IryōNLP[*] at MEDIQA-CORR 2024: Tackling the Medical Error Detection & Correction Task On the Shoulders of Medical Agents

Jean-Philippe Corbeil
Microsoft Health & Life Sciences
jcorbeil@microsoft.com

Abstract

In natural language processing applied to the clinical domain, utilizing large language models has emerged as a promising avenue for error detection and correction on clinical notes, a knowledge-intensive task for which annotated data is scarce. This paper presents MedReAct'N'MedReFlex, which leverages a suite of four LLM-based medical agents. The MedReAct agent initiates the process by observing, analyzing, and taking action, generating trajectories to guide the search to target a potential error in the clinical notes. Subsequently, the MedEval agent employs five evaluators to assess the targeted error and the proposed correction. In cases where MedReAct's actions prove insufficient, the MedReFlex agent intervenes, engaging in reflective analysis and proposing alternative strategies. Finally, the MedFinalParser agent formats the final output, preserving the original style while ensuring the integrity of the error correction process. One core component of our method is our RAG pipeline based on our ClinicalCorp corpora. Among other well-known sources containing clinical guidelines and information, we preprocess and release the open-source MedWiki dataset for clinical RAG application. Our results demonstrate the central role of our RAG approach with ClinicalCorp leveraged through the MedReAct'N'MedReFlex framework. It achieved the ninth rank on the MEDIQA-CORR 2024 final leaderboard.

1 Introduction

In natural language processing applied to the clinical domain, the accurate detection and correction of medical errors are paramount tasks with profound implications for patient care and safety. This paper introduces the multi-agent framework MedReAct'N'MedReFlex, meticulously engineered to tackle medical error detection and correction, as delineated in the MEDIQA-CORR 2024 competition.

Our framework integrates four distinct types of medical agents: MedReAct, MedReFlex, MedEval, and MedFinalParser, each playing a specialized role in the error identification and rectification process. Drawing inspiration from existing frameworks like ReAct (Yao et al., 2023) and Reflexion (Shinn et al., 2023), our framework orchestrates a structured approach to error handling.

Leveraging a Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2020) based on MedRAG (Xiong et al., 2024) and MedCPT (Jin et al., 2023), our approach operates over ClinicalCorp, an extensive corpora curated to encompass crucial clinical guidelines. Additionally, we introduce *MedWiki*, a collection of medical articles from Wikipedia. By integrating these resources, our approach seeks to advance state-of-the-art clinical NLP by offering a comprehensive solution tailored to the intricate nuances of medical error handling. Furthermore, this paper documents the construction and release of *MedWiki*, a substantial repository comprising over 1.3 million article chunks. Additionally, we detail the assembly of the ClinicalCorp, a comprehensive corpus comprising *MedWiki* along with other clinical guideline datasets, such as parts of the MedCorp corpora (Xiong et al., 2024) and parts of the guidelines (Chen et al., 2023).

Our main contributions are:

- We designed a multi-agent framework named *MedReAct'N'MedReFlex* to solve the medical error detection & correction task (MEDIQA-CORR 2024) based on four types of medical agents: *MedReAct*, *MedReFlex*, *MedEval* and *MedFinalParser*. We deployed this framework on ClinicalCorp using a retrieval-augmented generation approach.

[*]The team name *iryō* comes from the japanese for medical or healthcare.

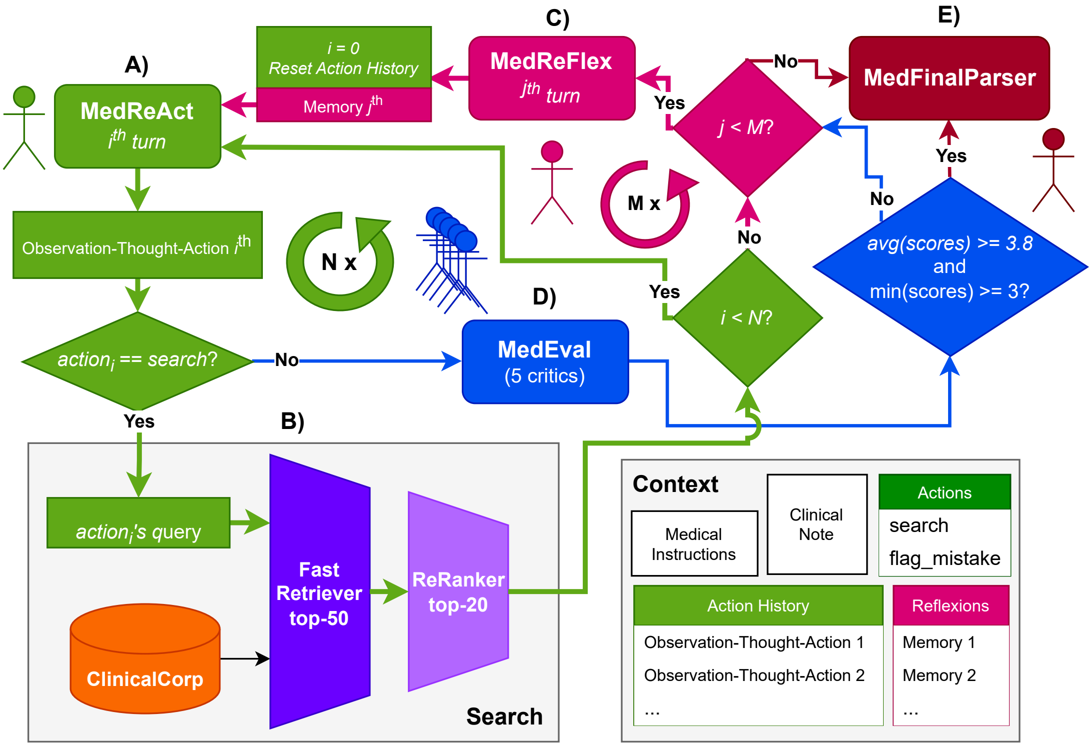

Figure 1: Schema of *MedReAct'N'MedReFlex* along the context of the clinical error correction task accessible to all medical agents: *MedReAct*, *MedReFlex*, *MedEval* and *MedFinalParser*. A) The *MedReAct* agent first provides an observation, a thought and an action. B) In the case of a *search* action, it triggers a semantic search over ClinicalCorp using *MedReAct*'s query. Then, the MedReAct agent loops up to N times (green inner loop) or until a *final_mistake* action is provided. C) After N unsuccessful searches from *MedReAct*, the *MedReFlex* agent reflects on the current situation and suggests a solution (pink outer loop). Then, *MedReAct* might start again. D) Once *MedReAct* selects the *final_mistake* action, the five *MedEval* agents review the answer and give a score between 1 and 5 (blue line). E) If the average equals or surpasses 3.8 and the minimum above or equal to 3, the *MedFinalParser* agent formats the final answer into a JSON object. If the answer is unsatisfactory, *MedReFlex* is triggered instead. If *MedReFlex* reaches unsuccessfully the M^{th} turns, *MedFinalParser* concludes that there is no error.

- We released the open-source *MedWiki*[1], a version of Wikipedia 2022-12-22 focused solely on medical articles. This RAG-ready dataset contains about 1.3M chunks from more than 150K articles, which represents about 3% of the original corpus.

- We provided the recipe to assemble our large corpora *ClinicalCorp* for RAG applications in the clinical domain, containing more than 2.3M chunks.

- We released a RAG-ready version[2] of the open-source *guidelines* used to pre-train *Meditron* (Chen et al., 2023), containing more than 710K chunks across eight open-source datasets.

- We released our codebase on GitHub[3].

2 Related Work

2.1 Medical Large Language Models

Since the emergence of ChatGPT by OpenAI in December 2022, the landscape of large language models (LLMs) has witnessed a proliferation of both private and public initiatives, leading to the development of increasingly sophisticated models. OpenAI's journey from the GPT3.5-turbo architecture, as reported by Brown et al. and Ouyang et al., culminated in the release of GPT-4 and its turbo variant (Achiam et al., 2023). Similarly, Google introduced Gemini, available in Nano, Pro, and Ultra configurations (Team et al., 2023), alongside its open-source Gemma model (Team et al., 2024). Anthropic contributed to this landscape with

[1] hf.co/datasets/jpcorb20/medical_wikipedia
[2] hf.co/datasets/jpcorb20/rag_epfl_guidelines
[3] github.com/microsoft/iryonlp-mediqa-corr-2024

Claude3, offered in three sizes, ranging from Haiku to Opus. Other notable LLMs include Mistral and Mixtral (Jiang et al., 2023, 2024), as well as Llama 2 (Touvron et al., 2023) and Yi (Young et al., 2024). These general-purpose LLMs, such as GPT-4, have demonstrated solid in-context learning capabilities in the medical field Nori et al. (2023).

Researchers have developed various open-source LLMs with diverse capabilities in the medical NLP domain. Examples include ClinicalCamel (Toma et al., 2023), Med42 (Christophe et al., 2023), PMC-Llama (Wu et al., 2023a), BioMedGPT (Zhang et al., 2023), Meditron (Chen et al., 2023), Apollo (Wang et al., 2024), OpenMedLM (Garikipati et al., 2024), and BioMistral (Labrak et al., 2024). Google also contributed Med-PaLM 2, a specialized LLM tailored for medical tasks (Singhal et al., 2023).

In this study, we employed OpenAI's GPT-4, specifically version turbo 0125, due to its proven state-of-the-art performances in various domains, its functional capabilities, and its large context window of 128K tokens. These attributes make it an ideal foundation for our approach. For instance, Nori et al. (2023) demonstrated that utilizing in-context learning with GPT-4 — relying on prompt engineering (i.e. few-shot learning (Brown et al.), chain-of-thought (Wei et al., 2022; Kojima et al., 2022), self-consistency (Wang et al., 2022) and shuffling multiple choice (Ko et al., 2020)) — achieves state-of-the-art performances on medical question-answering tasks, surpassing specialized models like Med-PaLM 2. We relied on a similar approach as our early baseline for medical error detection and correction, discarding the self-consistency and the shuffling techniques since both do not apply to generative tasks. Nonetheless, we have observed low results from which we hypothesized that this approach using only parametric knowledge is lacking reliable knowledge (Mallen et al., 2023; Ovadia et al., 2023; Kandpal et al., 2023), which we addressed by applying agentic methods in a retrieval-augmented generation framework.

2.2 Agentic Methods

Researchers have devised several agentic methods to enhance LLMs' responses and reasoning capabilities, such as ReAct (Yao et al., 2023), Reflexion (Shinn et al., 2023), DSPy (Khattab et al., 2023) and self-discovery (Zhou et al., 2024). Additionally, multi-agent paradigms (Wu et al., 2023b) have found application in the medical domain (Tang et al., 2023). Our approach draws inspiration from the Reflexion framework (Shinn et al., 2023), which we adapted into our *MedReFlex* agent. Specifically, we implemented a *MedReAct* agent — inspired by the ReAct approach (Yao et al., 2023) — to generate trajectories in our environment. However, this agent realizes its sequence of actions in a different order (i.e., observation, thought, and action), enabling streamlined execution.

Given the reliance of the Reflexion framework on feedback mechanisms, we incorporated an LLM-based metric into our *MedEval* medical agents. Evaluation metrics based on prompting strong LLMs, such as GPT-4 (Liu et al., 2023), have demonstrated a high correlation with human judgment. Similar findings have been reported in the medical NLP literature (Xie et al., 2023). Our evaluation protocol involves prompting five GPT-4 reviewers with task-specific criteria: validity, preciseness, confidence, relevance, and completeness. The average and minimum of their scores are both utilized as success criteria, capturing an unbiased final score and the evaluators' confidence, respectively.

2.3 Retrieval-Augmented Generation

Before the advent of LLMs, authors have proposed the retrieval-augmented generation (RAG) framework as a mechanism to incorporate non-parametric memory for knowledge-intensive tasks. This framework, as elucidated by Lewis et al. (Lewis et al., 2020), leverages both sparse (Robertson et al., 2009) and dense (Reimers and Gurevych, 2019) retrieval methods. In the medical NLP domain, MedCPT (Jin et al., 2023) serves as a prominent retrieval approach, augmented by a reranking stage based on a cross-encoder model. Notably, Xiong et al. (Xiong et al., 2024) conducted a comprehensive study on RAG applications in the medical domain, culminating in developing the MedRAG framework and the MedCorp corpora. Our approach builds upon these foundations, employing the MedCPT retrieval techniques and two corpora from MedCorp.

A pivotal aspect of RAG is its search engine's collection of indexed documents. The *guidelines* corpora, part of the *GAP-replay* corpora, was curated to train Meditron (Chen et al., 2023). This corpus comprises web pages describing medical guidelines from reputable healthcare websites like the World Health Organization. The *StatPearls* and

Textbooks datasets, included in the *MedCorp* corpora used in MedRAG (Xiong et al., 2024), encompass documents from clinical decision support tools and medical textbooks (Jin et al., 2021). While *Wikipedia* and *PubMed* datasets within *MedCorp* offer extensive data (i.e. more than 55M documents), we opted for efficiency by focusing on the smaller *PubMed* subset in the *guidelines* corpora and our *MedWiki* corpus.

3 Methodology

3.1 MEDIQA-CORR Task

The goal of the medical error detection and correction task (Ben Abacha et al., 2024a) from the clinical note is threefold: detect the presence of an error, locate the sentence containing the error and generate a corrected version of that sentence. The input of the dataset (Ben Abacha et al., 2024b) is a clinical note of several sentences containing a medical description of a patient's condition, test results, diagnosis, treatment and other aspects. There are two parts for the validation and test sets: *MS* from Microsoft and *UW* from the University of Washington. As a primary evaluation metric, the organizers asked to utilize the aggregation score defined by Abacha et al. (2023) over Rouge-1, BertScore and BLEURT, demonstrating a higher correlation with human judgement.

3.2 ClinicalCorp Corpora

Our corpus is detailed in Table 1.

guidelines We aggregated 13 datasets — which are open-source or closed-source — from the *guidelines* corpora. We adapted and ran the scrappers from the Meditron GitHub repository to gather the closed-source datasets. Then, we chunked the resulting documents using LangChain's recursive-character text splitter (Chase, 2022) with a chunk size of 1,000 characters and an overlap of 200 characters, as used for *StatPearls* (see next section).

MedCorp We gathered two of the four datasets contained in *MedCorp* from MedRAG (Xiong et al., 2024): *StatPearls* and *Textbooks*. The former was downloaded, cleaned and chunked using *MedRAG* GitHub repository, while the latter was readily available on the *HuggingFaceHub*[4].

MedWiki We filtered the 2022-12-22 Wikipedia dump[5] pre-processed into chunks by *Cohere* for medical articles only. To select the medical articles, we leveraged an available fine-tuned BerTopic[6] (Grootendorst, 2022), trained on the same Wikipedia dump. We associated its 2,3K topics to the medical domain based on the topics' word representations — e.g. topic *1850* is related to the medical field, and it corresponds to the word representations: shingles, herpesvirus, chickenpox, herpes, smallpox, zoster, immunity, infectivity, inflammation, and viral. We made these predictions by prompting *GPT3.5-turbo 0613* with a temperature of 1.0 followed by a majority vote over five predictions. If at least four were positive, we declared the topic medically relevant. In the manual verification of about 50 diverse medical terms on the resulting collection, we observed a near-perfect coverage of Wikipedia's articles related to diseases, treatments, bacteria, or drugs. Only two topics were missing[7], corresponding to one single example from the manual test. Given that our goal is to reduce the size of this dataset and use it in an RAG application, we added these topics manually. We obtained a corpus of 150K articles and nearly 1.4M chunks.

3.3 Semantic Search

We followed the MedCPT approach (Jin et al., 2023) in two stages (see step *B* in Figure 1), which is composed of a fast bi-encoder retrieving stage followed by a cross-encoder reranking stage.

We implemented the first stage on a ChromaDB instance, in which we loaded *ClinicalCorp*. This stage aims to find relevant documents while maintaining a good accuracy/latency trade-off. This vector database embeds documents using a fast bi-encoder model (Reimers and Gurevych, 2019). Then, we provide a query to fetch the closest documents under a given distance, computed with the hierarchical navigable small world approximation (HNSW, by Malkov and Yashunin (2018)). We experimented with three bi-encoders from the *HuggingFaceHub*: *sentence-transformers/all-MiniLM-L6-v2* (default), *NeuML/pubmedbert-base-embeddings-matryoshka* and MedCPT original Query/Article encoders. According to our initial experiments, we discarded *all-MiniLM-L6-v2* because we noticed a critical lack of knowledge about medical terminology hindering its accuracy despite a very low latency. NeuML's model and MedCPT's

[4] `hf.co/datasets/MedRAG/textbooks`

[5] `hf.co/datasets/Cohere/wikipedia-22-12`

[6] `hf.co/MaartenGr/BERTopic_Wikipedia`

[7] Index *509* related to biological taxonomy and *806* related to yeasts.

Table 1: Datasets gathered to construct ClinicalCorp.

Dataset	Source	Status	# Documents	# Chunks
Guidelines (Chen et al., 2023)	WikiDoc	open	33,058	360,070
	PubMed (guidelines only)	open	1,627	124,971
	National Institute for Health and Care Excellence	open	1,656	87,904
	Center for Disease Control and Prevention	open	621	70,968
	World Health Organization	open	223	33,917
	Canadian Medical Association	open	431	18,757
	Strategy for Patient-Oriented Research	open	217	11,955
	Cancer Care Ontario	open	87	2,203
	Drugs.com	close	6,711	37,255
	GuidelineCentral	close	1,285	2,451
	American Academy of Family Physicians	close	60	130
	Infectious Diseases Society of America	close	54	7,785
	Canadian Paediatric Society	close	43	1,123
MedCorp (Xiong et al., 2024)	StatPearls	close	9,379	307,187
	Textbooks (Jin et al., 2021)	open	18	125,847
ClinicalCorp (Ours)	MedWiki	open	150,380	1,139,464
	All	**mix**	**205,850**	**2,331,987**

are Bert-based models of 768 hidden dimensions and 12 layers, a slow architecture to generate sentence embeddings. However, NeuML fine-tuned a recent model using the Matryoshka Representation Learning technique (Kusupati et al., 2022), allowing to truncate dimensions down to 256 dimensions of the 768 embeddings, which significantly accelerated the computations. Our experiments employ this MRL encoder with truncation at 256 dimensions as a trade-off between accuracy and latency.

We implemented the reranking stage following the cross-encoder approach from MedCPT (Jin et al., 2023). Our early experimentation demonstrated the superiority of this model compared to NeuML's MRL bi-encoder with all 768 dimensions as a reranker.

4 MedReAct'N'MedReFlex Framework

Unlike previous multi-agent frameworks (Wu et al., 2023b; Tang et al., 2023), our approach diverges from a free conversation format to adopt a fixed design schema, as illustrated in Figure 1. Within this structured framework, each medical agent intervenes at a specific step, facilitating a systematic and coordinated approach to address the error detection and correction task. Central to our methodology are four distinct medical agents: MedReAct, MedReFlex, MedEval, and MedFinalParser.

4.1 MedReAct Agent

The MedReAct agent (see step *A* in Figure 1), inspired by the ReAct framework (Yao et al., 2023), operates cyclically, beginning with an observation of the current context, followed by a thoughtful analysis, and concluding with an action (*search* or *final_mistake*). This agent generates a trajectory of up to N steps if the action is always a *search* with different queries.

We also experimented with adding two other actions (*get_doc_by_id* and *next_results_from_query*), but MedReAct systematically ignored them.

4.2 MedEval Agent

Upon MedReAct's selection of the *final_mistake* action, the MedEval agents (see step *D* in Figure 1), akin to the GPT-Eval approach (Liu et al., 2023), evaluate the proposed solution. Five GPT-4-based evaluators assess the answer based on criteria such as validity, preciseness, confidence, relevance, and completeness. The ensemble of evaluators ensures comprehensive and unbiased feedback, contributing to robust error detection and correction. We leverage the average final score as well as the minimum review score. We added this condition on the minimum score to capture the confidence of the evaluation. If one reviewer gave a much lower score than the others, we experimentally observed that it was often a signal of lower confidence in the final answer.

4.3 MedReFlex Agent

In scenarios where MedReAct's actions fail to yield satisfactory outcomes, the MedReFlex agent (see step C in Figure 1), drawing from the Reflexion framework (Shinn et al., 2023), intervenes. This agent engages in reflective analysis to reassess the situation. By considering contextual cues, past interactions and all five reviews, MedReFlex proposes alternative strategies to address the identified challenges. This iterative process allows for adaptive decision-making and fosters resilience in error detection and correction tasks.

4.4 MedFinalParser Agent

Suppose the average score provided by the MedEval agents exceeds or equals 4, and the minimum score surpasses or equals 3. In that case, the MedFinalParser agent (see step D in Figure 1) proceeds to format the final answer into a JSON object. This agent also ensures the conservation of the original style of the clinical note, which the MedReAct agent tends to disrupt by copying the writing style of the search documents. Conversely, if the answer falls short of the predetermined thresholds, MedReFlex is triggered for further refinement. If MedReFlex's interventions prove ineffective after the M^{th} turn, the MedFinalParser agent concludes that no errors exist, ensuring the integrity of the error correction process.

5 Results

5.1 Results for the Competition

MedReAct'N'MedReFlex achieved the 9^{th} rank during the MEDIQA-CORR 2024 official testing period, corresponding to an aggregation score of 0.581. Nonetheless, we thoroughly optimize our method in the following sections. To complete these experiments in a reasonable amount of time, we randomly sample 50 examples from the MS validation set.

5.2 Agentic Method Comparison

In Table 2, we compared the MedReAct agent only against using our proposed method MedReAct'N'MedReFlex. Our approach achieves more than a few absolute percent across metrics. We also experimented with a baseline inspired from Nori et al. (2023) (i.e. "MedPrompt") with in-context learning prompting alone, but the results were drastically lower.

Metric	**MedReAct**	**MedReAct'N'MedReFlex**
ROUGE-1	0.504	0.568
BERTScore	0.580	0.642
BLEURT	0.531	0.588
Aggregate	0.539	0.599

Table 2: Comparison between MedReAct agent only with up to 10 turns against MedReAct'N'MedReFlex with 4 turns for MedReAct and 5 turns for MedReFlex, leveraging the optimal search configuration (retrieval top-k at 50 and reranking top-k at 20).

5.3 Semantic Search Optimization

After the end of the MEDIQA-CORR 2024 shared task, we carried out a thorough analysis of our semantic search engine. The main parameters to tune are retrieval top-k, reranking top-k and the source included in ClinicalCorp.

5.3.1 Retrieval Top-K

In Figure 2, we illustrate the performances across many retrieval top-k values employing a fixed reranking top-k of 20. For the official ranking of the MEDIQA-CORR 2024, we set this value to 300. However, we observe here that this setting is sub-optimal. A retrieval top-k of 50 improves the final performances by a few absolute percent. We interpret this observation as indicative of a misalignment between our task and the fine-tuning of the MedCPT reranker. The more documents we provide to the reranking model (e.g. 200 or 300), the more low-relevance documents are put in the top 20 by the reranker output.

Nonetheless, a reranking without surplus documents — i.e. retrieval top-k of 20 with a reranking top-k of 20 — remains sub-optimal, mainly in contrast to using 50 documents. In Figure 3, we provide the associated average latency for one react step in seconds. We notice that the latency seems to scale with the order of magnitude of the retrieval top-k, with a value of 20 and 50 having 17 seconds on average, while 100, 200 and 300 are around 20 seconds. We expected that the reranking of 300 examples against 100, for instance, would lead to noticeable latency, but it is negligible in contrast to the retrieval from ChromaDB over our 2.3M chunks.

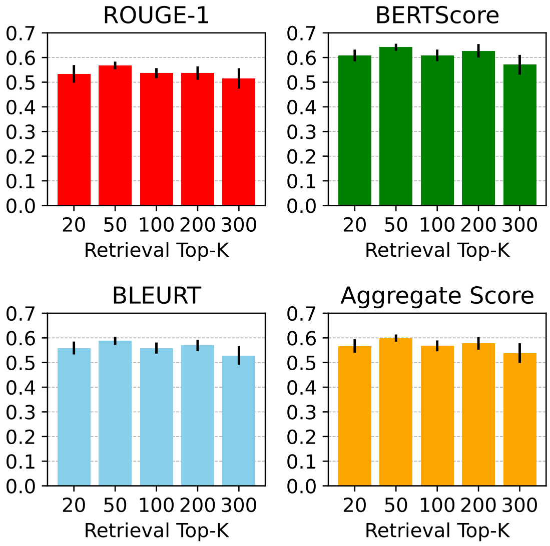

Figure 2: Performances across many retrieval top-k values with a reranking top-k set at 20 over 3 runs.

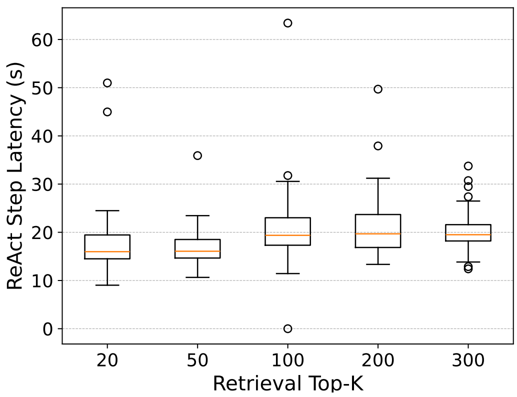

Figure 3: ReAct step average latency per retrieval top-k with a reranking top-k set at 20.

We also show the average amount of MedReFlex turns and the average of the sum of all MedReAct turns in Figure 4. Overall, the trends are similar, with 4.8 total ReAct turns on average, but there is a slight increase in the average and variance for top-k values of 200 and 300. Therefore, these settings are underperforming and slower regarding latency

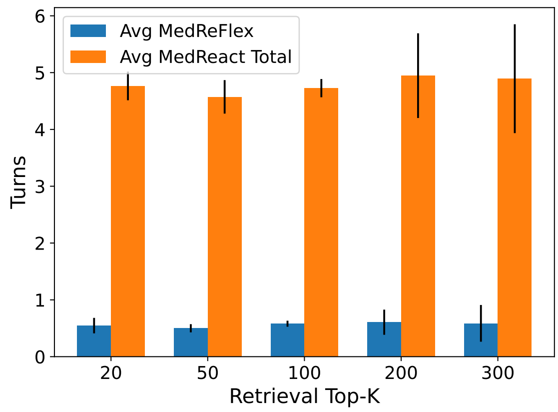

Figure 4: Average turns of MedReAct and MedReFlex according to various retrieval top-k with a reranking top-k set at 20.

and the number of turns needed to reach an answer.

Overall, the retrieval top-k of 50 leads to higher performances across all metrics and reduced latency and number of turns required by our algorithm.

5.3.2 Reranker Top-K

In Figure 5, we fix the retrieval top-k at 300 and compute the performances across three reranking top-k values: 5, 10 and 20. Since the context window of the LLM limits us, we constraint the maximum of K to 20, given that these K documents are injected in the prompt up to N times for each MedReAct step. According to Figure 5, we observe that the more documents we provide in the prompt, the more we increase the performances — the aggregate score gains close to 10% absolute when augmenting from 5 to 20 documents.

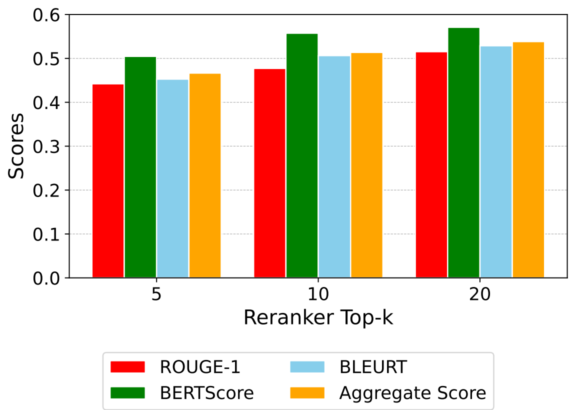

Figure 5: Reranker top-K with a retrieval top-k set at 300.

5.3.3 Sources in ClinicalCorp

We measure the impact of each source in ClinicalCorp in Figure 6. First, we observe that *MedWiki* is the lowest-performing source of documents with an aggregation score of nearly 0.47. *guidelines* and *Textbooks* provide a similar accuracy at about 0.51 in aggregate score. Finally, *StatPearls* leads to the highest score close to the full ClinicalCorp. Given our small validation set of 50 examples, we consider it a better practice to keep all ClinicalCorp for our task since more edge cases might appear at test time.

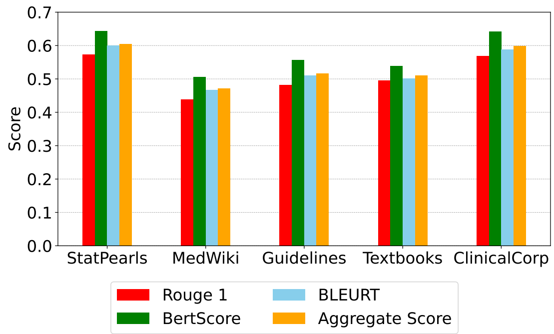

Figure 6: Performances per source in ClinicalCorp with the retrieval top-k set at 50 and the reranking top-k set at 20.

We show in Figure 7 the distributions of sources from ClinicalCorp in general in comparison to the distributions of sources' chunks used by one run of MedReAct'N'MedReFlex. We observe, in general, a much larger utilization of the *StatPearls*' chunks in contrast to the *MedWiki*'s chunks, while we remark similar distributions for the other two datasets. These results align with the previous analysis demonstrating a higher performance from using only *StatPearls*.

5.4 MedEval Aggregation Thresholds

In Figure 8, we show the impact of applying different thresholds to the average and minimum review scores on the performance. For the minimum score criterion, we choose the integer values of 2.0, 3.0 and 4.0. We select values for the average score criterion: 3.0, 3.2, 3.5, 3.8, 4.0 and 4.2. We do not compute the performances for combinations where the minimum threshold is higher than the average threshold for mathematical consistency. We observe an optimal setting for a minimum evaluation score of 3.0 with a range of average evaluation scores in $[3.5, 3.8]$.

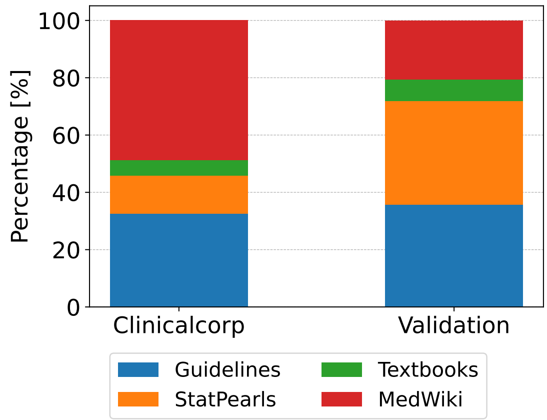

Figure 7: Distribution of sources' chunks in ClinicalCorp against appearances of these chunks' sources in one run of MedReAct'N'MedReFlex.

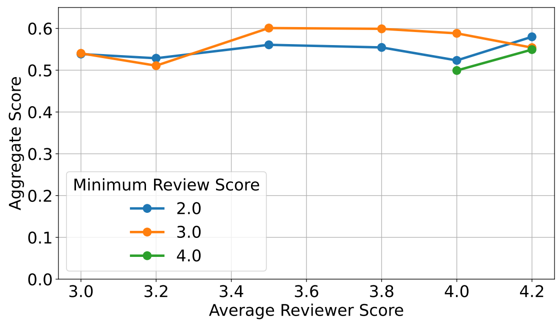

Figure 8: Aggregate scores across different MedEval's average and minimum thresholds with the retrieval top-k set at 50 and the re-ranker top-k set at 20. We omitted average thresholds that are strictly lower for consistency for a given minimum threshold.

6 Conclusion

In this paper, we introduced MedReAct'N'MedReFlex, a multi-agent framework developed for the MEDIQA-CORR 2024 competition aimed at medical error detection and correction in clinical notes. The framework incorporates four specialized medical agents: MedReAct, MedReFlex, MedEval, and MedFinalParser, leveraging the RAG framework and our ClinicalCorp. We detail the construction of our ClinicalCorp, including diverse clinical datasets such as *guidelines*, *Textbooks*, and *StatPearls*. Additionally, we released MedWiki, a corpus comprising Wikipedia medical articles. Our framework achieved the ninth rank in the competition with an aggregation score of 0.581. Through optimization experiments, we identified sub-optimal settings at the time, demonstrating substantial performance improvements with a retrieval top-k of 50, a reranking top-k of 20, an average review threshold of 3.8, and a minimum review threshold of 3. As future work, we envision refining the chunking strategy on the ClinicalCorp, applying further prompt engineering of the medical agents, and conducting a deeper analysis of the interactions between the MedReAct'N'MedReFlex's agents.

Acknowledgments

We would like to thank Jean-Michel Attendu, François Beaulieu, and Paul Vozila from Microsoft Health & Life Sciences team for their support, as well as the ClinicalNLP workshop's organizers, reviewers and other participants.

References

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513.

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

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

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

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners.

Harrison Chase. 2022. Langchain. https://github.com/langchain-ai/langchain. Version 1.2.0, Released on October 17, 2022.

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.

Anurag Garikipati, Jenish Maharjan, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Qingqing Mao, and Ritankar Das. 2024. Openmedlm: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, Heather Miller, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham M Kakade, Prateek Jain, et al. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Medicine*, 84(88.3):77–3.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.

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

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023b. Autogen: Enabling next-gen llm applications via multiagent conversation framework. *arXiv preprint arXiv:2308.08155*.

Yiqing Xie, Sheng Zhang, Hao Cheng, Zelalem Gero, Cliff Wong, Tristan Naumann, and Hoifung Poon. 2023. Enhancing medical text evaluation with gpt-4. *arXiv preprint arXiv:2311.09581*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. 2023. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed H Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. *arXiv preprint arXiv:2402.03620*.