HSE NLP Team at MEDIQA-CORR 2024 Task: In-Prompt Ensemble with Entities and Knowledge Graph for Medical Error Correction

Airat Valiev
HSE University / Moscow, Russia
aa.valiev@hse.ru

Elena Tutubalina
Kazan State University / Kazan, Russia
HSE University / Moscow, Russia
tutubalinaev@gmail.com

Abstract

This paper presents our LLM-based system designed for the MEDIQA-CORR @ NAACL-ClinicalNLP 2024 Shared Task 3, focusing on medical error detection and correction in medical records. Our approach consists of three key components: entity extraction, prompt engineering, and ensemble. First, we automatically extract biomedical entities such as therapies, diagnoses, and biological species. Next, we explore few-shot learning techniques and incorporate graph information from the MeSH database for the identified entities. Finally, we investigate two methods for ensembling: (i) combining the predictions of three previous LLMs using an AND strategy within a prompt and (ii) integrating the previous predictions into the prompt as separate 'expert' solutions, accompanied by trust scores representing their performance. The latter system ranked second with a BERTScore score of 0.8059 and third with an aggregated score of 0.7806 out of the 15 teams' solutions in the shared task.

1 Introduction

Medical records play a crucial role in healthcare systems as they capture essential patient information, including diagnoses, treatments, and outcomes. Medical texts are characterized by complex terminology, context-specific knowledge, and significant implications. Detecting and rectifying errors within clinical notes necessitates domain expertise and reasoning. This task presents a complex challenge that demands precise analysis and understanding of the medical domain.

In recent years, Large Language Models (LLMs) have revolutionized the field of natural language processing (NLP) by demonstrating unprecedented performance across a wide range of tasks. These models, often based on Transformer (Vaswani et al., 2017; Devlin et al., 2018), have become the cornerstone of modern NLP research (Pan et al., 2023). LLMs excel in key areas such as semantic understanding and contextualization (Radford et al., 2018), multimodal capabilities (Livne et al., 2023), few-shot and zero-shot learning (Dang et al., 2022), as well as various medical applications including disease diagnosis (Schubert et al., 2023), drug discovery (Livne et al., 2023), and medical records processing (Guevara et al., 2024).

Automated fact-checking has garnered significant attention due to the escalating challenge posed by misinformation. Traditionally, fact-checking has relied on manual verification conducted by human experts, primarily focusing on general-domain texts like Wikipedia articles and news reports (Zhang and Gao, 2023; Quelle and Bovet, 2024). Recently, LLMs have offered the capability to analyze false statements and provide an assessment of their factual accuracy by leveraging their pre-trained knowledge and contextual understanding (Wang and Shu, 2023; Guan, 2021; Lewis et al., 2020; Chen et al., 2021). Several methodologies have been proposed to enhance the overall performance in LLMs, and the most notable ones are Chain of Thought (CoT) (Zhang, 2023).

In this work, we utilize several key approaches for medical records correction using LLMs (see Figure 1). These approaches include entity extraction and normalization (Miftahutdinov et al., 2020, 2021; Sung et al., 2022), few-shot learning techniques (Brown et al., 2020), graph-based knowledge incorporation (Fei et al., 2021), and ensembling strategies (Wang et al., 2022). We investigate the application of these approaches to enhance the accuracy of medical error correction.

The paper is organized as follows. Section 1 presents shared task and data overview. We describe our approach with three key components and state-of-the-art (SoTA) models in Section 2. Experiments with baselines and our model are presented in Section 3.3.4. Finally, we discuss the results and conclude the work in Sections 4 and 5, respectively.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 470–482

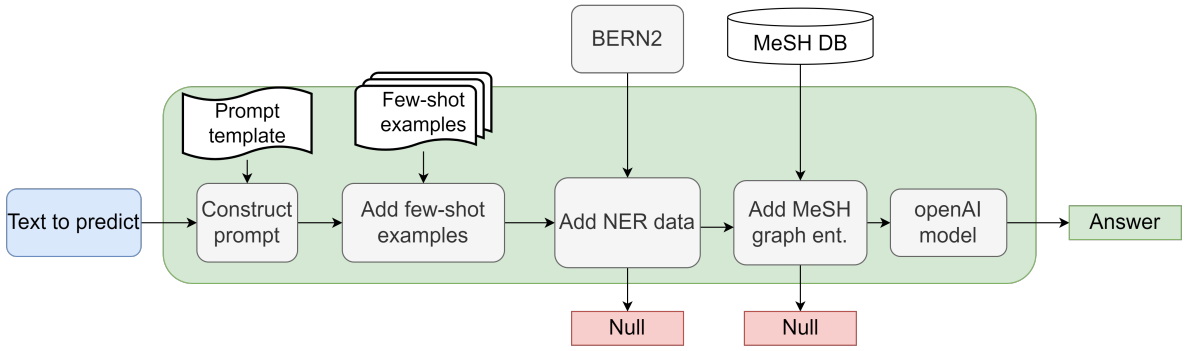

Figure 1: The system overview. The process can be described as follows: the system begins by receiving medical text as input. Initially, a prompt template is utilized, supplemented with a small number of few-shot examples (either 2 or 5). The Named Entity Recognition (NER) model is then employed to identify and extract named entities within the large language model's context. Subsequently, potential replacements for these extracted entities are sought within the Medical Subject Headings (MeSH) thesaurus. The prompt, enriched with these replacements, is passed to the selected OpenAI model. Finally, the model's output is returned and stored in the prediction file. This constitutes the overall operation of the system.

2 Task and Data Overview

The MEDIQA-CORR 2024 shared task (Ben Abacha et al., 2024) focuses on analyzing snippets of clinical text to address specific subtasks related to medical error detection and correction. These subtasks include:

1. Binary Classification: The first subtask involves determining whether the given clinical text contains a medical error. This step requires evaluating the accuracy, consistency, and factual correctness of the information presented in the text.

2. Span Identification: If a medical error is identified, the goal is to locate the specific text span associated with the error. This step is crucial for precisely pinpointing the erroneous segment within the clinical text.

3. Natural Language Generation (Correction): Once the medical error is identified and its location is determined, the task is to generate a free-text correction for the identified error. The generated correction should be contextually appropriate, accurate, and concise, effectively addressing the error in the clinical text.

We focus on the latter subtask, which encompasses all three subtasks mentioned.

The dataset provided by the organizers, known as the MS Training Set, consists of 2,189 clinical texts. Additionally, there is the 'MS' Validation Set comprising 574 clinical texts and the 'UW' Validation Set comprising 160 clinical texts (Abacha et al., 2024). The test portion of the dataset is formed by combining clinical texts from both collections. Each clinical text in the dataset is labeled as either correct or containing one error. More formally, the task involved in this dataset is as follows:

1. Predicting whether a given text contains an error or not. The error flag is represented by 1 if the text contains an error and 0 if it is error-free.

2. For texts flagged as containing errors, extract the sentence that contains the error.

3. Generating a corrected version of the identified error sentence.

3 Method

The error correction method of Figure 1, proposed in the current work, is straightforward and consists of three major steps: we first prepare the data to make predictions: extract named entities from texts, and search for the term replacements. Then we form the prompt for the model from the template, add a few examples and additional data (NER results, MeSH terms), and then use LLMs to make predictions.

3.1 Data preparation

Let us first discuss the first step. Before predicting, some preparations were made with the input texts, including Named Entity Recognition (NER), and

Figure 2: An example of the result obtained from Named Entity Recognition (NER).

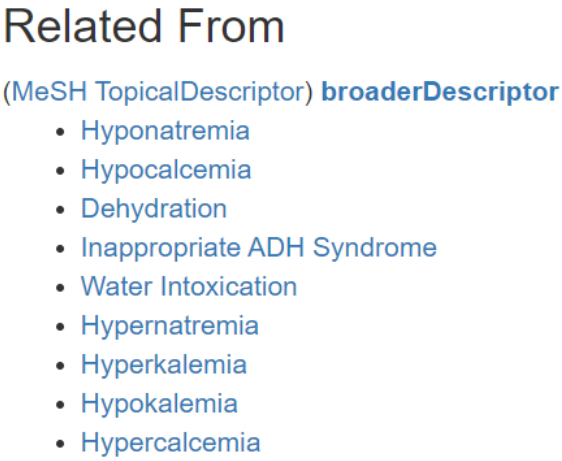

Figure 3: An example for the term D014883 (water-electrolyte imbalance) related entities, extracted from the MeSH database.

possible term replacement data extraction from the MeSH thesaurus.

3.1.1 Biomedical entities

Biomedical concepts, such as diseases, symptoms, drugs, genes, and proteins, are critical for many biomedical applications, including drug discovery (Khrabrov et al., 2022), clinical decision making (Sutton et al., 2020; Peiffer-Smadja et al., 2020), and biomedical research (Lee et al., 2016; Tutubalina et al., 2017; Soni and Roberts, 2021; Sakhovskiy et al., 2021; Sakhovskiy and Tutubalina, 2022; Miftahutdinov et al., 2020, 2021).

For NER, we use the BERN2 (Advanced Biomedical Entity Recognition and Normalization) model (Sung et al., 2022) is a neural biomedical named entity recognition and normalization tool. BERN2 significantly improves upon its predecessor (Kim et al., 2019) by employing a multi-task NER model and neural network-based entity linking (EL) models, resulting in faster and more accurate inference.

Using this tool, we extracted named entities (with MeSH identifiers) such as diagnosis, therapy, biological species, and medical entities. You can see an example of such extraction in Figure 2.

3.1.2 MeSH: Medical Subject Headings

MeSH is a hierarchically organized and concept-based vocabulary produced by the National Library of Medicine (NLM) (Mao and Lu, 2017). Its primary purpose is to facilitate indexing, cataloging, and searching of biomedical and health-related information. MeSH plays a crucial role in various NLM databases, including MEDLINE/PubMed and the NLM Catalog. MeSH consists of standardized keywords that describe the subject matter of journal articles, clinical notes, and other biomedical texts. These terms are carefully curated and organized to ensure consistency and accuracy. Researchers, librarians, and information specialists use MeSH to index and retrieve relevant literature. By assigning MeSH terms to documents, they enhance search precision and recall. MeSH thesaurus could be applied to perform Biomedical Literature Indexing (like in MEDLINE/PubMed (von Korff, 2022)), Concept Mapping, and Synonyms (MeSH provides a standardized way to map synonyms and related terms, for different synonyms of a medical condition to be linked to a single MeSH term), and investigating cross-lingual clinical entity linking using MeSH concepts. Highlights the importance of MeSH in linking biomedical entities across languages. MeSH serves as a foundational resource for organizing and accessing biomedical knowledge. Its controlled vocabulary ensures consistency and precision, benefiting researchers, clinicians, and information professionals.

In the presented work, we use the MeSH database to perform the knowledge graph search - for the extracted entities with available MeSH IDs, we've found their possible replacements (the example is in Figure 3) (other entities on the same relation level with the parent term node) to present them to the LLM as clues about possible errors in a text.

3.2 Dataset description

The statistical data about the dataset can be seen from the Table 1. In total, 2,923 texts (2,189 texts

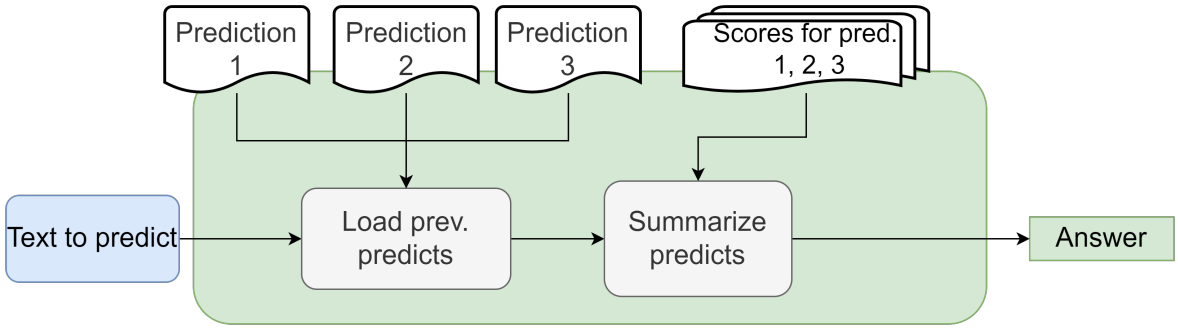

Figure 4: The solution ensembling overview. In this approach, we use previous predictions of different models for each input text and resulting prediction scores, and a new template. We evaluated three major ensembling strategies, including AND (all three models found an error), majority of votes, and weighted approach (weight prediction by each prediction score), in the validation stage, but decided to make the final prediction using AND strategy.

	Train	Val MS	Val UW	Test
Texts	2 189	574	160	925
NER ent.	3,3	3,3	3,3	6,5
MeSH terms	2,1	2,1	2,1	2,2

Table 1: Dataset statistics by the number of texts and found entities.

in the train part + 574 texts in the MS validation part + 160 texts in the UW validation part), the BERN2 model found 9,682 named entities with MeSH IDs, an average of 3.3 entities per single text. An average of 2.1 MeSH term replacements were found using MeSH graph search. The train part included 1,219 texts with errors and 970 correct entries, the MS validation part consisted of 80 correct and 80 with errors, and the UW validation included 319 entries with errors and 255 correct.

The test data part consisted of 925 text entries. During the test part processing, the BERN2 model extracted 6,032 MeSH IDs (avg. 6.5 terms per text), with an average of 2.2 replacements extracted from the MeSH thesaurus.

3.3 Making predictions

After the preparation step, we move forward to make the predictions and find the texts with medical errors. We have studied and used three general LLM-based approaches for prediction making:

1. Ordinary prompting (2-shot and 5-shot)
2. Prediction ensembling (ensemble of 3 solutions)
3. In-prompt ensembling (expert opinions with trust scores)

In this section, we first discuss the ordinary solution with different OpenAI models (GPT3.5-turbo, GPT4, GPT4-turbo preview) (Yenduri et al., 2022) and simple prompts. These models continually improve the instruction following ability and have broader general knowledge and advanced reasoning capabilities. The solution idea is simple, as we discussed earlier: the model receives the prompt prefix containing the behavior rules for the model (see Appendix 1), 2 of 5 examples (texts and expected output from the training dataset part), and the text to analyze along with the NER information (found named entities) and replacement entities from the MeSH graph. All significant parts of the template are highlighted in color. A few shot examples fixed set (2 or 5) were selected from the Train data split to present the data with and without corrections needed equally.

3.3.1 Ordinary prompting (2-shot)

The first solution (as illustrated in Figure 1), with the 2-shot template, consists of a prefix (2-shot prompt prefix from Appendix A1), text to predict, and additional data: a list of found named entities with additional info from the BERN2 model.

3.3.2 Ordinary prompting (5-shot)

The second solution with the 5-shot prompt template, is constructed from a 5-shot prompt prefix from Appendix A.1.3, the text to predict and additional data: NER results and MeSH graph data with possible entity replacements. The process scheme is illustrated in Figure 1.

3.3.3 Prediction ensembling

Decision ensembling is different from previously discussed approaches. In this variant, as it is

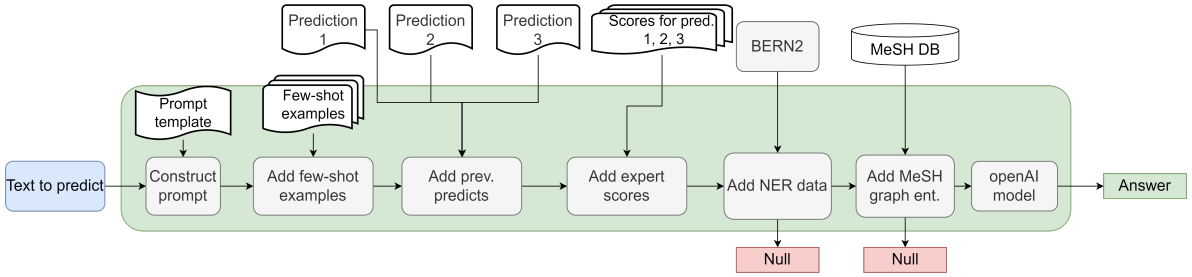

Figure 5: The system overview. The medical text inputs into the system. First of all, we use the prompt template and add 2 or 5 few-shot examples. The NER model finds named entities for the large language model. Then we find possible replacements for extracted entities in the MeSH thesaurus. Here we also use previous predictions of different models for each input text and resulting prediction scores, adding these 'expert' opinions with expert trust scores, to the prompt. The generated prompt is passed to the openAI model of our choice, and we return the result to the prediction file.

shown in Figure 4, we simply construct the prediction from the three top-score previous predictions, based on AND strategy: for each text entry we decide the error exists, if only the error is found in all the three previous predictions - in this case we include the error sentence number and correction from the previous prediction with the highest score. If at least one model has predicted this sentence as correct, we count it as containing no errors. This strategy slightly improved the resulting score: 0.62 -> 0.64.

3.3.4 In-prompt ensembling

In this approach, we have combined the idea of basic prompting, few-shot learning, and an ensemble of experts. We again add information about NER entities and MeSH graph replacements, but because of the ensembling approach evaluated, we also include predictions from the top three previous submissions (model predictions with the highest score), calling it 'expert's solutions'. We also add three expert trust scores - these are the test scores for these submissions, to help the model estimate the expert opinion correctness indirectly.

We added the test predictions of the three previous models. Still, in the case of a real data evaluation, this ensemble could be formed from the three different models and their predictions, and trust scores could be obtained from the validation scores.

The result of this ensemble addition could be the following: "Expert 1 with trust score (weight1): (outputs1), expert 2 with trust score (weight2): (outputs2), expert 3 with trust score (weight): (outputs3)." Prompt prefix (Appendix A.2.3) and process scheme 5 are included. The ensemble example with the real data is the following:

- Expert 1 with trust score 0,72: "Error exists: ||Yes|| Correction: || Patient's symptoms are suspected to be due to acute gastroenteritis.|| Error sentence number: ||10||",
- Expert 2 with trust score 0,69: "Error exists: ||Yes|| Correction: || Patient's symptoms are suspected to be due to typhoid fever.|| Error sentence number: ||10||",
- Expert 3 with trust score 0,68: "Error exists: ||No|| Correction: ||None|| Error sentence number: ||None||"

4 Baselines

During the model development and preparation, we explored various baselines. In addition to the above-mentioned methods, we initially investigated a simpler BERT-based approach (Devlin et al., 2018) and utilized other LLMs such as self-hosted LLaMA-based Med42 70b (Christophe et al., 2023) and Meditron 7b (Chen et al., 2023).

The BERT model, specifically the PubMedBERT-base checkpoint (Gu et al., 2021), was trained for 10 epochs on a subset of the training data. However, it performed poorly on the validation data, achieving a score of approximately 0.57 even on the task of text classification for error presence, which is a binary classification problem. This subpar performance can be attributed to a limited number of training examples and the wide variation in replaceable terms and diverse themes found in medical texts. Due to these unsatisfactory results in the validation phase, we decided not

Table 2: Evaluation results. Here 'ens' stands for an ensemble of 3 previous solutions and these predict scores, 'NER' - for named entities from the text, and 'MeSH' - for the related terms from the MeSH thesaurus. The general approach is shown in Figure 1, the prediction ensemble - in Figure 4, and an ensemble of experts in Figure 5.

Base model name	Prompt	Additional data	AggrScore	R1F	BERTScore	BLEURT	AggrC
gpt-3.5t	General 2-shot	NER	0.31	0.35	0.38	0.34	0.24
gpt-4-t-1401-preview	5-shot	NER	0.55	0.55	0.55	0.55	0.41
gpt-4-t-preview-0125	5-shot	NER + meSH	0.62	0.62	0.60	0.62	0.53
-	-	Ens. of 3 predicts	0.64	0.64	0.62	0.63	0.54
gpt-4-t-preview-0125	Ensemble prompt	NER+ens	0.68	0.68	0.67	0.68	0.52
gpt-4-t-0125-preview	Ensemble prompt	NER+ens+MeSH	0.69	0.71	0.67	0.69	0.51
gpt-4	Ensemble prompt	NER+ens+MeSH	0.72	0.74	0.69	0.72	0.55
gpt-4-t-0125-preview	Ensemble prompt	NER+ens+MeSH	**0.78**	**0.81**	**0.76**	**0.78**	**0.51**

to proceed with evaluating the model's precision on the test data and instead moved on to explore alternative solution methods.

The LLaMA-based models exhibited better performance and were successful in identifying and correcting misplaced terms, achieving an aggregated score of approximately 0.43 on the validation data. However, these models disregarded certain in-prompt rules and ensemble solutions. Consequently, despite not showing any positive performance improvements with the addition of NER data and graph entities, they were excluded from the test submission.

5 Experiments and Results

The evaluation results of our error correction systems are shown in Table 2. The aggregate score is the main evaluation score to rank the participating systems. We've used the following scripts[1] for evaluation. More specifically about the metrics used for evaluation:

- NLG (Natural Language Generation) metrics: ROUGE(Lin, 2004), BERTScore (Zhang et al., 2019), BLEURT(Sellam et al., 2020), their Aggregate-Score (Mean of ROUGE-1-F, BERTScore, BLEURT-20), and their Composite Scores (AggrC) for the evaluation of Sentence Correction.

- The Composite score is the mean of individual scores computed as follows for each text:
  - 1 point if both the system correction and the reference correction are "NA";
  - 0 point if only one of the system or the reference is "NA".

- NLG metrics value in [0, 1] range (e.g., ROUGE, BERTScore, BLEURT, or Aggregate-Score) if both the system correction and reference correction are non-"NA" sentences.

- The Aggregate score is the main evaluation score to rank the participating systems(Abacha et al., 2023).

As we can see from table 2, we can observe that the more powerful language model, 'sophisticated' prompting, and additional data presented to the language model lead to better results: results improved from 0.62 to 0.72 and finally to 0.78, which is a Top-3 solution of an entire competition. One also can see that additional examples (2 vs 5 texts) in the few-shot section also increase performance: 0.31 vs 0.55. Also, the in-prompt ensembling technique improves final results greatly because the model can see the solutions from previous runs along with the scores for these runs, and correct the current prediction, which leads to more stable and reliable predictions and error corrections. We also could see the obvious trend of better performance with more complicated models: GPT 4 outperforms GPT 3.5 Turbo, and GPT 4 Turbo preview beats the ordinary GPT 4: 0.31 vs 0.55 vs 0.62, respectively.

The methodology delineated herein possesses the potential for expansion and further refinement through the incorporation of techniques such as the

[1] https://github.com/abachaa/MEDIQA-CORR-2024/tree/main/evaluation

Knowledge Graph, PromptKG (Xie et al., 2022), the meta-prompting approach, and the Chain of Thought (CoT) approach. Additionally, the integration of specialized models, specifically designed for error detection and error span identification, into the model pipeline could be achieved directly by utilizing the chaining techniques (e.g. langchain). This would serve to enhance the robustness and accuracy of the overall system.

6 Conclusion

In this work, we have addressed the issue of identifying and resolving error text in biomedical texts. We have proposed a system for the MEDIQA-CORR shared task by utilizing prompting, ensembling techniques, and LLMs. Our approach demonstrates that the problem can be solved using ordinary GPT models without pre-training, relying solely on in-context learning, along with the NER model and additional MeSH knowledge graph data. By employing an in-prompt ensemble of LLMs as experts and incorporating data from the MeSH knowledge graph and NER results, we achieved a high task aggregated score of 0.78059, securing the 3rd position on the official competition leaderboard. Our results highlight the effectiveness of our proposed prompting approach while also indicating areas for future improvement. Utilizing more advanced tools like full-scale RAG and fine-tuned biomedical LLMs could potentially enhance the quality of error correction. In addition, we plan to make all our code and data publicly accessible shortly after the publication of our paper.

Acknowledgments

The work of E.T. has been supported by the Russian Science Foundation grant # 23-11-00358. We would also like to thank the organizers of the MEDIQA-CORR task and the anonymous reviewers for their comments on this paper.

Limitations

Large Language Models (LLMs) have emerged as powerful tools for natural language understanding and generation. However, their effectiveness hinges on the quality and diversity of their training data. LLMs are typically trained on vast corpora of text from the internet, making manual curation infeasible. Consequently, they inherit any biases, inaccuracies, or limitations in their training data. Additionally, the success of LLMs has led to the generation of online content by these models, which may introduce hallucinated information.

Ethics Statement

Using databases for retrieval carries a drawback in that these sources may lack comprehensiveness and contain inaccuracies. LLMs are not immune to representation biases and the risk of producing potentially misleading outcomes, particularly in the healthcare sector.

All pre-trained language models and datasets utilized in this study are openly accessible for research purposes.

We honor and support the ACL Code of Ethics.

References

Asma Ben Abacha, Wen wai Yim, Velvin Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2024. 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 metrics for automated medical note generation. *arXiv preprint arXiv:2305.17364*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are factuality checkers reliable? adversarial meta-evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2082–2095, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Clément Christophe, Avani Gupta, Nasir Hayat, Praveen Kanithi, Ahmed Al-Mahrooqi, Prateek Munjal, Marco Pimentel, Tathagata Raha, Ronnie Rajan, and

Shadab Khan. 2023. Med42 - a clinical large language model.

Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. 2021. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in bioinformatics*, 22(3):bbaa110.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Shuo Guan. 2021. Knowledge and keywords augmented abstractive sentence summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 25–32, Online and in Dominican Republic. Association for Computational Linguistics.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. 2024. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1):6.

Kuzma Khrabrov, Ilya Shenbin, Alexander Ryabov, Artem Tsypin, Alexander Telepov, Anton Alekseev, Alexander Grishin, Pavel Strashnov, Petr Zhilyaev, Sergey Nikolenko, and Artur Kadurin. 2022. nablaDFT: Large-Scale conformational energy and hamiltonian prediction benchmark and dataset. *Phys. Chem. Chem. Phys.*, 24(42):25853–25863.

Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, 7:73729–73740.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. 2016. Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10):e0164680.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, and Alex Zhavoronkov. 2023. nach0: Multimodal natural and chemical languages foundation model. *arXiv preprint arXiv:2311.12410*.

Yuqing Mao and Zhiyong Lu. 2017. Mesh now: automatic mesh indexing at pubmed scale via learning to rank. *Journal of biomedical semantics*, 8:1–9.

Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2020. On biomedical named entity recognition: experiments in interlingual transfer for clinical and social media texts. In *European Conference on Information Retrieval*, pages 281–288. Springer.

Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, and Elena Tutubalina. 2021. Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics*, 37(21):3856–3864.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Nathan Peiffer-Smadja, Timothy Miles Rawson, Raheelah Ahmad, Albert Buchard, P Georgiou, F-X Lescure, Gabriel Birgand, and Alison Helen Holmes. 2020. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, 26(5):584–595.

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7:1341697.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Andrey Sakhovskiy, Zulfat Miftahutdinov, and Elena Tutubalina. 2021. KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal BERT-based models for adverse drug effects. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 39–43, Mexico City, Mexico. Association for Computational Linguistics.

Andrey Sakhovskiy and Elena Tutubalina. 2022. Multimodal model with text and drug embeddings for adverse drug reaction classification. *Journal of Biomedical Informatics*, 135:104182.

Marc Cicero Schubert, Wolfgang Wick, and Varun Venkataramani. 2023. Large language model-driven evaluation of medical records using medcheckllm. *medRxiv*, pages 2023–11.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Sarvesh Soni and Kirk Roberts. 2021. An evaluation of two commercial deep learning-based information retrieval systems for COVID-19 literature. *J. Am. Medical Informatics Assoc.*, 28(1):132–137.

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. Bern2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

EV Tutubalina, Z Sh Miftahutdinov, RI Nugmanov, TI Madzhidov, SI Nikolenko, IS Alimova, and AE Tropsha. 2017. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects. *Russian Chemical Bulletin*, 66:2180–2189.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Modest von Korff. 2022. Exhaustive indexing of PubMed records with medical subject headings. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10)*, pages 8–15, Marseille, France. European Language Resources Association.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. *arXiv preprint arXiv:2310.05253*.

Lijing Wang, Timothy Miller, Steven Bethard, and Guergana Savova. 2022. Ensemble-based fine-tuning strategy for temporal relation extraction from the clinical narrative. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 103–108, Seattle, WA. Association for Computational Linguistics.

Xin Xie, Zhoubo Li, Xiaohan Wang, Shumin Deng, Feiyu Xiong, Huajun Chen, and Ningyu Zhang. 2022. Promptkg: A prompt learning framework for knowledge graph representation learning and application. *CoRR, abs/2210.00305*.

Gokul Yenduri, M Ramalingam, G Chemmalar Selvi, Y Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G Deepti Raj, Rutvij H Jhaveri, B Prabadevi, Weizheng Wang, et al. 2022. Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv preprint arXiv:2305.10435*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Yifan Zhang. 2023. Meta prompting for agi systems. *arXiv preprint arXiv:2311.11482*.

A Appendix 1: prompt examples

A.1 2-shot prompt prefix

A.1.1 Introduction part

"'You are an AI model that checks biomedical records and corrects existing errors, based only on facts. Your goal is to read the medical record text and decide whether there are any errors. If yes, propose the corrected variant and indicate the error sentence number in the text. Correction is just the entire corrected sentence with NO additional explanations or words.

A.1.2 Few-shot examples

- Example 1: Text: "0 A 56-year-old man comes to the physician for a follow-up examination. 1 One month ago, he was diagnosed with a focal seizure, and treatment with a drug that blocks voltage-gated sodium channels was begun. 2 Today, he reports that he has not had any abnormal body movements, but he has noticed occasional double vision. 3 His serum sodium is 132 mEq/L, alanine aminotransferase is 49 U/L, and aspartate aminotransferase is 46 U/L. Medications include carbamazepine." Error exist: |||No||| Correction: |||None||| Error sentence number: |||None|||

- Example 2: Text: "0 A 53-year-old man comes to the physician because of a 1-day history of fever and chills, severe malaise, and cough with yellow-green sputum. 1 He

works as a commercial fisherman on Lake Superior. 2 Current medications include metoprolol and warfarin. 3 His temperature is 38.5 C (101.3 F), pulse is 96/min, respirations are 26/min, and blood pressure is 98/62 mm 4 Hg. 5 Examination shows increased fremitus and bronchial breath sounds over the right middle lung field. 6 After reviewing imaging, the causal pathogen was determined to be Haemophilus influenzae. 7 An x-ray of the chest showed consolidation of the right upper lobe." Error exists: |||Yes||| Correction: |||After reviewing imaging, the causal pathogen was determined to be Streptococcus pneumoniae.||| Error sentence number: |||6|||

A.1.3 Rules for the model

Output format if an error exists: Error exist: |||Yes||| Correction: |||<Correction text>||| Error sentence number: |||<Sentence number>|||

Output format if no error is present: Error exist: |||No||| Correction: |||None||| Error sentence number: |||None|||

Please make sure you complete the objective above with the following rules:

- 1. You should focus on errors in named entities like diagnoses, therapies, and biological species names.
- 2. You must not make things up, you should use only your medical knowledge and medical record data.
- 3. Remember, that you will be rewarded for correct corrections, but also fined for the wrong reports.
- 4. You shouldn't check and correct any spelling errors because only semantical errors are important to you.
- 5. For your convenience, you will see the list of named entities from the record and some info about them.
- 6. You will also see the enumerated sentences from the text - if an error is found, please provide the problematic sentence number.
- 7. Please provide NO explanation for your answer, just give me the error status and error corrections, if any, according to the Output format.

"'

A.2 5-shot prompt prefix

A.2.1 Introduction part

"'You are an AI model that checks biomedical records and corrects existing errors, based only on facts. Your goal is to read the medical record text and decide whether there are any errors. If yes, you should propose the corrected variant and indicate the error sentence number in the text. Correction is just the entire corrected sentence with NO additional explanations or words.

A.2.2 Few-shot examples

- Example 1: Text: "0 A 56-year-old man comes to the physician for a follow-up examination. 1 One month ago, he was diagnosed with a focal seizure, and treatment with a drug that blocks voltage-gated sodium channels was begun. 2 Today, he reports that he has not had any abnormal body movements, but he has noticed occasional double vision. 3 His serum sodium is 132 mEq/L, alanine aminotransferase is 49 U/L, and aspartate aminotransferase is 46 U/L. Medications include carbamazepine." Error exist: |||No||| Correction: |||None||| Error sentence number: |||None|||
- Example 2: Text: "0 A 53-year-old man comes to the physician because of a 1-day history of fever and chills, severe malaise, and cough with yellow-green sputum. 1 He works as a commercial fisherman on Lake Superior. 2 Current medications include metoprolol and warfarin. 3 His temperature is 38.5 C (101.3 F), pulse is 96/min, respirations are 26/min, and blood pressure is 98/62 mm 4 Hg. 5 Examination shows increased fremitus and bronchial breath sounds over the right middle lung field. 6 After reviewing imaging, the causal pathogen was determined to be Haemophilus influenzae. 7 An x-ray of the chest showed consolidation of the right upper lobe." An error exists: |||Yes||| Correction: |||After reviewing imaging, the causal pathogen was determined to be Streptococcus pneumoniae.||| Error sentence number: |||6|||
- Example 3: Text: "1 He complains of anxiety, nausea, abdominal cramping, vomiting, and diarrhea for three days. 2 He denies smoking, drinking alcohol, and using illicit drugs. 3 He appears restless. 4 His temperature is 37 C (98.6 F), pulse is 110/min, and 5 blood

pressure is 150/86 mm 6 Hg. 7 Physical examination shows dilated pupils, diaphoresis, and piloerection. 8 His abdominal exam shows diffuse mild tenderness. 9 There is no rebound tenderness or guarding. 10 Suspected overdose, recommend Naloxone administration. 11 His hemoglobin concentration is 14.5 g/dL 12 , leukocyte count is 8,000/mm, and platelet count is 250,000/mm3; serum studies and urinalysis show no abnormalities." An error exists: |||Yes||| Correction: |||Suspected overdose, recommend methadone administration.||| Error sentence number: |||10|||

- Example 4: Text: "0 A potassium hydroxide preparation is conducted on a skin scraping of the hypopigmented area. 1 Patient was treated with topical selenium sulfide based on the microscopy findings. 2 Microscopy of the preparation showed long hyphae among clusters of yeast cells." Error exist: |||No||| Correction: |||None||| Error sentence number: |||None|||

- Example 5: Text: "0 A 56-year-old man comes to the physician for a follow-up examination. 1 One month ago, he was diagnosed with a focal seizure, and treatment with a drug that blocks voltage-gated sodium channels was begun. 2 Medications include phenytoin. 3 Today, he reports that he has not had any abnormal body movements, but he has noticed occasional double vision. 4 His serum sodium is 132 mEq/L, alanine aminotransferase is 49 U/L, and aspartate aminotransferase is 46 U/L." Error exists: |||Yes||| Correction: |||Medications include carbamazepine.||| Error sentence number: |||2|||

A.2.3 Rules for the model

Output format if an error exists: Error exists: |||Yes||| Correction: |||<Correction text>||| Error sentence number: |||<Sentence number>|||

Output format if no error is present: Error exists: |||No||| Correction: |||None||| Error sentence number: |||None|||

Please make sure you complete the objective above with the following rules:

- 1. You should focus on errors in named entities like diagnoses, therapies, and biological species names.

- 2. You must not make things up, you should use only your medical knowledge and medical record data.

- 3. Remember, that you will be rewarded for correct corrections, but also fined for the wrong reports.

- 4. You shouldn't check and correct any spelling errors because only semantical errors are important to you.

- 5. For your convenience, you will see the list of named entities from the record and some info about them.

- 6. You will also see the enumerated sentences from the text - if an error is found, please provide the problematic sentence number.

- 7. Please provide NO explanation for your answer, just give me the error status and error corrections, if any, according to the Output format.

"'

A.3 Ensemble prompt prefix

A.3.1 Introduction part

"'You are an AI model that checks biomedical records and corrects existing errors, based only on facts. Your goal is to read the medical record text and decide whether there are any errors. If yes, propose the corrected variant and indicate the error sentence number in the text. Correction is just the entire corrected sentence with NO additional explanations or words.

A.3.2 Few-shot examples

- Example 1: Text: "0 A 56-year-old man comes to the physician for a follow-up examination. 1 One month ago, he was diagnosed with a focal seizure, and treatment with a drug that blocks voltage-gated sodium channels was begun. 2 Today, he reports that he has not had any abnormal body movements, but he has noticed occasional double vision. 3 His serum sodium is 132 mEq/L, alanine aminotransferase is 49 U/L, and aspartate aminotransferase is 46 U/L. Medications include carbamazepine." Error exist: |||No||| Correction: |||None||| Error sentence number: |||None|||

- Example 2: Text: “0 A 53-year-old man comes to the physician because of a 1-day history of fever and chills, severe malaise, and cough with yellow-green sputum. 1 He works as a commercial fisherman on Lake Superior. 2 Current medications include metoprolol and warfarin. 3 His temperature is 38.5 C (101.3 F), pulse is 96/min, respirations are 26/min, and blood pressure is 98/62 mm 4 Hg. 5 Examination shows increased fremitus and bronchial breath sounds over the right middle lung field. 6 After reviewing imaging, the causal pathogen was determined to be Haemophilus influenzae. 7 An x-ray of the chest showed consolidation of the right upper lobe.” Error exist: |||Yes||| Correction: |||After reviewing imaging, the causal pathogen was determined to be Streptococcus pneumoniae.||| Error sentence number: |||6|||

- Example 3: Text: “1 He complains of anxiety, nausea, abdominal cramping, vomiting, and diarrhea for three days. 2 He denies smoking, drinking alcohol, and using illicit drugs. 3 He appears restless. 4 His temperature is 37 C (98.6 F), pulse is 110/min, and 5 blood pressure is 150/86 mm 6 Hg. 7 Physical examination shows dilated pupils, diaphoresis, and piloerection. 8 His abdominal exam shows diffuse mild tenderness. 9 There is no rebound tenderness or guarding. 10 Suspected overdose, recommend Naloxone administration. 11 His hemoglobin concentration is 14.5 g/dL 12 , leukocyte count is 8,000/mm, and platelet count is 250,000/mm3; serum studies and urinalysis show no abnormalities.” Error exist: |||Yes||| Correction: |||Suspected overdose, recommend methadone administration.||| Error sentence number: |||10|||

- Example 4: Text: “0 A potassium hydroxide preparation is conducted on a skin scraping of the hypopigmented area. 1 Patient was treated with topical selenium sulfide based on the microscopy findings. 2 Microscopy of the preparation showed long hyphae among clusters of yeast cells.” Error exist: |||No||| Correction: |||None||| Error sentence number: |||None|||

- Example 5: Text: “0 A 56-year-old man comes to the physician for a follow-up examination. 1 One month ago, he was diagnosed with a focal seizure, and treatment with a drug that blocks voltage-gated sodium channels was begun. 2 Medications include phenytoin. 3 Today, he reports that he has not had any abnormal body movements, but he has noticed occasional double vision. 4 His serum sodium is 132 mEq/L, alanine aminotransferase is 49 U/L, and aspartate aminotransferase is 46 U/L.” Error exist: |||Yes||| Correction: |||Medications include carbamazepine.||| Error sentence number: |||2|||

A.3.3 Rules for the model

Output format if an error exists: Error exist: |||Yes||| Correction: |||<Correction text>||| Error sentence number: |||<Sentence number>|||

Output format if no error is present: Error exist: |||No||| Correction: |||None||| Error sentence number: |||None|||

Please make sure you complete the objective above with the following rules:

- 1. You should focus on errors in named entities like diagnoses, therapies, and biological species names.

- 2. You must not make things up, you should use only your medical knowledge and medical record data.

- 3. Remember, that you will be rewarded for correct corrections, but also fined for the wrong reports.

- 4. You shouldn’t check and correct any spelling errors because only semantical errors are important to you.

- 5. For your convenience, you will see the list of named entities from the record and some info about them.

- 6. You will see the enumerated sentences from the text - if an error is found, please provide also the problematic sentence number.

- 7. You will also see some possible solutions for this text from the other experts, along with the mean expert trust score for each opinion. You could take expert decisions into account, but with respect to the trust score (higher is better).

- 8. Please provide NO explanation for your answer, just give me the error status and error

corrections, if any, according to the Output format.

“”

B Appendix 2: Resources used

During the discussed approaches evaluation and prediction making, more than 5,600 API requests were made with 10,537,000 tokens transferred, and the total prediction cost was around $93,6.