

Edinburgh Clinical NLP at MEDIQA-CORR 2024: Guiding Large Language Models with Hints

Aryo Pradipta Gema^{1*} Chaeun Lee^{1*} Pasquale Minervini¹ Luke Daines²

T. Ian Simpson¹ Beatrice Alex^{3,4}

¹School of Informatics, University of Edinburgh ²Usher Institute, University of Edinburgh

³Edinburgh Futures Institute, University of Edinburgh

⁴School of Literatures, Languages and Cultures, University of Edinburgh

{aryo.gema, chaeun.lee, p.minervini, luke.daines}@ed.ac.uk

{ian.simpson, b.alex}@ed.ac.uk

Abstract

The MEDIQA-CORR 2024 shared task aims to assess the ability of Large Language Models (LLMs) to identify and correct medical errors in clinical notes. In this study, we evaluate the capability of general LLMs, specifically GPT-3.5 and GPT-4, to identify and correct medical errors with multiple prompting strategies. Recognising the limitation of LLMs in generating accurate corrections only via prompting strategies, we propose incorporating error-span predictions from a smaller, fine-tuned model in two ways: 1) by presenting it as a hint in the prompt and 2) by framing it as multiple-choice questions from which the LLM can choose the best correction. We found that our proposed prompting strategies significantly improve the LLM's ability to generate corrections. Our best-performing solution with 8-shot + CoT + hints ranked sixth in the shared task leaderboard. Additionally, our comprehensive analyses show the impact of the location of the error sentence, the prompted role, and the position of the multiple-choice option on the accuracy of the LLM. This prompts further questions about the readiness of LLM to be implemented in real-world clinical settings.¹

1 Introduction

Medical errors represent a major concern in the healthcare sector, leading to adverse patient outcomes and higher costs for healthcare providers. The detection and correction of such medical errors are critical in enhancing healthcare delivery and outcomes. Recognising the importance of efficient and precise medical documentation, the MEDIQA-CORR 2024 shared task (Ben Abacha et al., 2024a) is initiated to evaluate the potential of using Large Language Models (LLMs) as solutions to locate and correct medical errors within clinical notes.

In our study, we evaluated multiple prompting strategies such as In-context Learning (ICL) and Chain-of-Thought (CoT) to enhance the performance of LLMs, specifically focusing on GPT-3.5 and GPT-4 (OpenAI, 2023). We proposed incorporating a smaller fine-tuned language model, namely BioLinkBERT (Yasunaga et al., 2022), to aid LLMs in locating an error span in a clinical note. We incorporated the predicted error span in two ways: 1) by presenting it as a hint in the prompt to direct the error correction, and 2) by framing it as multiple-choice questions where the LLM can select the most probable correction.

Our findings revealed that the LLMs show noticeable improvements in their generation capability when presented with more ICL examples. Similarly, the CoT prompt also improves the error correction capability of the LLMs. Among the different reasoning styles we experimented with, the LLM performs the best with brief reasoning. Our prompt design, which provides a hint about the typical nature of the errors and a hint from the error span prediction, further improves the LLMs' ability to generate corrections. The combination of 8-shot ICL with Brief CoT reasoning and hints is the best-performing prompting strategy in the two provided validation sets. This pipeline ranked sixth in the shared task leaderboard. In summary, our contributions are as follows:

- A comprehensive analysis of the impact of ICL on the performance of LLMs for medical error correction.
- An extensive exploration of CoT to inject various reasoning styles into the LLM and their impact on the performance.
- Novel approaches to integrate the predictions of a smaller language model into the LLM generation.
- Sensitivity analyses of LLMs, highlighting how minor variations such as the error sentence loca-

*Equal contribution.

¹Our code is available at <https://github.com/aryopg/mediqa>

Category	Train		Valid		Test	
	MS	UW	MS	UW	MS	UW
No Error	970	0	255	80	-	-
Contain Error	1,219	0	319	80	-	-
Total	2,189	0	574	160	597	328

Table 1: Dataset statistics of each split, categorised by the source and presence of a medical error.

tion, the prompted role, and the multiple-choice positioning can influence generation capabilities.

2 Background

2.1 Task Description

MEDIQA-CORR 2024 task (Ben Abacha et al., 2024b) comprises three sub-tasks, each addressing a different aspect of medical error correction:

Binary classification: Detecting whether the clinical note contains a medical error.

Span Identification: Identifying the text span associated with a medical error if it exists.

Natural Language Generation: Generating a correction if a medical error exists.

Table 1 shows the statistics for each data split, organised by the source of the data and whether or not it contains a medical error. Each clinical note contains either one or no medical error.

The task uses accuracy for binary classification and span identification. The generated correction is evaluated using an aggregate Natural Language Generation (NLG) score, combining ROUGE-1 (Lin, 2004), BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020), which is best aligned with human judgement, among other NLG metrics (Ben Abacha et al., 2023).

2.2 Related work

LLMs have shown remarkable capabilities in many NLP tasks, including in the clinical domain. Liévin et al. (2022) evaluated LLMs with various prompting strategies, showing LLMs’ capability to answer complex medical questions. Falis et al. (2024) uses GPT-3.5 to generate accurate synthetic discharge summaries by prompting it with a list of diagnoses. Gema et al. (2024) also shows GPT-4 in zero-shot setting outperforms other fine-tuned LLMs in a natural language inference task for clinical trial data.

However, despite the increasing use of general LLMs, their performance varies widely depending on the nature of the task. For instance,

fine-tuned smaller encoder-based models (e.g., BioLinkBERT) still maintain the lead in tasks such as medical entity recognition (Kim et al., 2023). Gema et al. (2023) showed that domain-adapted LLaMA (Touvron et al., 2023) outperforms the state-of-the-art models in clinical outcome prediction tasks. Such studies show that fine-tuned models are still preferable, especially in discriminative tasks such as classification and entity recognition.

In this study, we seek to combine the generative capability of LLMs with the discriminative capability of a smaller fine-tuned language model. We compared our novel method with solutions that rely solely on prompting strategies (i.e., ICL and CoT).

3 System Overview

We experimented with three strategies:

End-to-end Prompting Strategy for Error Correction: This strategy treats all three subtasks as a single prompting task. The LLM simultaneously predicts if the clinical note contains an error, pinpointing its location, and proposing a correction.

Fine-tuning Error Span Prediction and MCQ-style Error Correction: This method splits the task into error span prediction and correction. It uses a fine-tuned model for error span prediction, followed by MCQ-style prompts for correction.

Hybrid Approach: As shown in Figure 1, This approach uses error span predictions from a fine-tuned model as correction hints injected into the end-to-end prompting strategy. This is our best-performing strategy in both validation and test sets.

The following sections outline the details for the **Error Span Prediction** and **Error Correction**.

3.1 Error Span Prediction

We noticed that medical errors appear predominantly in the form of diagnoses or treatments, instead of the patient’s factual information. This finding motivated us to fine-tune an encoder model to first detect an error span within the clinical note.

We trained BioLinkBERT and BERT² using a question-answering pipeline adapted from the Stanford Question Answering Dataset (SQuAD). We pre-processed the training and validation sets to align them with the SQuAD v1 format, which assumes that there is always an error span in the input. We introduced a template question, “Which part in the given clinical note is clinically incorrect?” in the question column of the SQuAD format. The

²Both base and large versions of the models

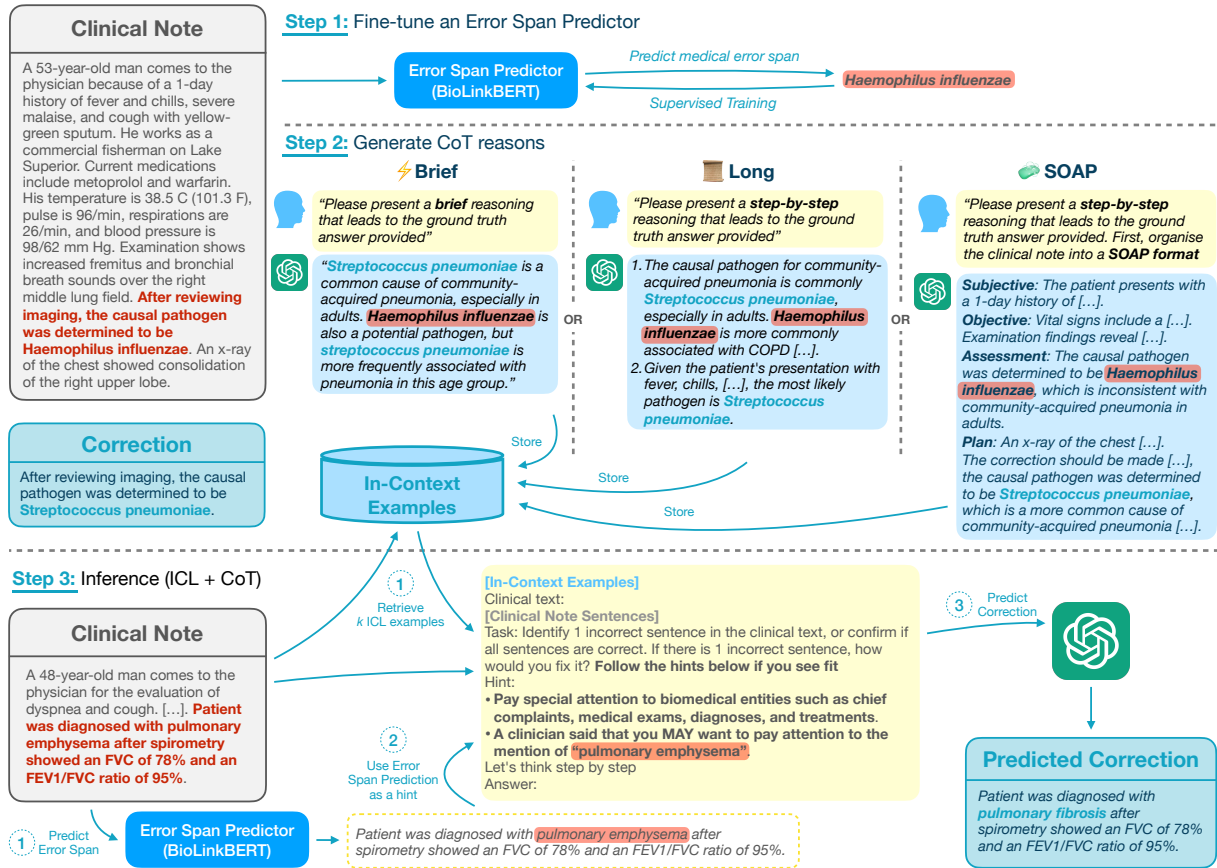


Figure 1: Schema of our best-performing strategy with In-Context Learning (ICL) and Chain-of-Thought (CoT) prompting strategies. The strategy involves fine-tuning BioLinkBERT on the training set for error span prediction. Then, we prompt GPT-3.5 with various reasoning templates to reason pairs of clinical notes and ground truth corrections to gather ICL examples with CoT reasons. Subsequently, this strategy leverages the ICL examples and error span predictions as a hint.

trained model predicts the start and end indices, which indicate the position of the predicted error span in the text.

We trained and evaluated the error span prediction models only on clinical notes that contained errors. We evaluated the models using exact match (EM) and token-based F1 score metrics, using the latter to choose the best checkpoint.

3.2 Error Correction

We experimented with GPT-3.5 and GPT-4 for the error correction step. We prompted the LLMs to return the outputs in JSON format for ease of post-processing. In rare cases where the outputs are not JSON-parseable, we default the prediction as if no error was found. We integrated the error span prediction to this error correction step in two ways:

3.2.1 Multiple-Choice Question prompt

As shown in Figure 2, this strategy involves two interactions with the LLM: 1) to construct an options

set and 2) to ask a multiple-choice question.

In the first interaction, the model generates potential replacement options for the identified error span. Here, the predicted error span is replaced with a placeholder "**<BLANK>**", and the LLM is tasked with generating n replacement candidates. During our experiments, we observed a pattern where the model often included the predicted error span or its synonyms in the options. To eliminate this redundancy, we added a directive prompt "*Do not include the **<predicted_error_span>** or its medical synonyms in your answer*".

In the second interaction, we query the LLM with an MCQ-style prompt, which presents the full clinical note, with the predicted error span replaced by "**<BLANK>**", and the options comprised of n LLM-generated options from the first interaction and the predicted error span (totalling $n + 1$ options). The LLM chooses the best correction among these options. Subsequently, we derive the error flag classification based on the LLM's re-

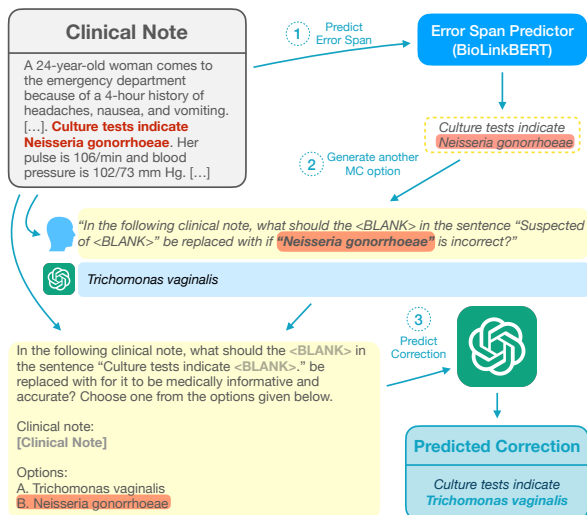


Figure 2: Schema of the Multiple-Choice Question prompt strategy.

sponse, 0 if it selects the predicted error span as the correct answer, or 1 if the model selects one of the other choices. We experimented with varying the number of answer choices to two and four options.

3.2.2 Hybrid Approach

As illustrated in Figure 1, the pipeline continues with the preparation of the ICL examples after the training for the error span prediction. For solutions that rely only on ICL examples and do not require CoT reasoning, we directly retrieve pairs of clinical notes and their respective ground-truth corrections as ICL examples. In contrast, CoT-based solutions require ICL examples with reasons provided. Inspired by He et al. (2023), we prompted GPT-3.5 (gpt-3.5-turbo-0613) to generate a reasoning for the ICL examples. We selected GPT-3.5 particularly because of its generation capability and clinical knowledge (Gema et al., 2024).

We experimented with three CoT reasoning templates: **Brief**, **Long**, and **SOAP**. All reasoning templates require the model to reason the ground-truth correction by identifying the incorrect span and providing the reasoning behind it. However, each format provides a different depth and structure of reasoning. The **Brief CoT** template prompts concise reasoning, the **Long CoT** template requires detailed step-by-step explanations, and the **SOAP CoT** template organises information according to Subjective, Objective, Assessment, and Plan sections before making corrections.

During inference, the solution uses a selected reasoning format with ICL examples to correct clinical notes. The model applies a reasoning strat-

egy to new scenarios based on the reasoned ICL examples which are retrieved using the BM25 algorithm (Robertson et al., 1995), selecting examples similar to the clinical note in question. We also integrate a hint about the typical nature of the errors, focusing the model’s attention on specific biomedical entities such as diagnoses and treatments (i.e., “Pay special attention to biomedical entities such as chief complaints, medical exams, diagnoses, and treatments.”). We denote this as “**Type hint**”. Finally, we leverage the error span prediction by adding it as another hint, denoted as “**Span hint**” (i.e., “A clinician said that you MAY want to pay attention to the mention of <predicted_error_span>”).

4 Results

Our experiments are structured as answers to sequential research questions. Firstly, we conducted experiments to find the best model for error span prediction, evaluating them on EM and F1 scores. Subsequently, we experimented with various prompting strategies for error correction, evaluating them on the macro-averaged accuracy and aggregate NLG scores across MS and UW datasets. The first error correction experiment starts with an end-to-end prompting approach, relying solely on the LLM capability with ICL and CoT to correct errors. We, then, experimented with integrating the error span prediction model into the error correction process via the MCQ-style prompt. Lastly, we experimented with the hybrid approach, integrating the error span prediction as a hint for the end-to-end prompting approach. We used GPT-3.5 in our error correction experiments on the validation sets³, choosing the best prompting strategy to be implemented with GPT-4 on the test set.

RQ1: How well are the smaller LMs performing in the error span detection?

As shown in Table 2, we experimented with general (i.e., BERT-base and -large) and domain-adapted models (i.e., BioLinkBERT-base and -large) for the error span prediction. We evaluated the models exclusively on a subset of the validation set that contains a medical error as stated in Subsection 3.1.

Among all models, BioLinkBERT-large showed the highest EM and F1 scores on the MS validation set, indicating a superior ability to predict error spans within clinical notes. This suggests that

³Due to a limited research budget.

Model	MS		UW	
	EM	F1	EM	F1
BERT-base	54.86	80.09	1.25	4.44
BERT-large	55.17	79.30	5.00	7.92
BioLinkBERT-base	55.17	81.33	6.25	12.29
BioLinkBERT-large	58.31	82.49	6.25	8.91

Table 2: Performance of fine-tuned error span prediction models. **Bold cell** indicates the highest score for the metric.

# shots	Acc _{flag}	Acc _{sent_id}	Score _{agg}
2	0.5089	0.3348	0.4139
4	0.5242	0.4215	0.4503
8	0.5268	0.4526	0.5038

Table 3: Performance of GPT-3.5 using different numbers of ICL examples on validation sets. **Bold cell** indicates the highest score for the metric.

the domain-adaptive pretraining that BioLinkBERT has undergone contributes to its performance in medical error detection tasks. However, all models struggle to accurately predict error spans on the UW validation set. Recognising this, we trained BioLinkBERT-large on the MS train dataset and 25% of the UW validation dataset as the error span prediction model for the subsequent experiments.

RQ2: Can LLMs perform well end-to-end solely with prompting strategies?

Before leveraging the error span prediction, we began our error correction experiment by solely relying on the LLM with prompting strategies to correct errors without any help from the error span prediction. This prompt-only end-to-end approach serves as the baseline for our proposed solutions.

RQ2.1: Do more ICL examples improve the LLM’s performance?

Firstly, we experimented with varying the number of ICL examples on GPT-3.5’s performance across MS and UW validation sets. We did not report 0-shot performance as the LLM failed to generate a parseable answer, indicating that the LLM failed to complete the task without any examples. As shown in Table 3, we observe a trend where the performance of the LLM improves in all metrics as the number of shots increases, with the 8-shot setting performing the best. Our subsequent experiments will use the 8-shot ICL setup.

Type Hint	Acc _{flag}	Acc _{sent_id}	Score _{agg}
✗	0.5527	0.4472	0.4467
✓	0.5268 (-0.03)	0.4526 (+0.01)	0.5038 (+0.06)

Table 4: Performance of GPT-3.5 using 8-shot prompt with or without a type hint on validation sets. Values in parentheses indicate the performance difference against the LLM that does not receive a type hint. **cyan** indicates improvement, **red** indicates decrease. **Bold cell** indicates the highest score for the metric.

CoT	Acc _{flag}	Acc _{sent_id}	Score _{agg}
None	0.5268	0.4526	0.5038
Brief	0.5866 (+0.06)	0.4989 (+0.05)	0.5389 (+0.04)
Long	0.6074 (+0.08)	0.4717 (+0.02)	0.4930 (-0.01)
SOAP	0.5186 (-0.01)	0.4058 (-0.05)	0.4228 (-0.08)

Table 5: Performance of GPT-3.5 using 8-shot and type hint prompt with various CoT formats on validation sets. Values in parentheses indicate the performance difference against the LLM that does not use CoT reasoning. **cyan** indicates improvement, **red** indicates decrease. **Bold cell** indicates the highest score for the metric.

RQ2.2: Adding a hint about the typical error

In our first experiment, we observed that the LLMs tend to correct non-essential errors (e.g., grammatical and unit errors). Thus, we prompted the LLM with a hint about the typical form of the errors (i.e., “Pay special attention to biomedical entities such as chief complaints, medical exams, diagnoses, and treatments.”). Table 4 shows the performance comparison between a prompt with and without this hint. When a hint is provided, there is a decrease in the error flag accuracy by 0.03 which may indicate that there are medical errors that are not one of the specified biomedical entities. However, this is compensated by improvements in both sentence ID accuracy and the aggregate NLG score, with the latter seeing a notable increase of 0.06. This indicates that while the hint may slightly hinder the model’s binary classification ability, it correctly directs the focus of the LLM in locating the error.

RQ2.3: Chain-of-Thought with various formats

Table 5 evaluates the effect of different Chain-of-Thought (CoT) formats on GPT-3.5’s performance. The absence of CoT (None) serves as a baseline against which the Brief, Long, and SOAP formats are compared. The Brief CoT format leads to improvements across all metrics, particularly in sentence ID accuracy and the aggregate NLG score, underscoring the benefit of concise, targeted rea-

Prompting Strategy	Acc _{flag}	Acc _{sent_id}	Score _{agg}
8-shot + Brief CoT	0.5866	0.4989	0.5389
MCQ (2 options)	0.6131	0.6029	0.6492
MCQ (4 options)	0.6087	0.5944	0.6448

Table 6: Performance of GPT-3.5 with the MCQ-style prompt on validation sets. **Bold cell** indicates the highest score for the metric.

soning in enhancing model performance. The Long format, while offering the highest accuracy in error flagging, exhibits a decrease in the aggregate score, suggesting that excessive detail may detract from overall correction quality. Conversely, the SOAP format results in declines across all metrics, highlighting that detailed and structured reasoning approaches may not necessarily be beneficial and may even hinder the model’s effectiveness.

RQ3: Can LLMs perform if provided with a span hint?

After the experiments with different prompting setups, we experimented with integrating the error span prediction into the error correction process.

RQ3.1: Can LLMs perform better with MCQ-style prompts?

As shown in Table 6, MCQ-style prompt using error span prediction improved performance over end-to-end systems. This can be attributed to two reasons. First, the MCQ-style prompt provides options that match the specificity of the predicted error span in the original clinical note, limiting the LLMs’ tendency to generate generic corrections. Second, the MCQ-style prompt addresses the LLMs’ tendency to be verbose by limiting corrections to a specific error span.

RQ3.2: Can end-to-end LLMs perform better when provided with a span hint?

In our RQ2 experiments with end-to-end systems, we observed limitations in the LLM’s ability to accurately locate errors within the clinical notes. While in RQ3.1, we noticed that integrating error span predictions helped improve the LLM’s performance. These insights motivated us to integrate the error span predictions from fine-tuned models to the end-to-end LLM solution. We denoted this solution as the “Hybrid approach”, as mentioned in Subsubsection 3.2.2, leveraging the “Span hint” from the error span prediction.

CoT	Span Hint	Acc _{flag}	Acc _{sent_id}	Score _{agg}
MCQ (2 opt)	✓	0.6131	0.6029	0.6492
MCQ (4 opt)	✓	0.6087	0.5944	0.6448
None	✗	0.5268	0.4526	0.5038
	✓	0.5671 (+0.04)	0.5543 (+0.10)	0.7348 (+0.23)
Brief	✗	0.5866	0.4989	0.5389
	✓	0.5610 (-0.03)	0.5454 (+0.05)	0.7385 (+0.20)
Long	✗	0.6074	0.4717	0.4930
	✓	0.6048 (-0.00)	0.4651 (-0.01)	0.4822 (-0.01)
SOAP	✗	0.5186	0.4058	0.4228
	✓	0.5237 (+0.01)	0.4310 (+0.03)	0.4884 (+0.07)

Table 7: Performance of GPT-3.5 using 8-shot and type hint prompt with various CoT format and with or without receiving span hint on validation sets. Values in parentheses indicate the performance difference against the solution that does not receive a span hint. **cyan** indicates improvement, **red** indicates decrease. **Bold cell** indicates the highest score for the metric.

Prompting Strategy	Acc _{flag}	Acc _{sent_id}	Score _{agg}
8-shot + Hints	0.5243	0.4649	0.6274
8-shot + Brief CoT + Hints	0.6681	0.5924	0.6634
MCQ (2 options)	0.6573	0.5957	0.6267
MCQ (4 options)	0.5935	0.5232	0.5882

Table 8: Results of GPT-4 with either ICL + CoT + hinted prompt or Multiple-Choice-Question prompt on test sets. The models are compared based on the aggregate NLG score.

Integrating a span hint into the end-to-end LLM prompt resulted in improvements across all metrics, as shown in Table 7. Notably, span hint significantly improved the aggregate NLG scores of Brief CoT and no-CoT solutions. However, span hint did not improve Long CoT solution, suggesting that the reasoning style may influence the LLM’s ability to leverage span hints.

Despite MCQ prompts demonstrating higher accuracy in error sentence identification, “Brief CoT” prompts combined with ICL, type hint, and span hints showed a higher aggregate NLG score, emphasising the different strengths of the two strategies. This indicates that the hybrid approach harnesses the LLM’s generative capabilities, while the fine-tuned error span prediction model helps direct these corrections to the appropriate error locations.

Performance on Test Set

We submitted our four best-performing solutions to be evaluated on the holdout test set. As shown in Table 8, we can observe a similar trend as in the validation set experiments. The 2-options MCQ prompts show strong performance in accurately

identifying the error-containing sentence. The 8-shot + Brief CoT + Hints method performs better, especially in the aggregate NLG score. This suggests that while MCQ prompts effectively direct the model’s focus, enabling accurate detection of errors, they may slightly constrain the model’s generative capability. Overall, these results highlight the benefit of using concise CoT reasoning in LLMs as well as providing guidance via targeted hints. Our best-performing pipeline, 8-shot + Brief CoT + Hints, ranked sixth in the shared task leaderboard based on the aggregate NLG score.

5 Post-hoc Analyses

Commonly reported NLG metrics tend to not be well correlated with human judgement, especially in the clinical domain (Ben Abacha et al., 2023). To understand the limitations of LLMs for clinical note correction, we extend beyond the reported performance metrics by analysing the sensitivity of LLMs to the data and prompt, as well as the common mistakes that LLMs tend to commit.⁴

5.1 Sensitivity

It is a well-known fact that the performance of an LLM may differ massively given slight differences in the way we prompt it (Voronov et al., 2024). We analysed factors observed in the data and prompt that may contribute to performance differences.

5.1.1 Sensitivity to the position of error sentence in the clinical note

We investigated the sensitivity of the model performance to the position of the error sentence within a given clinical note, dividing them into three cases; if the error sentence is in the first sentence (“beginning”), the last sentence (“end”), or in between the first and the last sentences (“middle”).

Figure 3 illustrates the relationship between the NLG metrics and the error sentence position, along with the proportion of the error sentence location. We can observe that ROUGE 1, BERTScore, and BLEURT scores do not vary significantly based on the position of the error sentence. This observation is quantitatively supported by the Kruskal-Wallis H-Test and the post-hoc Dunn’s test results shown in Appendix D. The test results reveal that the LLM’s ability to generate accurate corrections is not impacted by where the error appears in the input, which is a desirable trait.

⁴Post-hoc analyses are conducted on the validation sets.

Role	Acc _{flag}	Acc _{sent_id}	Score _{agg}
Clinician assistant	0.5610	0.5454	0.7385
No role	0.5570 (-0.00)	0.5416 (-0.00)	0.7504 (+0.01)
Assistant	0.5509 (-0.01)	0.5442 (-0.00)	0.7504 (+0.01)
Medical student	0.5539 (-0.01)	0.5468 (+0.00)	0.7484 (+0.01)
Nurse	0.5763 (+0.02)	0.5615 (+0.02)	0.7424 (+0.00)
Clinical note vericator	0.5554 (+0.01)	0.5438 (-0.00)	0.7518 (+0.01)
Clinician	0.5793 (+0.02)	0.5615 (+0.02)	0.7615 (+0.02)

Table 9: Performance of our best-performing solution when prompted with different roles via the system prompt (i.e., “You are «a role» tasked to ...”) on the validation sets.

Generated Option Position	Acc _{flag}	Acc _{sent_id}	Score _{agg}
A	0.6131	0.6029	0.6492
B	0.6368	0.6265	0.6380

Table 10: Results of the sensitivity analysis of MCQ-style prompt to the position of the LLM-generated option in the 2 options setting on validation sets.

5.1.2 Sensitivity to the role described in the system prompt

Owing to their instruction-following ability, LLMs are capable of playing a role as prompted by the user (Wang et al., 2023). In the clinical domain, we tend to prompt an LLM to answer a query as a healthcare professional, such as a clinician. In this analysis, we explored how the role prompted or the lack thereof may affect the performance of the LLM in generating corrections. We modify the system prompt (i.e., “You are «a role» tasked to ...”) with various role options. Table 9 details the varying performances of the best-performing 8-shot + Brief CoT + hints solution when prompted with different roles. The LLM performs best when prompted to role-play as a “clinician”. This phenomenon, known as *In-Context Impersonation* (Salewski et al., 2024), highlights that role-playing should be examined when developing a prompt-based solution.

5.1.3 Sensitivity to the position of the multiple choice options

Table 10 shows the outcome of a sensitivity analysis, based on the relative positioning of the LLM-generated option and the predicted error span within the original text for the systems with MCQ-type prompts. Both binary classification accuracy and error sentence prediction accuracy were improved when the LLM-generated option was positioned as option B, as opposed to option A. On the other hand, the aggregate score for correction

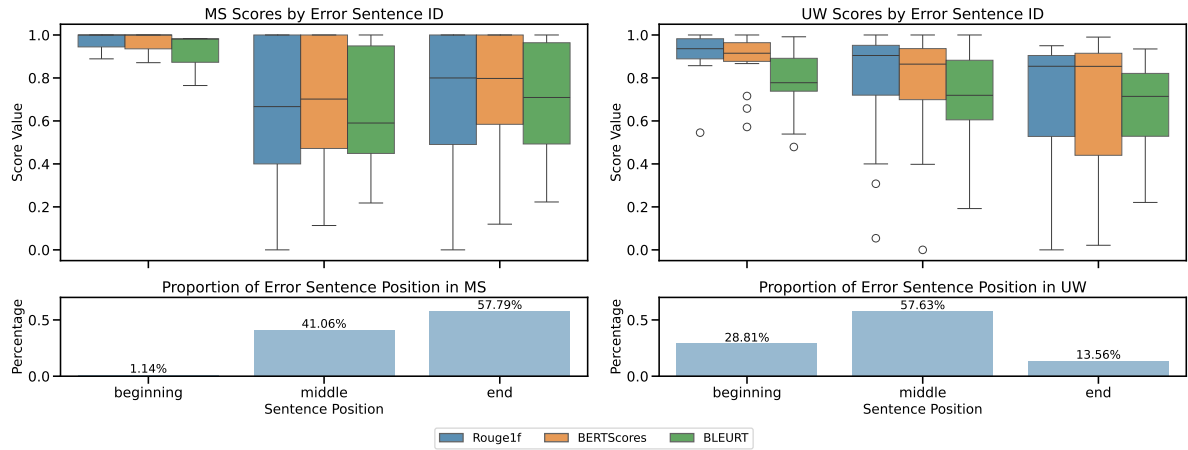


Figure 3: Boxplots of the distribution of ROUGE 1, BERTScore, and BLEURT with respect to the position of the error sentence for MS (left) and UW (right) datasets. “beginning” denotes that the error sentence is at index 0, “end” at the end, while “middle” is in between “beginning” and “end”.

reveals a higher score when the LLM-generated option was positioned as option A, achieving a score of 0.6492. This observation of *selection bias* echoes findings by previous studies (Pezeshkpour and Hruschka, 2023; Zheng et al., 2023).

5.2 Common LLM mistakes

We qualitatively evaluated the common mistakes found in the generated reasons and corrections.

Corrections of marginal effects LLMs occasionally make minor corrections to clinical notes that, although technically correct, do not significantly affect the correctness. Changes, such as altering “3” to “three” or fixing grammatical mistakes, might enhance readability but are not clinically significant. LLMs also tend to add adjectives, such as “acute” to “pyelonephritis”, adding specificity desirable in clinical settings but not always favourably reflected in NLG metrics.

Near-accurate corrections LLMs often suggest near-accurate corrections that lack the required specificity. For example, fixing an error sentence with the generic “antiplatelet therapy” instead of “aspirin” misses the required precision, even though aspirin is an antiplatelet therapy. Likewise, proposing to “Start anticoagulation therapy” instead of the more explicit “dalteparin” lacks specificity. These near-accurate adjustments underscore the difficulty LLMs encounter in achieving the specificity of the ground truth label.

Mistake due to incomplete context LLMs struggle to fix errors in clinical notes when details are

lacking. One example is when the LLM mistakenly suggests changing “pulmonary fibrosis” to “chronic obstructive pulmonary disease”. Both conditions share very similar early symptoms that are difficult to differentiate even for clinicians (Chilosi et al., 2012). Another example involves incorrectly adjusting a malnutrition patient’s Body Mass Index (BMI) from 30 to 18. Albeit a BMI of 18 signals malnutrition, it deviates from the ground truth label 13. These instances underscore the complexity of the MEDIQA-CORR task, as well as medical error correction in general which is very challenging to do without additional context even for human clinicians.

In summary, the sensitivity and qualitative analyses highlight the current limitations of LLMs in the clinical domain, which prompt further questions about the readiness of LLMs to be implemented in real-world clinical settings.

6 Conclusion

This study explores strategies for using LLMs to detect and correct medical error for the MEDIQA-CORR 2024 shared task. In addition to the comprehensive evaluation of prompting strategies based on different reasoning styles, we experiment with integrating error-span predictions from a fine-tuned model. Our best-performing system includes a fine-tuned BioLinkBERT-large for error-span prediction and GPT-4 for error correction. By harnessing LLMs’ generative abilities with 8-shot ICL and Brief CoT and presenting predicted error span as a hint in the prompt, our best-performing solu-

tion ranked sixth in the shared task leaderboard. Our post-hoc analyses offer insights into the use of LLM in medical error correction, including sensitivity to error location, role-playing bias, and common types of mistakes made by LLMs.

Limitations

The scope of our study was exclusively confined to GPT-based models, namely GPT-3.5 and GPT-4. The reported findings may differ across different types of LLMs. Furthermore, we independently explored various prompting strategies, such as CoT and MCQ prompt. We did not investigate the effect of integrating MCQ prompt with CoT reasoning. This unexplored combination may offer additional improvements in the LLM’s error correction capabilities.

Our post-hoc analyses also reveal a significant limitation of LLMs in clinical settings. Despite the advancements demonstrated through our proposed methodologies, the study underscores that LLMs may not be ready for deployment in real-world clinical environments without human oversight. The analysis highlights the critical need for human supervision, especially given the potential risks associated with inaccuracies in medical documentation and the consequent impacts on patient care. This limitation calls for further research into enhancing the reliability of LLMs as well as the evaluation metrics before considering their implementation in sensitive areas such as healthcare.

Acknowledgements

APG and CL were supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. PM was partially funded by ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence), EPSRC (grant no. EP/W002876/1), an industry grant from Cisco, and a donation from Accenture LLP; and is grateful to NVIDIA for the GPU donations. BA was partially funded by Legal and General PLC as part of the Advanced Care Research Centre and by the Artificial Intelligence and Multimorbidity: Clustering in Individuals, Space and Clinical Context (AIM-CISC) grant NIHR202639. For the purpose of open access, The authors have applied a Creative Commons attribution (CC BY) licence to any author-accepted manuscript version arising.

Experiments from this work are conducted mainly on the Edinburgh International Data Facility⁵ and supported by the Data-Driven Innovation Programme at the University of Edinburgh.

References

- Asma Ben Abacha, Wen wai Yim, Velvin Fu, Zhaoyi Sun, Fei Xia, and Meliha Yetisgen. 2024a. Overview of the mediqua-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*.
- Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. [An investigation of evaluation methods in automatic medical note generation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada. Association for Computational Linguistics.
- Marco Chilosi, Venerino Poletti, and Andrea Rossi. 2012. The pathogenesis of copd and ipf: distinct horns of the same devil? *Respiratory research*, 13:1–9.
- Matúš Falis, Aryo Pradipta Gema, Hang Dong, Luke Daines, Siddharth Basetti, Michael Holder, Rose S Penfold, Alexandra Birch, and Beatrice Alex. 2024. Can gpt-3.5 generate and code discharge summaries? *arXiv preprint arXiv:2401.13512*.
- Aryo Pradipta Gema, Luke Daines, Pasquale Minervini, and Beatrice Alex. 2023. Parameter-efficient fine-tuning of llama for the clinical domain. *arXiv preprint arXiv:2307.03042*.
- Aryo Pradipta Gema, Giwon Hong, Pasquale Minervini, Luke Daines, and Beatrice Alex. 2024. [Edinburgh clinical nlp at semeval-2024 task 2: Fine-tune your model unless you have access to gpt-4](#).
- Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2023. Using natural language explanations to improve robustness of in-context learning for natural language inference. *arXiv preprint arXiv:2311.07556*.
- Hyunjae Kim, Hyeon Hwang, Chaeun Lee, Minju Seo, Wonjin Yoon, and Jaewoo Kang. 2023. Exploring approaches to answer biomedical questions: From pre-processing to gpt-4 notebook for the bioasq lab at clef 2023. In *CEUR Workshop Proceedings*, volume 3497, pages 132–144. CEUR-WS.

⁵<https://edinburgh-international-data-facility.ed.ac.uk/>

- Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.
- Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.
- Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2024. In-context impersonation reveals large language models’ strengths and biases. *Advances in Neural Information Processing Systems*, 36.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. [BLEURT: learning robust metrics for text generation](#). *CoRR*, abs/2004.04696.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766*.
- Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. [LinkBERT: Pretraining language models with document links](#). In *Proceedings of the 60th Annual*

Parameter	Value
Model Name	gpt-3.5-turbo-0613
API Version	2023-03-15-preview
Temperature	0
Top P	0
Frequency Penalty	0
Presence Penalty	0
Max new token	256

Table 11: GPT-3.5 API call hyperparameters to generate Natural Language Explanations.

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A Experimental setup

All fine-tuning experiments were run on a single NVIDIA A100-40GB GPUs. We used the HuggingFace’s transformer library (Wolf et al., 2020). The validation set was utilised to determine the best checkpoint.

In-context examples were retrieved from the Training set. Additionally, the validation set was used to evaluate and select the optimal prompt design. For the test submission, we also retrieved In-context examples from the MS and UW validation sets.

B Hyperparameters

B.1 GPT-3.5 Hyperparameters for the generation of Natural Language Explanation

We prompted GPT-3.5 (model name: gpt-3.5-turbo-0613) with hyperparameters as shown in Table 11. The generation process took approximately 2 hours and cost \$2.

B.2 GPT-4 generation hyperparameters

During inference on the test set, we prompted GPT-4 (model name: gpt-4-turbo) as shown in Figure 1 Step 3. We set temperature=0 to ensure that the model’s generation is deterministic. The maximum generation length is 512, allowing longer

CoT reasons. One generation process took approximately 2 hours and cost \$35.

C Prompt Examples

Here, we provide examples of the prompts used in our experiments. The black text within the box represents the prompt input text, the red text represents the prediction of the models, and the blue text represents the ground truth.

C.1 Prompt for In-Context Learning and Chain-of-Thought

C.1.1 System Prompt

System Prompt for the In-Context Learning and Chain-of-Thought Experiments

You are a clinician assistant tasked with reviewing clinical texts that have been corrupted by an evil third party. Each clinical text may contain either one perturbed sentence with clinical or factual inaccuracies, or no errors at all. Your objective is to help the clinician in identifying any perturbed sentence and provide a correction. Please respond in JSON format with the following structure:

- 'reason': Describe the reasoning behind identifying a specific sentence as incorrect or all clinical text as correct.
- 'incorrect_sentence_id': If you identify an incorrect sentence, provide its ID here. If all sentences are correct, respond with -1.
- 'correction': If an incorrect sentence is identified, provide a corrected sentence or an explanation of the inaccuracy. If all sentences are correct, respond with 'NA'.

When evaluating the text, focus specifically on clinical or factual inaccuracies. This could include incorrect medical information, factual errors related to patient care, or erroneous data interpretations. Your detailed assessment and correction are critical for ensuring the reliability and accuracy of our clinical documentation. You have to be more action-oriented, don't suggest planning, discussion, or something that is not action-oriented.

C.1.2 CoT reasons

Brief Chain-of-Thought

Streptococcus pneumoniae is not typically associated with skin lesions with a blue livid center, pale intermediate zone, and a dark red peripheral rim. *Mycoplasma pneumoniae* is a common cause of atypical pneumonia and can present with skin rashes like the ones described in the text.

Long Chain-of-Thought

The incorrect sentence is sentence 24, which states 'Pathology confirms *Streptococcus pneumoniae* is the causal organism.' This is inaccurate because *Streptococcus pneumoniae* is a common cause of community-acquired pneumonia, but in this case, the clinical presentation with skin lesions and the absence of a productive cough suggests a more likely diagnosis of atypical pneumonia caused by *Mycoplasma pneumoniae*. *Mycoplasma pneumoniae* is known to cause atypical pneumonia with extrapulmonary manifestations such as skin lesions, which aligns better with the patient's presentation. Therefore, the correction should be 'Pathology confirms *Mycoplasma pneumoniae* is the causal organism.'

SOAP Chain-of-Thought

Subjective: The patient presents with fever, nausea, nonproductive cough, headaches, generalized fatigue, muscle and joint pain, and increasing shortness of breath.

Objective: Vital signs show a slightly elevated temperature, normal pulse, respirations, and blood pressure. Skin lesions with specific characteristics are noted on the extremities. Laboratory studies reveal hemoglobin, leukocyte count, electrolyte levels, and other values. An x-ray of the chest is performed.

Assessment: The patient likely has an infectious process given the symptoms and laboratory findings.

Plan: Further evaluation and treatment are warranted.

Inconsistency: The incorrect sentence is in the Assessment section where *Streptococcus pneumoniae* is mentioned as the causal organism, which is inconsistent with the typical causative agent for atypical pneumonia, which is *Mycoplasma pneumoniae*.

C.1.3 Chain-of-Thought Prompt

Brief Chain-of-Thought Prompt Example

«System Prompt»

«In-Context Learning Examples»

Clinical text:

0 A 78-year-old man is brought in to the emergency department by ambulance after his wife noticed that he began slurring his speech and had developed facial asymmetry during dinner approximately 30 minutes ago.

1 His past medical history is remarkable for hypertension and diabetes.

2 His temperature is 99.1 F (37.3 C), blood pressure is 154/99

3 mmHg, pulse is 89/min, respirations are 12/min, and oxygen saturation is 98% on room air.

4 CTA of the head is obtained after neurologic exam reveals right upper and lower extremity weakness and an asymmetric smile.

Task: Identify 1 incorrect sentence in the clinical text, or confirm if all sentences are correct. If there is 1 incorrect sentence, how would you fix it? Follow the hints below if you see fit

Hint:

- Pay special attention to biomedical entities such as chief complaints, medical exams, diagnoses, and treatments. The mistake often lies within these areas.
- The clinician said that you MAY want to pay attention to the mention of 'CTA of the head'. If you believe that the mention is incorrect, ONLY SWAP this mention with something more probable. DO NOT MODIFY the sentence in any other way.

Let's think step by step

Answer:

```
{
  "reason": "The mention of 'CTA of the head' is not the most appropriate imaging study for the patient's presentation of acute stroke symptoms with slurred speech, facial asymmetry, and extremity weakness. A non-contrast CT of the head is typically the initial imaging study to evaluate for acute intracranial haemorrhage or ischemic stroke. Therefore, replacing 'CTA of the head' with 'CT of the head' would align better with the standard diagnostic approach for acute stroke evaluation.",
  "incorrect_sentence_id": "4",
  "correction": "CT of the head is obtained after neurologic exam reveals right upper and lower extremity weakness and an asymmetric smile.",
}
```

C.2 Option Generation Prompt

Multiple-Choice Question Prompt

C.2.1 MCQ 2 options

Prompt Used to Generate MCQ Answer Options

Your job is to review a clinical note that potentially contains a medical error.

In the following clinical note, what should the <BLANK> in the sentence "Suspected of <BLANK>." be replaced with if "primary ciliary dyskinesia" is incorrect? Do not answer with "primary ciliary dyskinesia" or its medical synonyms in your answer. Output your response in JSON format, with keys 'option'.

Clinical note:

A 4-year-old boy is brought to the physician in December for episodic shortness of breath and a nonproductive cough for 3 months. These episodes frequently occur before sleeping, and he occasionally wakes up because of difficulty breathing. His mother also reports that he became short of breath while playing with his friends at daycare on several occasions. He is allergic to peanuts. He is at the 55th percentile for height and weight. Vital signs are within normal limits. Examination shows mild scattered wheezing in the thorax. An x-ray of the chest shows no abnormalities. Suspected of <BLANK>.

Generated answer:

```
{
  "option": "asthma"
}
```


C.2.2 MCQ 4 options

Prompt Used to Generate MCQ Answer Options

Your job is to review a clinical note that potentially contains a medical error.

In the following clinical note, what should the <BLANK> in the sentence “Suspected of <BLANK>.” be replaced with if “primary ciliary dyskinesia” is incorrect? Do not answer with “primary ciliary dyskinesia” or its medical synonyms in your answer. Output your response in JSON format, with keys ‘option_1’, ‘option_2’ and ‘option_3’.

Clinical note:

A 4-year-old boy is brought to the physician in December for episodic shortness of breath and a nonproductive cough for 3 months. These episodes frequently occur before sleeping, and he occasionally wakes up because of difficulty breathing. His mother also reports that he became short of breath while playing with his friends at daycare on several occasions. He is allergic to peanuts. He is at the 55th percentile for height and weight. Vital signs are within normal limits. Examination shows mild scattered wheezing in the thorax. An x-ray of the chest shows no abnormalities. Suspected of <BLANK>.

Generated answer:

```
{
  "option_1": "asthma"
  "option_2": "bronchiolitis"
  "option_3": "pulmonary embolism"
}
```

C.3 Inference Prompt Multiple-Choice Question Prompt

C.3.1 MCQ 2 options

Inference Prompt for Multiple-Choice Question style with 2 options

Your job is to review a clinical note that potentially contains a medical error.

In the following clinical note, what should the <BLANK> in the sentence “Suspected of <BLANK>.” be replaced with for it to be medically informative and accurate? Choose one from the options given below. Output your response in JSON format, with a key ‘Answer’.

Clinical note:

A 4-year-old boy is brought to the physician in December for episodic shortness of breath and a nonproductive cough for 3 months. These episodes frequently occur before sleeping, and he occasionally wakes up because of difficulty breathing. His mother also reports that he became short of breath while playing with his friends at daycare on several occasions. He is allergic to peanuts. He is at the 55th percentile for height and weight. Vital signs are within normal limits. Examination shows mild scattered wheezing in the thorax. An x-ray of the chest shows no abnormalities. Suspected of <BLANK>.

Options:

- A. asthma
- B. primary ciliary dyskinesia

Generated answer: {

```
"Answer": "A. asthma"
}
```

C.3.2 MCQ 4 options

Inference Prompt for Multiple-Choice Question style with 4 options

Your job is to review a clinical note that potentially contains a medical error.

In the following clinical note, what should the <BLANK> in the sentence “Culture tests indicate <BLANK>.” be replaced with for it to be medically informative and accurate? Choose one from the options given below. Output your response in JSON format, with a key ‘Answer’.

Clinical note:

A 4-year-old boy is brought to the physician in December for episodic shortness of breath and a nonproductive cough for 3 months. These episodes frequently occur before sleeping, and he occasionally wakes up because of difficulty breathing. His mother also reports that he became short of breath while playing with his friends at daycare on several occasions. He is allergic to peanuts. He is at the 55th percentile for height and weight. Vital signs are within normal limits. Examination shows mild scattered wheezing in the thorax. An x-ray of the chest shows no abnormalities. Suspected of <BLANK>.

Options:

- A. asthma
- B. primary ciliary dyskinesia
- C. bronchiolitis
- D. pulmonary embolism

Generated answer: {
"Answer": "A. asthma"
}

	MS			UW		
	ROUGE 1	BERTScore	BLEURT	ROUGE 1	BERTScore	BLEURT
H	6.0749	5.0249	7.2848	5.6821	3.6073	2.3457
p	0.0480	0.0811	0.0262	0.0584	0.1647	0.3095

Table 12: Summary of Kruskal-Wallis H-Test results for sentence position impact on ROUGE 1, BERTScore, and BLEURT metrics. Statistically significant differences ($p < 0.05$) are highlighted in cyan.

	MS			UW		
	ROUGE 1	BERTScore	BLEURT	ROUGE 1	BERTScore	BLEURT
beginning-middle	0.1751	0.3121	0.1389	0.3596	0.3464	0.7118
middle-end	0.2137	0.2479	0.1192	0.5251	1.0000	1.0000
beginning-end	0.3923	0.6258	0.3586	0.0609	0.2849	0.4757

Table 13: Summary of Post-hoc Dunn’s Test results for sentence position impact on ROUGE 1, BERTScore, and BLEURT metrics. No significant differences observed.

D Statistics of “Sensitivity to the position of error sentence in the clinical note”

The analysis was split into two main tests: the Kruskal-Wallis H-Test to identify overall differences across sentence positions and the Post-hoc Dunn’s Test to investigate pairwise differences between sentence positions.

The Kruskal-Wallis H-Test was applied to compare the distributions of scores for ROUGE 1, BERTScore, and BLEURT across three sentence positions (beginning, middle, end) within clinical notes from the validation sets of MS and UW. As shown in Table 13, statistically significant differences were found in the MS dataset for ROUGE 1 and BLEURT metrics, suggesting sensitivity to sentence positioning.

Following the Kruskal-Wallis H-Test, a Post-hoc Dunn’s Test was performed to conduct pairwise comparisons between sentence positions for each evaluation metric. The Post-hoc Dunn’s Test revealed no statistically significant differences between any pairwise comparisons of sentence positions for all evaluated metrics, suggesting that while overall differences exist, specific pairwise comparisons did not reach statistical significance.