rred: A radiology report error detector based on deep learning framework. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 41–52.
- Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. [Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine](#). *NPJ Digital Medicine*, 7(1):20.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.
- S M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv e-prints*, pages arXiv–2401.
- Ozan Unlu, Jiyeon Shin, Charlotte J Maily, Michael F Oates, Michela R Tucci, Matthew Varugheese, Kavishwar Waghlikar, Fei Wang, Benjamin M Scirica, and Alexander J Blood. 2024. Retrieval augmented generation enabled generative pre-trained transformer 4 (GPT-4) performance for clinical trial screening. *medRxiv*, pages 2002–2024.
- Cyril Zarka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, and Euan Ashley. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):A10a2300068.

A The Pipeline’s Prompts Templates

The instruction template of each prompt throughout the pipeline is demonstrated in Tables 3, 4, and 5.

B The Output of the Pipeline

The model’s performance at each stage is demonstrated with an example for each of the three case scenarios. Figure 2 displays an example for each case scenario, including the original note as an input to the pipeline and the corrected sentence. The outputs throughout the pipeline’s phases are illustrated as follows: at the Preprocessing B.1, Clinical Reasoning Rationale/CoT of the Diagnosis and Clinical Management B.2, Verify CRR with the EBM B.3, and Medical Errors Detection and Correction B.4 stage.

B.1 Preprocessing

Figure 3 illustrates the outputs at stage one, the Preprocessing, for the first case scenario medical errors that are in particular related to clinical notes with level one context. Figures 7 and 11, represent the second and third case scenarios, respectively, both of which are in particular related to clinical notes with level two context.

B.2 Clinical Reasoning Rationale/CoT of the Diagnosis and Clinical Management

The outputs of this phase utilizing Prompts 2 and 3 are depicted in Figures 4, 8, and 12 for the three case scenarios.

B.3 Verify CRR with the EBM

At stage 3, the verified CRR output of clinical notes corresponding to the first, second, and third case scenarios is shown in Figures 5, 9, and 13, respectively.

Table 3: The templates of Prompts 1,2, and 3.

Prompt	Instruction Template
1	<p>I will give you a clinical note. You have to delete the shortest sentence that shows the cause or diagnosis, following to these conditions:</p> <ol style="list-style-type: none"> 1) If the clinical note mentions any of clinical management actions (treatment, clinical care plan, or any intervention,...etc.) related to (management of past medical history, management history of present illness, diagnosis), then do not delete anything. Give this label 0. 2) Else, then delete the sentence that shows the cause and diagnosis. Give this label 1 3) Print the assigned labels 1 or 0. 4) Print the deleted part if applicable. 5) Print the full final note.
2	<ol style="list-style-type: none"> 1) Based on Evidence-Based Medicine, use step-by-step deduction to create a differential diagnosis and then use step by step deduction to identify both of the most likely causing (Pathogen name of the bacteria, worm, virus, fungi,...etc., poison,... etc.) and diagnosis separately. The answer should also be definitive to one cause and one diagnosis without requiring any further clinical investigating action. 2) Then, step by step, deduce the most correct (treatment, clinical care plan, clinical management, intervention) <p>You are designed to output JSON. The JSON should be structured like this:</p> <pre>{ "Differential Diagnosis Step by Step": { "Step 1": ..., "Step 2": ..., "Step N": ... }, "Differential Diagnosis": { "Most Likely Cause": ..., "Explanation": ... }, "Treatment Step by Step": { "Step 1": ..., "Step 2": ..., "Step N": ... }, "Definitive Diagnosis": ..., "Treatment": { "Definitive Treatment": ... } }</pre>
3	<ol style="list-style-type: none"> 1) Use this interpretable clinical reasoning rationale you have produced for this clinical note: cot 2) Based on the interpretable clinical reasoning rationale, If the clinical note mentions a diagnosis or a medical condition that is based on a clinical presentation or findings that are not directly connected to each other in most common clinical contexts, then there should be a medical error in the diagnosis. 3) Delete the diagnosis or a medical condition-related keyword from the clinical note. 4) Print the deleted keyword if applicable. 5) Print the full final note, where the deleted keyword should be masked with this label -> "0" <p>You are designed to output JSON. has to be structured like this:</p> <pre>{{ "DeletedKeyword": ..., "FullFinalNote": ... }}</pre>

Table 4: The template of Prompt 4.

Prompt	Instruction Template
4	<p>You have to verify your interpretable clinical reasoning rationale of the diagnosis you have produced of its related clinical note. The verification should be done by generating questions that target and retrieve information from the most appropriate clinical practice guidelines.</p> <ul style="list-style-type: none"> -Make the query address the name of the guideline you want to retrieve that response from. -If you want to check for the diagnosis of clinical findings, make the query address the related clinical findings you want to check for the diagnosis. -Make the directed query address the most likely correct (cause, diagnosis). -Make the directed query address the recommendations part of the guideline related to (diagnosis, clinical management, treatment, drug of choice) -Search from the directed guidelines. -Return the information you gained. -Compare your interpretable clinical reasoning rationale with the retrieved information from the guideline; if there is a discrepancy, show it. -If there is a major discrepancy, take the retrieved information as ground truth and print out the final CoT after being revised. <p>You are designed to output JSON. It has to be structured like this:</p> <pre> {{ "VerificationQueries": { "Query 1": ..., "Query 2": ..., "Query 3": ..., "Query N": ... }, "RetrievedInformation": { "Response 1": ..., "Response 3": ..., "Response N": ... }, "Comparison": { "Clinical Findings": ..., "Causes": ..., "Treatment": ... }, "Discrepancy": ... (could be nullable), "FinalCoT": { "Differential Diagnosis Process": { "Step 1": ..., "Step 2": ..., "Step 3": ..., "Step N": ... }, "Definitive Cause": { "Most Likely Pathogen/Cause": ... }, "Definitive Diagnosis": ..., "Treatment Plan": { "Step 1": ..., "Step 2": ..., "Step 3": ..., "Step 4": ..., "Step N": ... } } }} </pre>

Table 5: The template of Prompt 5.

Prompt	Instruction Template
5	<p>1) Use this interpretable clinical reasoning rationale you have produced as a ground truth {verified_cot}</p> <p>2) Compare if the clinical note matches the ground truth to tell if the clinical note has a medical error in (diagnosis (pathogen, poison, disease), clinical management (treatment, clinical care plan, intervention (order certain lab test, transfer, certain image by name, procedure)).).</p> <p>3) Identify any discrepancy between the ground truth and the clinical note.</p> <p>4)Then, if there is anything in the clinical note related to either diagnosis or cause that is not available (referenced) in the ground truth reference, then label it as a medical error. And skip the steps related to clinical management.</p> <p>5)Then else, if there is anything in the clinical note related to clinical management after diagnosis that is not available (referenced) in the ground truth reference, specifically in (clinical management-related sections), then label it as a medical error. And skip the steps related to the diagnosis or cause. If there is a medical error, identify its type (diagnosis, cause, or clinical management) and print it, identify the exact related shortest part and print it, and correct it with the shortest possible correction. Do not change the format of the corrected part. Only correct the related keyword. Then, if the error type is related to clinical management-related errors, the corrected sentence should be definitive to the exact needed medication, procedure, image,..... etc., not general. Not as a recommendation. Correct the note directly with the most correct probable needed audit. If the error type is related to diagnosis, cause, or clinical management, consider this error correction to be edited on the final corrected note. The priority to add the correction of diagnosis and cause first to be considered. Consider one correction only, depending on the context. Finally print out the corrected final note.</p> <p>The clinical note you have to correct is split into sentences with an index for each. The correction you return includes the error flag, the error location, and the sentence correction.</p>

B.4 Medical Errors Detection and Correction

At stage 4, using Prompt 5. The final outputs include the "error flag" to identify if an error exists, "error location", and "sentence correction" if there is an error within a clinical note. See Figures 6, 10, and 14 corresponding to the first, second, and third case scenarios, respectively.

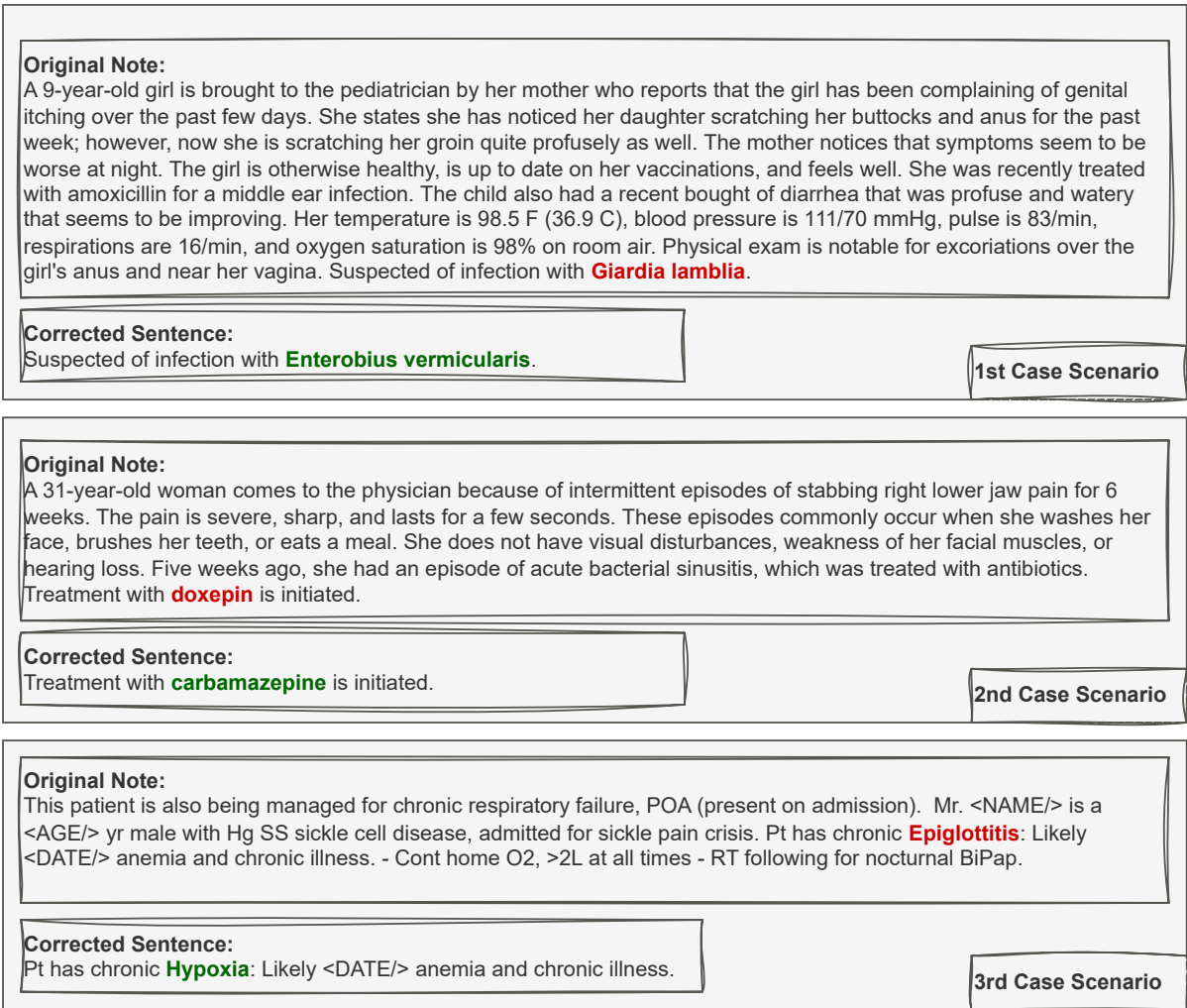


Figure 2: Example for each type of case scenario.

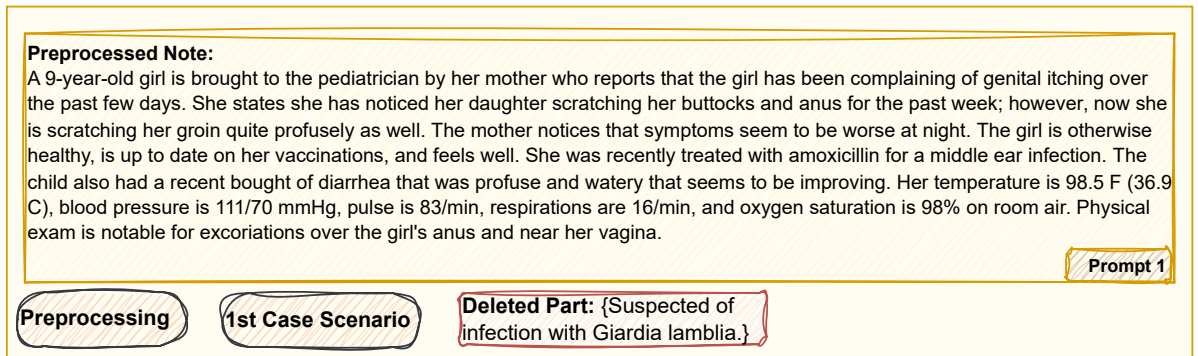


Figure 3: The outputs of clinical notes with the first case scenario, level one notes. At stage 1, the Preprocessing, using Prompt 1.

CRR Round 1:
{Differential Diagnosis Step by Step': **{Step 1':** 'Begin by considering common causes of genital itching in children, including irritant contact dermatitis, infections (bacterial, viral, fungal, parasitic), and infestations.'
'Step 2': 'Note the symptom of worsening itching at night, which is a characteristic feature of certain parasitic infections.'
'Step 3': 'Acknowledge the recent use of antibiotics (amoxicillin) and a bout of diarrhea, which could predispose to opportunistic infections or alter normal flora, but these are less likely to cause localized itching primarily at night.'
'Step 4': 'Consider the physical exam findings of excoriations over the genital and anal area, which suggest an infestation rather than a generalized dermatitis or systemic illness.'
'Step 5': 'Given the symptom of itching worsening at night and the specific locations of the excoriations, focus on parasitic infestations that exhibit nocturnal activity.'
'Differential Diagnosis': **{Most Likely Cause':** 'Enterobius vermicularis (Pinworm) Infestation',
'Explanation': 'The symptoms of nocturnal itching, particularly around the anus, and the presence of excoriations in a child, are classic for pinworm infestation. Pinworms are a common cause of genital itching in children and are known for increased activity and egg-laying at night, which markedly exacerbates symptoms during these hours.'
'Treatment Step by Step': **{Step 1':** 'Confirm the diagnosis if possible, typically by using the tape test to collect eggs for microscopic identification. However, given the classic presentation, empirical treatment may be initiated without confirmation.'
'Step 2': 'Prescribe a single dose of albendazole or mebendazole, with a second dose to be taken in two weeks to eradicate any subsequent hatchlings.'
'Step 3': 'Educate the family on the importance of strict hygiene measures to avoid reinfection. This includes frequent handwashing, especially after using the bathroom and before eating, daily showering, changing underwear and bed linens frequently, and avoiding nail-biting.'
'Step 4': 'Advise treatment of all family members living in the household to prevent reinfection or spread, as pinworms are highly contagious.'
'Definitive Diagnosis': 'Enterobiasis (Pinworm Infestation)',
'Treatment': **{Definitive Treatment':** 'Albendazole or mebendazole, with a repeat dose in 2 weeks, coupled with stringent personal and household hygiene measures.}}

Prompt 2

Clinical Note (deleted diagnosis with 0): A 9-year-old girl is brought to the pediatrician by her mother who reports that the girl has been complaining of genital itching over the past few days. She states she has noticed her daughter scratching her buttocks and anus for the past week; however, now she is scratching her groin quite profusely as well. The mother notices that symptoms seem to be worse at night. The girl is otherwise healthy, is up to date on her vaccinations, and feels well. She was recently treated with amoxicillin for a middle ear infection. The child also had a recent bout of diarrhea that was profuse and watery that seems to be improving. Her temperature is 98.5 F (36.9 C), blood pressure is 111/70 mmHg, pulse is 83/min, respirations are 16/min, and oxygen saturation is 98% on room air. Physical exam is notable for excoriations over the girl's anus and near her vagina. Suspected of infection with **0**.

Prompt 3

CRR Round 2:
{Differential Diagnosis Step by Step': **{Step 1':** 'Given the symptoms of genital itching, particularly noted to be worse at night, and the presence of excoriations over the anus and near the vagina, parasitic infections are highly suspected.'
'Step 2': 'The recent history of diarrhea could suggest a gastrointestinal infection or infestation, which may be related to the itching if the pathogen has a cutaneous manifestation or provokes such a reaction.'
'Step 3': 'The recent use of amoxicillin for an ear infection introduces the possibility of a secondary yeast infection due to antibiotic use; however, the description of symptoms being worse at night and the specific location of itching are less typical for a yeast infection.'
'Step 4': 'Considering the girl's age and the specific symptoms described, common infections of the perianal region in children should be considered, such as pinworm (Enterobius vermicularis) infestation.'
'Differential Diagnosis': **{Most Likely Cause':** 'Enterobius vermicularis (pinworm) infestation',
'Explanation': 'The nocturnal itching described, along with the location of the excoriations, strongly suggests Enterobius vermicularis as the causative agent. Pinworm eggs are deposited around the anus by the adult female worm, causing intense itching, especially at night. The history of improving diarrhea could be coincidental or related to the stress on the child's gastrointestinal system.'
'Treatment Step by Step': **{Step 1':** 'Confirm the diagnosis by identifying pinworm eggs. This is typically done with the 'tape test,' where clear adhesive tape is applied to the perianal area in the morning before washing or using the toilet, then examined microscopically for eggs.'
'Step 2': 'Treat with a specific antiparasitic medication. The first-line treatment for pinworm infestation is either albendazole or mebendazole, given as a single dose and then repeated in two weeks to prevent reinfection.'
'Step 3': 'Educate the family on hygiene measures to prevent reinfection and the spread to other family members. This includes washing hands and nails thoroughly, especially after using the toilet and before eating; changing underwear, bed linen, and towels daily; and avoiding scratching the perianal area.'
'Step 4': 'Consider treating all household members simultaneously to prevent reinfection, given the high contagion rate among individuals living in close quarters.'
'Definitive Diagnosis': 'Enterobius vermicularis (pinworm) infestation',
'Treatment': **{Definitive Treatment':** 'Treatment consists of administering a single dose of albendazole or mebendazole, repeated in two weeks. Alongside, stringent personal and household hygiene practices should be emphasized to prevent reinfection and spread.}}

Prompt 2/ 2nd time

CRR of Diagnosis and Clinical Management

1st Case Scenario

Figure 4: The outputs of clinical notes with the first case scenario, level one notes. At phase 2, The CRR of Diagnosis and Clinical Management, using Prompts 2 and 3.

Verified CRR:

{'Verification Queries': {'Query 1': 'What are the clinical practice guidelines for diagnosing Enterobius vermicularis (pinworm) infestation?', 'Query 2': 'What are the first-line treatments for Enterobius vermicularis (pinworm) infestation according to current clinical guidelines?', 'Query 3': 'What hygiene measures are recommended in clinical practice guidelines to prevent pinworm reinfection and spread?', 'Query 4': 'Is the 'tape test' recommended by clinical practice guidelines for the diagnosis of pinworm infestation?', 'Query 5': 'Are there any guidelines that recommend treating all household members when one person is diagnosed with pinworm infestation?'}, 'Retrieved Information': {'Response 1': 'Clinical practice guidelines recommend the diagnosis of Enterobius vermicularis (pinworm) infestation to be confirmed through visualization of worms or by identifying eggs using the 'tape test'.', 'Response 2': 'Guidelines confirm that the first-line treatment for pinworm infestation is either albendazole or mebendazole, given as a single dose and repeated in two weeks.', 'Response 3': 'To prevent reinfection and spread, guidelines recommend stringent hygiene measures including washing hands, changing bed linens and underwear frequently, and maintaining clean living environments.', 'Response 4': 'Yes, the 'tape test' is specifically recommended for diagnosing Enterobius vermicularis infestation.', 'Response 5': 'Clinical practice guidelines do recommend considering the treatment of all household members simultaneously to prevent reinfection, reflecting the high contagion rate of pinworm among individuals living in close contact.'}, 'Comparison': {'Clinical Findings': 'The clinical findings and the use of the 'tape test' for diagnosis align with clinical practice guidelines.', 'Causes': 'The correlations between symptoms and Enterobius vermicularis as the cause are consistent with the expected clinical presentation.', 'Treatment': 'The treatment plan including the use of albendazole or mebendazole and the emphasis on hygiene measures, matches the recommendations from the clinical practice guidelines.'}, 'Discrepancy': None, 'Final CoT': {'Differential Diagnosis Process': {'Step 1': 'Given the symptoms of genital itching, particularly noted to be worse at night, and the presence of excoriations over the anus and near the vagina, parasitic infections are highly suspected.', 'Step 2': 'The recent history of diarrhea could suggest a gastrointestinal infection or infestation, which may be related to the itching if the pathogen has a cutaneous manifestation or provokes such a reaction.', 'Step 3': 'The recent use of amoxicillin for an ear infection introduces the possibility of a secondary yeast infection due to antibiotic use; however, the description of symptoms being worse at night and the specific location of itching are less typical for a yeast infection.', 'Step 4': 'Considering the girl's age and the specific symptoms described, common infections of the perianal region in children should be considered, such as pinworm (Enterobius vermicularis) infestation.'}, 'Definitive Cause': {'Most Likely Pathogen/Cause': 'Enterobius vermicularis (pinworm) infestation'}, 'Definitive Diagnosis': 'Enterobius vermicularis (pinworm) infestation', 'Treatment Plan': {'Step 1': 'Confirm the diagnosis by identifying pinworm eggs with the 'tape test'.', 'Step 2': 'Treat with a specific antiparasitic medication, either albendazole or mebendazole, given as a single dose and then repeated in two weeks.', 'Step 3': 'Educate the family on hygiene measures to prevent reinfection and the spread to other family members.', 'Step 4': 'Consider treating all household members simultaneously to prevent reinfection.'}}}

Prompt 4

CRR of Diagnosis and Clinical Management 1st Case Scenario

Figure 5: The verified CRR output of clinical notes, corresponding to the first case scenario and level one notes. At stage 3, the CRR of Diagnosis and Clinical Management, using Prompt 4. The Final CoT is the verified CRR. CRR; Clinical Reasoning Rationale

Final Note:

{'error_flag':1, 'error_location':10, 'sentence_correction':"Suspected of Enterobius vermicularis (pinworm) infestation."}

Prompt 5

Medical Errors Detection and Correction 1st Case Scenario

Figure 6: The last structured output includes: "error flag" to identify if an error exists. And "error location" and "sentence correction" if an error within a clinical note corresponds to the first case scenario and level one notes. At stage 4, the Medical Errors Detection and Correction, using Prompt 5.

Preprocessed Note:

A 31-year-old woman comes to the physician because of intermittent episodes of stabbing right lower jaw pain for 6 weeks. The pain is severe, sharp, and lasts for a few seconds. These episodes commonly occur when she washes her face, brushes her teeth, or eats a meal. She does not have visual disturbances, weakness of her facial muscles, or hearing loss. Five weeks ago, she had an episode of acute bacterial sinusitis, which was treated with antibiotics. Treatment with doxepin is initiated.

Prompt 1

Preprocessing 2nd Case Scenario Deleted Part: {}

Figure 7: The outputs of clinical notes with the second case scenario, level two clinical notes. At stage 1, the Preprocessing, using Prompt 1.

CRR Round 1:

{**Differential Diagnosis Step by Step**: {**Step 1**: 'Given the location of the pain (right lower jaw) and the trigger factors (washing face, brushing teeth, eating), conditions involving the trigeminal nerve should be considered, as well as dental causes.', **Step 2**: 'The absence of visual disturbances, weakness of facial muscles, or hearing loss rules out conditions that involve other cranial nerves or central nervous system disorders.', **Step 3**: 'The recent history of acute bacterial sinusitis suggests a possible link between sinus inflammation and the jaw pain, but the nature of the pain described is not typical for sinusitis radiating pain.', **Step 4**: 'The nature of the pain (severe, sharp, and lasts for a few seconds) is highly suggestive of Trigeminal neuralgia, a disorder of the trigeminal nerve.', **Step 5**: 'Considering the treatment initiated (doxepin), which is an antidepressant used off-label for chronic pain management, further supports the neuralgia hypothesis over other dental or sinus-related causes.'},

Differential Diagnosis: {**Most Likely Cause**: 'Trigeminal Neuralgia',

Explanation: "The patient's symptoms of intermittent episodes of stabbing pain in the right lower jaw triggered by facial movements or activities, along with her recent sinusitis episode and the absence of other neurological signs, strongly suggest Trigeminal Neuralgia. This condition involves the trigeminal nerve and is characterized by sudden, severe, brief, stabbing recurrent episodes of facial pain in one or more divisions of the nerve."},

Treatment Step by Step: {**Step 1**: 'Confirm the diagnosis of Trigeminal Neuralgia through a detailed patient history and physical examination to rule out other causes of facial pain.', **Step 2**: 'Initiate pharmacological treatment with carbamazepine or oxcarbazepine as the first-line treatment for trigeminal neuralgia, adjusting the dosage according to patient response and tolerance.', **Step 3**: 'Monitor for side effects of medication, such as dizziness, nausea, and diplopia.', **Step 4**: 'If the patient does not respond to or cannot tolerate the first-line medications, consider alternative treatments such as gabapentin, pregabalin, or low doses of tricyclic antidepressants like amitriptyline or nortriptyline.', **Step 5**: 'In cases where medication does not provide adequate relief, or for individuals who prefer not to take medication long term, refer the patient to a neurosurgeon for consideration of surgical options such as microvascular decompression or gamma knife radiosurgery.', **Step 6**: 'Provide patient education regarding the nature of the disease, the importance of medication adherence, and possible side effects.'},

Definitive Diagnosis: 'Trigeminal Neuralgia',

Treatment: {**Definitive Treatment**: 'First-line pharmacological treatment with carbamazepine or oxcarbazepine, with consideration for surgical intervention in refractory cases.'}}

Prompt 2

Clinical Note (deleted diagnosis with 0): A 31-year-old woman comes to the physician because of intermittent episodes of stabbing right lower jaw pain for 6 weeks. The pain is severe, sharp, and lasts for a few seconds. These episodes commonly occur when she washes her face, brushes her teeth, or eats a meal. She does not have visual disturbances, weakness of her facial muscles, or hearing loss. Five weeks ago, she had an episode of acute bacterial sinusitis, which was treated with antibiotics. Treatment with doxepin is initiated.

Prompt 3

CRR Round 2:

{**Differential Diagnosis Step by Step**: {**Step 1**: 'Given the location of pain in the right lower jaw and its triggering by mechanical stimuli such as washing the face and brushing teeth, the differential diagnosis initially includes dental causes like dental abscess, temporomandibular joint disorders, and neuralgias.', **Step 2**: 'The absence of dental pathology signs (e.g., toothache, swelling), temporomandibular joint dysfunction symptoms (e.g., jaw clicking, difficulty chewing), and the description of the pain as intermittent, stabbing, and severe, directs the differential towards neuralgias.', **Step 3**: "Excluding facial muscle weakness, hearing loss, and visual disturbances rules out conditions such as Bell's palsy, acoustic neuroma, and multiple sclerosis.", **Step 4**: 'The history of acute bacterial sinusitis treated with antibiotics suggests a possible relationship; however, the lack of ongoing sinusitis symptoms decreases the likelihood of this being the direct cause.', **Step 5**: 'The pain experienced during common stimuli like eating and brushing teeth specifically implicates triggers typical for trigeminal neuralgia.'},

Differential Diagnosis: {**Most Likely Cause**: 'Trigeminal Neuralgia', **Explanation**: 'Given the characteristics of the pain (severe, sharp, and lasts for a few seconds) occurring in the distribution of the trigeminal nerve, and triggered by touch, combined with the exclusion of other facial or dental pathologies, trigeminal neuralgia is the most likely diagnosis.'},

Treatment Step by Step: {**Step 1**: 'Initiate pharmacologic treatment with anticonvulsants, considering carbamazepine as the first-line treatment due to its effectiveness in relieving neuralgic pain.', **Step 2**: 'If carbamazepine is not tolerated or contraindicated, other medications such as gabapentin, lamotrigine, or oxcarbazepine may be considered as alternatives.', **Step 3**: 'Monitor for side effects of the medication and adjust dosage accordingly to optimize pain control while minimizing adverse effects.', **Step 4**: 'In cases where medication is ineffective or side effects are intolerable, referral to a neurosurgeon for evaluation for surgical intervention, such as microvascular decompression, may be considered.', **Step 5**: 'Provide patient education on trigeminal neuralgia, including potential triggers to avoid and the importance of adherence to medication for pain management.'},

Definitive Diagnosis: 'Trigeminal Neuralgia',

Treatment: neuralgic conditions, but the first-line treatment for trigeminal neuralgia should be carbamazepine, due to its proven efficacy in this condition. Other options include gabapentin and surgical interventions for refractory cases.'}}

Prompt 2/ 2nd time

CRR of Diagnosis and Clinical Management

2nd Case Scenario

Figure 8: The outputs of clinical notes with the second case scenario, level two notes. At stage 2, The CRR of Diagnosis and Clinical Management, using Prompts 2 and 3.

Verified CRR:

{**VerificationQueries**: {**Query 1**: 'What are the clinical practice guidelines for diagnosing trigeminal neuralgia?', **Query 2**: 'What are the first-line and alternative treatments for trigeminal neuralgia according to the latest clinical practice guidelines?', **Query 3**: 'Are there any recommended diagnostic criteria for trigeminal neuralgia that include the absence of other neurological signs?', **Query 4**: 'Does the clinical practice guideline recommend carbamazepine as the first-line medication for trigeminal neuralgia? What are the alternatives if carbamazepine is contraindicated or not tolerated?', **Query 5**: 'What are the indications for surgical intervention in trigeminal neuralgia according to clinical practice guidelines?'},

Retrieved Information: {**Response 1**: 'The latest clinical practice guidelines for diagnosing trigeminal neuralgia involve the identification of characteristic pain symptoms, such as sudden, brief, and severe episodes of facial pain along the distribution of the trigeminal nerve, without other neurological deficits.', **Response 2**: 'According to the guidelines, carbamazepine is recommended as the first-line treatment for trigeminal neuralgia, followed by oxcarbazepine. Gabapentin or lamotrigine are considered alternatives when the first-line treatments are not tolerated.', **Response 3**: 'The guidelines specify that diagnosis should consider the absence of other neurological signs to rule out secondary causes of facial pain.', **Response 4**: 'Yes, the guidelines confirm carbamazepine as the first-line therapy for trigeminal neuralgia and list gabapentin, lamotrigine, and oxcarbazepine as alternatives.', **Response 5**: 'Surgical intervention, such as microvascular decompression, is recommended for patients who do not respond to pharmacological treatment or who cannot tolerate the medication.'},

Comparison: {**Clinical Findings**: 'The differential diagnosis steps align well with guidelines, focusing on specific pain characteristics and ruling out other pathologies.',

Causes: 'The identified cause, trigeminal neuralgia, is directly supported by the absence of other neurological deficits and the pain's distribution, which matches guideline criteria.',

Treatment: 'The initial treatment plan aligns with guidelines, recommending carbamazepine as first-line treatment. Alternatives and surgical intervention for refractory cases are also in agreement with recommended practices.'},

Discrepancy: None,

Final CoT: {**Differential Diagnosis Process**: {**Step 1**: 'Given the location of pain in the right lower jaw and its triggering by mechanical stimuli such as washing the face and brushing teeth, the differential diagnosis initially includes dental causes like dental abscess, temporomandibular joint disorders, and neuralgias.', **Step 2**: 'The absence of dental pathology signs (e.g., toothache, swelling), temporomandibular joint dysfunction symptoms (e.g., jaw clicking, difficulty chewing), and the description of the pain as intermittent, stabbing, and severe, directs the differential towards neuralgias.', **Step 3**: 'Excluding facial muscle weakness, hearing loss, and visual disturbances rules out conditions such as Bell's palsy, acoustic neuroma, and multiple sclerosis.', **Step 4**: 'The history of acute bacterial sinusitis treated with antibiotics suggests a possible relationship; however, the lack of ongoing sinusitis symptoms decreases the likelihood of this being the direct cause.', **Step 5**: 'The pain experienced during common stimuli like eating and brushing teeth specifically implicates triggers typical for trigeminal neuralgia.'},

Definitive Cause: {**Most Likely Pathogen/Cause**: 'Trigeminal Neuralgia'},

Definitive Diagnosis: 'Trigeminal Neuralgia',

Treatment Plan: {**Step 1**: 'Initiate pharmacologic treatment with anticonvulsants, considering carbamazepine as the first-line treatment due to its effectiveness in relieving neuralgic pain.', **Step 2**: 'If carbamazepine is not tolerated or contraindicated, other medications such as gabapentin, lamotrigine, or oxcarbazepine may be considered as alternatives.', **Step 3**: 'Monitor for side effects of the medication and adjust dosage accordingly to optimize pain control while minimizing adverse effects.', **Step 4**: 'In cases where medication is ineffective or side effects are intolerable, referral to a neurosurgeon for evaluation for surgical intervention, such as microvascular decompression, may be considered.', **Step 5**: 'Provide patient education on trigeminal neuralgia, including potential triggers to avoid and the importance of adherence to medication for pain management.'}}}

Prompt 4

CRR of Diagnosis and Clinical Management **2nd Case Scenario**

Figure 9: The verified CRR output of clinical notes corresponds to the second case scenario and level two notes. At stage 3, the CRR of Diagnosis and Clinical Management, using Prompt 4. The Final CoT is the verified CRR. CRR; Clinical Reasoning Rationale

Final Note:

```
{
  "error_flag": 1,
  "error_location": 5,
  "sentence_correction": "Treatment with carbamazepine is initiated."
}
```

Prompt 5

Medical Errors Detection and Correction **2nd Case Scenario**

Figure 10: The last structured output includes: "error flag" to identify if an error exists. And "error location" and "sentence correction" if an error within a clinical note corresponds to the second case scenario and level two notes. At stage 4, the Medical Errors Detection and Correction, using Prompt 5.

Preprocessed Note:
 This patient is also being managed for chronic respiratory failure, POA (present on admission). Mr. <NAME/> is a <AGE/> yr male with Hg SS sickle cell disease, admitted for sickle pain crisis. Pt has chronic **Epiglottitis**: Likely <DATE/> anemia and chronic illness. - Cont home O2, >2L at all times - RT following for nocturnal BiPap.

Prompt 1

Preprocessing 3rd Case Scenario Deleted Part: {}

Figure 11: The outputs of clinical notes with the third case scenario, level two clinical notes. At stage 1, the Preprocessing, using Prompt 1.

CRR Round 1:
 {'**Differential Diagnosis Step by Step**': {'Step 1': 'The patient's history of Hg SS sickle cell disease and chronic respiratory failure suggests complications related to these conditions.', 'Step 2': 'The presentation of sickle pain crisis indicates an exacerbation of the sickle cell disease, potentially triggering acute chest syndrome or other vaso-occlusive complications.', 'Step 3': 'Chronic epiglottitis, chronic anemia, and chronic illness might indicate ongoing or recurrent infections, inflammation, or complications from sickle cell disease.'}, '**Differential Diagnosis**': {'**Most Likely Cause**': 'Complications associated with Hg SS sickle cell disease.'}, '**Explanation**': 'Given the patient's chronic conditions and the presentation of a sickle pain crisis, complications of sickle cell disease are likely contributing to the current state. These can include acute chest syndrome, exacerbated by chronic respiratory failure, or other vaso-occlusive events leading to organ damage and pain. Chronic epiglottitis could be a result of recurrent infections or chronic inflammation.'}, '**Treatment Step by Step**': {'Step 1': 'Manage the sickle pain crisis with hydration, oxygen therapy, and pain management tailored to the patient's needs.', 'Step 2': 'Evaluate the need for transfusion or exchange transfusion to address anemia and reduce sickle hemoglobin levels.', 'Step 3': 'Continue home oxygen therapy and monitor the need for nocturnal BiPAP to manage chronic respiratory failure.', 'Step 4': 'Assess the epiglottitis and consider antibiotics if a bacterial infection is suspected, along with steroids to reduce inflammation.', 'Step 5': 'Long-term, explore disease-modifying treatments for sickle cell disease, such as hydroxyurea or stem cell transplantation, to reduce the frequency of vaso-occlusive crises and complications.'}, '**Definitive Diagnosis**': 'Exacerbation of Hg SS sickle cell disease leading to acute sickle pain crisis and potential acute chest syndrome, coupled with chronic respiratory failure.', '**Treatment**': {'**Definitive Treatment**': 'The treatment involves managing the acute crises and the chronic conditions concurrently. This includes pain management, possibly transfusions, continued oxygen therapy and possibly BiPAP for respiratory support, antibiotics and steroids if an infection or significant inflammation is present, and considering disease-modifying therapies for long-term management of sickle cell disease.'}}}

Prompt 2

Clinical Note (deleted diagnosis with 0): This patient is also being managed for chronic respiratory failure, POA (present on admission). Mr. <NAME/> is a <AGE/> yr male with Hg SS sickle cell disease, admitted for sickle pain crisis. Pt has chronic **0**: Likely <DATE/> anemia and chronic illness. - Cont home O2, >2L at all times - RT following for nocturnal BiPap.

Prompt 3

CRR Round 2:
 {'**Differential Diagnosis Step by Step**': {'Step 1': 'Identify key clinical features: Patient with known Hg SS sickle cell disease presented for pain crisis, has chronic anemia, chronic respiratory failure, requires home oxygen therapy (>2L), and nocturnal BiPAP.', 'Step 2': 'Consideration of common complications in sickle cell disease: Acute pain crisis, acute chest syndrome, pulmonary hypertension, and chronic organ damage.', 'Step 3': 'Analysis of provided information regarding respiratory management suggests chronic respiratory complications, possibly pulmonary hypertension or hypoxemia secondary to chronic sickle cell lung disease.', 'Step 4': 'Given the chronic nature of the respiratory failure requiring oxygen and BiPAP, identify potential causes such as vaso-occlusive episodes leading to pulmonary infarction or chronic acute chest syndromes contributing to progressive lung function decline.'}, '**Differential Diagnosis**': {'**Most Likely Cause**': 'Pulmonary Hypertension secondary to chronic sickle cell disease.'}, '**Explanation**': 'Chronic anemia and repeated vaso-occlusive crises can lead to pulmonary hypertension, a common complication in patients with chronic sickle cell disease. This would also explain the requirement for continuous oxygen therapy and nocturnal BiPAP to manage resultant chronic respiratory failure.'}, '**Treatment Step by Step**': {'Step 1': 'Continue current supportive measures, including oxygen therapy and nocturnal BiPAP, to manage chronic respiratory failure.', 'Step 2': 'Management of sickle cell disease exacerbations with hydration, pain control, possibly blood transfusions to reduce sickling episodes.', 'Step 3': 'Consider initiating or optimizing specific therapies for pulmonary hypertension, which could include phosphodiesterase-5 inhibitors (e.g., sildenafil) or other pulmonary vasodilators.', 'Step 4': 'Regular follow-up with a comprehensive sickle cell management team, including specialists in hematology, pulmonology, and potentially cardiology.', 'Step 5': 'Evaluation and possible adjustment of current hydroxyurea therapy to reduce the frequency of pain crises and other complications related to sickle cell disease.', 'Step 6': 'Encourage vaccinations and prophylactic antibiotics as appropriate to reduce the risk of infections leading to acute chest syndrome.'}, '**Definitive Diagnosis**': 'Chronic Respiratory Failure secondary to Pulmonary Hypertension in the setting of Sickle Cell Disease.', '**Treatment**': {'**Definitive Treatment**': 'Optimization of sickle cell disease management including pain crisis management, chronic respiratory failure support with oxygen and BiPAP, consideration of pulmonary hypertension targeted therapy, hydroxyurea therapy optimization, and preventive measures for complications.'}}}

Prompt 2/ 2nd time

CRR of Diagnosis and Clinical Management 3rd Case Scenario

Figure 12: The outputs of clinical notes with the third case scenario, level two notes. At stage 2, The CRR of Diagnosis and Clinical Management, using Prompts 2 and 3.

Verified CRR:

{**VerificationQueries:** {Query 1: 'What are the American Thoracic Society guidelines for diagnosing pulmonary hypertension in the context of sickle cell disease?', Query 2: 'What are the clinical findings indicative of pulmonary hypertension as per the latest guidelines?', Query 3: "What are the National Heart, Lung, and Blood Institute's recommendations for the management of sickle cell disease complications, particularly pulmonary hypertension?", Query 4: 'What are the recommended treatments for chronic respiratory failure secondary to pulmonary hypertension in sickle cell disease patients according to the most recent clinical practice guidelines?'},

Retrieved Information: {Response 1: 'The American Thoracic Society guidelines recommend echocardiographic screening for pulmonary hypertension in sickle cell disease patients, with confirmatory diagnosis requiring right heart catheterization.', Response 2: 'Clinical findings indicative of pulmonary hypertension include dyspnea, chest pain, and signs of right heart failure. Echocardiography may show elevated tricuspid regurgitant jet velocity suggesting pulmonary hypertension.', Response 3: 'The National Heart, Lung, and Blood Institute recommends comprehensive care including hydroxyurea therapy to reduce sickle cell-related complications, aggressive treatment of acute chest syndrome, and evaluation for pulmonary hypertension in patients with sickle cell disease.', Response 4: 'Recommended treatments for chronic respiratory failure in sickle cell disease include oxygen therapy, treatment of underlying pulmonary hypertension with medications such as phosphodiesterase-5 inhibitors, and management of sickle cell disease exacerbations.'},

Comparison: {Clinical Findings: 'Echocardiographic screening and right heart catheterization for pulmonary hypertension align with proposed diagnosis considerations.'},

Causes: 'Chronic anemia and vaso-occlusive crises leading to pulmonary hypertension are consistent with suggested complications of sickle cell disease.',

Treatment: 'Supportive measures, specific therapies for pulmonary hypertension, hydroxyurea optimization, and preventive measures align with guideline recommendations.'},

Discrepancy: None,

Final COT: {**Differential Diagnosis Process:** {Step 1: 'Identify key clinical features: Patient with known Hg SS sickle cell disease presented for pain crisis, has chronic anemia, chronic respiratory failure, requires home oxygen therapy (>2L), and nocturnal BiPAP.', Step 2: 'Consideration of common complications in sickle cell disease: Acute pain crisis, acute chest syndrome, pulmonary hypertension, and chronic organ damage.', Step 3: 'Analysis of provided information regarding respiratory management suggests chronic respiratory complications, possibly pulmonary hypertension or hypoxemia secondary to chronic sickle cell lung disease.', Step 4: 'Given the chronic nature of the respiratory failure requiring oxygen and BiPAP, identify potential causes such as vaso-occlusive episodes leading to pulmonary infarction or chronic acute chest syndromes contributing to progressive lung function decline.'},

Definitive Cause: 'Most Likely Pathogen/Cause: 'Pulmonary Hypertension secondary to chronic sickle cell disease.'},

Definitive Diagnosis: 'Chronic Respiratory Failure secondary to Pulmonary Hypertension in the setting of Sickle Cell Disease.', 'Treatment Plan': {Step 1: 'Continue current supportive measures, including oxygen therapy and nocturnal BiPAP, to manage chronic respiratory failure.', Step 2: 'Management of sickle cell disease exacerbations with hydration, pain control, possibly blood transfusions to reduce sickling episodes.', Step 3: 'Consider initiating or optimizing specific therapies for pulmonary hypertension, which could include phosphodiesterase-5 inhibitors (e.g., sildenafil) or other pulmonary vasodilators.', Step 4: 'Regular follow-up with a comprehensive sickle cell management team, including specialists in hematology, pulmonology, and potentially cardiology.', Step 5: 'Evaluation and possible adjustment of current hydroxyurea therapy to reduce the frequency of pain crises and other complications related to sickle cell disease.', Step 6: 'Encourage vaccinations and prophylactic antibiotics as appropriate to reduce the risk of infections leading to acute chest syndrome.'}}}

Prompt 4

CRR of Diagnosis and Clinical Management **3rd Case Scenario**

Figure 13: The verified CRR output of clinical notes corresponds to the third case scenario and level two notes. At stage 3, the CRR of Diagnosis and Clinical Management, using Prompt 4. The Final CoT is the verified CRR. CRR; Clinical Reasoning Rationale

Final Note:

{"error_flag":1,
"error_location":3,
"sentence_correction": "Pt has chronic Hypoxia likely due to pulmonary hypertension secondary to sickle cell disease."}

Prompt 5

Medical Errors Detection and Correction **3rd Case Scenario**

Figure 14: The last structured output includes: "error flag" to identify if an error exists. And "error location" and "sentence correction" if an error within a clinical note corresponds to the third case scenario and level two notes. At stage 4, the Medical Errors Detection and Correction, using Prompt 5.