Overview of the MEDIQA-CORR 2024 Shared Task on Medical Error Detection and Correction

Asma Ben Abacha, Wen-wai Yim

Microsoft Health AI {abenabacha,yimwenwai}@microsoft.com

Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen

University of Washington {velvinfu,zhaoyis,fxia,melihay}@uw.edu

Abstract

Automatic detection and correction of medical errors enables a more rigorous validation of medical documentation as well as clinical notes generated by large language models. Such solutions can ensure the accuracy and medical coherence of clinical texts and enhance patient care and health outcomes. The MEDIQA-CORR 2024 shared task focused on detecting and correcting different types of medical errors in clinical texts. Seventeen teams participated in the shared task and experimented with a broad range of approaches and models. In this paper, we describe the MEDIQA-CORR task, datasets, and the participants' results and methods.

1 Introduction

A recent survey study from three US health care organizations showed that 1 in 5 patients who read a clinical note reported finding a mistake and 40% perceived the mistake as serious. Among the very serious errors reported by patients, the most common category of mistakes was related to current or past diagnoses. Other very serious patient-reported mistakes included inaccurate description of medical history, medications or allergies, physical examination, test results, notes on the wrong patient, and sidedness (left vs right) (Bell et al., 2020).

Giardina et al. (2018) focused on diagnostic errors and analyzed patient- and family-reported error narratives to explore factors that contribute to diagnostic errors. Problems related to patient-physician interactions emerged as major contributors.

The probability of such errors is expected to increase in medical documents and clinical notes generated by Large Language models (LLMs) to assist healthcare professionals in their daily documentation tasks.

On a general, coarse-grained, level, LLMs have shown the ability to imitate clinical reasoning while forming mostly accurate diagnoses (Savage et al., 2024). However, one of the main challenges in integrating LLMs in medical workloads is their potential to generate misleading or incorrect information (Tang et al., 2023). Rigorous validation processes are essential to mitigate these risks and make LLMs safe(r) to use for medical content generation (Karabacak and Margetis, 2023).

One important aspect of this validation is medical common-sense checking to validate the coherence and soundness of the generated medical reasoning. However, most previous studies on common sense detection have focused on the general domain (Wang et al., 2020; Once et al., 2021).

In this task, we tackle the problem of identifying and correcting (common sense) medical errors in clinical notes. From a human perspective, identifying and correcting these errors requires medical expertise, specialized knowledge, and sometimes practical experience. To the best of our knowledge, this task is the first to address the automatic validation and correction of clinical notes.

2 Task Description

The MEDIQA-CORR 2024 shared task¹ addresses the problem of identifying and correcting (common sense) medical errors in clinical notes. From a human perspective, identifying and correcting these errors require medical expertise and knowledge.

In the task data, each clinical text is either correct or contains one error. The task consists of three subtasks:

- **A:** Predicting the error flag (1: the text contains an error, 0: the text has no errors)
- **B:** Extracting the sentence that contains the error for flagged texts (-1: the text contains no error; Sentence ID: if the text contains an error)

¹https://sites.google.com/view/mediqa2024/
mediqa-corr

	DIAGNOSIS	CAUSAL ORGANISM	MANAGEMENT	TREATMENT	PHARMACOTHERAPY
ERROR	A 17-year-old boy is brought to the physician by his mother because of increasingly withdrawn behavior for the last two years. His mother reports that in the last 2-3 years of high school, her son has spent most of his time in his room playing video games. He does not have any friends and has never had a girifriend. He usually refuses to attend family dinner and avoids contact with his siblings. The patient states that he prefers being on his own. When asked how much playing video games means to him, he replies that "it's okay." When his mother starts crying during the visit, he appears indifferent. Physical and neurologic examinations show no other abnormalities. Suspected of autism spectrum disorder. On mental status examination, his thought process is organized and logical. His affect is flattened.	A 64-year-old man is brought to the emergency department because of fever, chills, shortness of breath, chest pain, and a productive cough with bloody sputum for the past several days. He has metastatic pancreatic cancer and is currently undergoing polychemotherapy. His temperature is 38.3 C (101 F). Pulmonary examination shows scattered inspiratory crackles in all lung fields. A CT scan of the chest shows multiple nodules, cavities, and patchy areas of consolidation. Histoplasma capsulatum was determined as the causal pathogen. A photomicrograph of a specimen obtained on pulmonary biopsy is shown.	A 42-year-old woman comes to the physician because of a low-grade fever and generalized fatigue for a week. During this period, she has passed decreased amounts of urine. Two months ago, she underwent a renal allograft transplant because of reflux nephropathy. There is no family history of serious illness () Oral fluconazole is administered. Patient was recommended intravenous immunoglobulin therapy as a next step in management.	A 47-year-old woman comes to the physician because of easy bruising and fatigue. She appears pale. Her temperature is 38 C (100.4 F). Examination shows a palm-sized hematoma on her left leg. Abdominal examination shows an enlarged liver and spleen. Based on the following findings, patient was treated with platelet transfusion. Her hemoglobin concentration was 9.5 g/dL, leukocyte count was 12,300/mm3, platelet count was 55,000/mm3, and fibrinogen concentration was 120 mg/dL. Cytogenetic analysis of leukocytes showed a reciprocal translocation of chromosomes 15 and 17.	A 67-year-old man with type 2 diabetes mellitus and benign prostatic hyperplasia comes to the physician because of a 2-day history of sneezing and clear nasal discharge. He has had similar symptoms occasionally in the past. His current medications include metformin and tamsulosin. Examination of the nasal cavity shows red, swollen turbinates. The patient is given diphenhydramine.
	Suspected of schizoid personality disorder.	Aspergillus fumigatus was determined as the causal pathogen.	Patient was recommended methylprednisolone therapy as a next step in management.	Based on the following findings, patient was treated with all-trans retinoic acid.	The patient is given desloratadine.

Figure 1: Examples from the MEDIQA-CORR MS training set.

C: Generating a corrected sentence for flagged texts

3 Data Creation

We created a new dataset of 3,848 clinical texts with injected errors such as diagnosis, causal organism, management, treatment, and pharmacotherapy (Ben Abacha et al., 2024). The dataset includes two types of texts: clinical texts from publicly available data (MS collection) and de-identified clinical notes from the University of Washington Medical Center (UW collection). The UW dataset was built using new de-identified notes and requires signing a data usage agreement. The MS dataset was built by transforming the MedQA medical questionanswering dataset (Jin et al., 2020) with manual error injections and text modifications that leveraged the clinical notes and the multiple-choice questions.

The MS training set contains 2,189 clinical texts. Figure 1 presents examples from the MS training data. The MS validation set contains 574 clinical texts and the UW validation set contains 160 clinical texts. The final test set consists of 597 clinical texts from the MS collection and 328 clinical texts from the UW dataset.

4 Evaluation

4.1 Evaluation Metrics

We rely on Accuracy for Error Flag Prediction (subtask A) and Error Sentence Detection (subtask B).

For the evaluation of Sentence Correction (subtask C), we selected three automatic metrics that highly correlate with human judgments on clinical texts based on recent studies (Ben Abacha et al., 2023a,b). These metrics are: ROUGE-1 (Lin, 2004), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2020).

Similar to MEDIQA-Chat (Ben Abacha et al., 2023) and MEDIQA-SUM 2023 (Yim et al., 2023), we used the aggregate (average) score from ROUGE-1, BLEURT-20, and BERTScore (microsoft/deberta-xlarge-mnli) as the main score to rank the participating systems.

We also computed a Composite score as follows

	Team	Affiliation	Subtasks	Paper	Code
l	WangLab	University of Toronto, Canada	1, 2, 3	(Toma et al., 2024)	1 2
2	PromptMind	Google, USA	1, 2, 3	(Gundabathula and Kolar, 2024)	
3	HSE NLP	Higher School of Economics University, Russia	1, 2, 3	(Valiev and Tutubalina, 2024)	3
4	KU-DMIS	Korea University	1, 2, 3	(Hwang et al., 2024)	4
5	Maven	Pune Institute of Computer Technology, India	1, 2, 3	(Jadhav et al., 2024)	5
6	Edinburgh Clin- ical NLP	University of Edinburgh, Scotland	1, 2, 3	(Gema et al., 2024)	6
7	knowlab_AIMed	University College London & The University of Hong Kong	1, 2, 3	(Wu et al., 2024)	7
8	EM_Mixers	Emory University, USA	1, 2, 3	(Rajwal et al., 2024)	8
9	IryoNLP	Microsoft, Canada	1, 2, 3	(Corbeil, 2024)	
10	IKIM	Institute for AI in Medicine, Germany	1, 2, 3	-	10
11	CLD-MEC	Princess Sumaya University for Technol- ogy, Jordan	1, 2, 3	(Alzghoul et al., 2024)	11
12	romarcg	IDSIA, Switzerland	1, 2, 3	-	12
13	mekki	Um6p College Of Computing, Morocco	1, 2, 3	-	13
14	MediFact	National University of Computer and	1, 2, 3	(Saeed, 2024)	14
		Emerging Sciences, Pakistan			
15	harivm	University of California, Los Angeles (UCLA), USA	1, 2, 3	-	15
16	VerbaNexAI	Pontificia Universidad Javeriana, Colombia	1, 2	(Pajaro et al., 2024)	16
17	nlp-lab-iu	Indiana University Bloomington, USA	1,2	-	17
	2 https:/	<pre>1 https://github.com/bowan //github.com/satyakesav/medical 3 https://github.com/Rebell 4 https://github.com/HwangHye 5 https://github.com/HwangHye</pre>	-error-de -Leader/m eoni/MEDI(tection-and-correction ediga-corr DA-CORR-2024	
		5https://github.com/abhayshan	-		
		6 https://github.com/ 7 https://github.com/wuzl01/Kn			
	8 https://git	hub.com/swati-rajwal/EM_Mixers			4
		9 https://github.com/jpcork			
		10 https://github.com/dadaan	nin/MEDIQA	-CORR-2024	
		<pre>11 https://github.com/Rer</pre>	nadzghoul/	CLD-MEC	
		12 https://github.com/OWLmx/	/mediqa202	24_medicorr	
		13 https://github.com/4mekk	-		
	14	ttps://github.com/NadiaSaeed/M			
		15 https://github.co			
		16https://github.com/Da			
	1	7 https://github.com/dhananjay-	srivasta	/a/MEDIQA-CORR	

Table 1: MEDIQA-CORR 2024: Participating teams, subtasks, papers, and codes.

for each text: (i) 1 point if both the system correction and the reference correction are "NA": i.e., both the reference and system agree that the text has no errors, (ii) 0 points if only one of the system or the reference is "NA" (i.e., disagreement on error presence), and (iii) Aggregate-Score if both the system and reference agree that the sentence has an error.

Our evaluation scripts are available online².

4.2 Code Verification

For additional validation, we required the submission of the code in addition to the models' outputs/runs. The participants shared their private codes with the organizers on GitHub following provided guidelines.

4.3 Baseline System

We built a GPT-4-based baseline system, with deterministic outputs (temperature=0), using the following prompt for the three subtasks:

• The following is a medical narrative about a patient. You are a skilled medical doctor reviewing the clinical text. The text is either correct or contains one error. The text has a sentence per line. Each line starts with the sentence ID, followed by a pipe character then the sentence to check. Check every sentence of the text. If the text is correct return the following output: CORRECT. If the text has

 $^{^2 \}rm https://github.com/abachaa/MEDIQA-CORR-2024/tree/main/evaluation$

Team	Error Flag Accuracy	Error Sentence Detection Accuracy
WangLab *	0.8649	0.8357
PromptMind	0.6216	0.6086
HSE NLP	0.5222	0.5200
KU-DMIS	0.6346	0.6151
Maven *	0.5600	0.5200
Edinburgh Clinical NLP	0.6692	0.6108
knowlab_AIMed	0.6941	0.6195
EM_Mixers	0.6800	0.6400
IryoNLP	0.6714	0.6097
IKIM	0.6778	0.5903
CLD-MEC	0.5665	0.4908
romarcg	0.5016	0.3784
mekki	0.5395	0.3632
MediFact	0.7373	0.6000
harivm	0.5027	0.1924
nlp-lab-iu	0.5124	0.0497
VerbaNexAI	0.5103	0.4865
Baseline (GPT-4)	0.6562	0.5503

Table 2: Official Results of Error Flag Prediction (Subtask A) and Error Sentence Detection (Subtask B). * Potential use of MS test data.

a medical error, return the sentence id of the sentence containing the error, followed by a space, and a corrected version of the sentence.

5 Official Results

5.1 Participating Teams

The MEDIQA-CORR 2024 shared task attracted 112 registered teams from academy and industry. Among them, seventeen teams submitted their codes and runs following the challenge rules. Table 1 presents the teams that participated in the three subtasks. We limited the number of submitted runs to 20 runs per team.

5.2 Results & Approaches

The main results of the challenge are presented in Table 2 and Table 3.

The WangLab team (Toma et al., 2024) achieved the best Accuracy of 0.8649 in Error Flag Prediction (subtask A) and 0.8357 in Error Sentence Detection (subtask B). They also achieved the best Aggregate-Score of 0.7891 and Aggregate-Composite of 0.7746 in Sentence Correction (subtask C). The WangLab team used two different methods for the MS and UW datasets. They leveraged the MedQA medical question-answering dataset (Jin et al., 2020) to isolate questions resembling those in the MS data. This likely led to test data leakage as the MedQA dataset was used to build the MS subset. They employed DSPy (Khattab et al., 2023), a framework for automating the optimization of LLM programs, to refine a series of modules aimed at detecting and correcting errors. They also implemented a distinct set of DSPy modules to develop LLM-based programs for error identification and correction in the UW dataset.

The PromptMind team (Gundabathula and Kolar, 2024) achieved the second best aggregate score of 0.7866 in error sentence correction with 0.6216 error flag accuracy and 0.6086 error sentence detection accuracy using a prompt-based in-context learning strategy. They combined the results of GPT-4 and Claude-3 Opus models to generate the error flag, error sentence ID, and corrected sentence.

The third best aggregate score was obtained by the HSE NLP team (Valiev and Tutubalina, 2024) with an in-prompt ensemble approach with named entity recognition and knowledge graph for medical error checking. Their approach consists of three key components: entity extraction, prompt engineering, and ensemble. First, they automatically extract biomedical entities such as therapies, diagnoses, and biological species. Next, they explore few-shot learning techniques and incorporate graph information from the MeSH database for the identified entities. Finally, they investigate two methods for ensembling: (i) combining the predictions of three previous LLMs using an AND strat-

Team	ROUGE1	BERTSCORE	BLEURT	AggregateComposite	AggregateScore	Rank
WangLab *	0.7755	0.8087	0.7831	0.7746	0.7891	1
PromptMind	0.8070	0.8058	0.7470	0.5739	0.7866	2
HSE NLP	0.7795	0.8059	0.7564	0.5117	0.7806	3
KU-DMIS	0.7288	0.7672	0.7047	0.5709	0.7336	4
Maven *	0.7031	0.7437	0.7522	0.5239	0.7330	5
Edinburgh	0.6780	0.7435	0.7111	0.5629	0.7109	6
Clini-						
calNLP						
knowlab_AIN	1 0.6435	0.6767	0.6542	0.5731	0.6581	7
EM_Mixers	0.5713	0.5952	0.5959	0.5475	0.5875	8
IryoNLP	0.5607	0.5916	0.5905	0.5283	0.5810	9
IKIM	0.5233	0.5644	0.5882	0.5500	0.5587	10
CLD-	0.4273	0.4837	0.5318	0.3448	0.4809	11
MEC						
romarcg	0.4323	0.4574	0.4608	0.3227	0.4501	12
mekki	0.4180	0.4592	0.4679	0.3997	0.4483	13
MediFact	0.4540	0.4441	0.4386	0.5353	0.4456	14
harivm	0.1431	0.1345	0.2563	0.1766	0.1780	15
Baseline (GPT-4)	0.5559	0.5801	0.5900	0.4726	0.5754	-

Table 3: Official Results of Error Sentence Correction (Subtask C). The teams are ranked according to AggregateScore. * Potential use of MS test data.

egy 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 in BERTScore (0.8059) and third in aggregated score (0.7806), with an error flag accuracy of 0.5222 and an error sentence detection accuracy of 0.5200.

The KU-DMIS team (Hwang et al., 2024) generated a Chain-of-Thought reasoning dataset using GPT-4 and MEDIQA-CORR dataset. Subsequently, they fine-tuned Meerkat-7B with this generated dataset to enhance its error detection and correction capabilities. The fine-tuned model achieved an aggregate score of 0.7336 in error sentence correction, with a 0.6346 error flag accuracy and 0.6151 error sentence detection accuracy.

The Maven team (Jadhav et al., 2024) conducted Named Entity Recognition (NER) using GEMINI to identify words representing diseases or vaccines in the text. After masking these identified words, the team implemented the Retrieval-Augmented Generation (RAG) model on external datasets . If the RAG score fell below a certain threshold, they passed the input to the model, which was created by quantizing Palmyra-20b (Team, 2023) using 4bit quantization and then fine-tuned it using the QLoRA technique on MedQA data (possible test data leakage). If the word provided by Palmyra or RAG model matched the word detected by NER, no error was detected. Otherwise, if a different word was obtained, it was replaced with the masked word identified by NER. Finally the error sentence is mapped with the sentence Id to get the output in desired format.

The Edinburgh Clinical NLP team (Gema et al., 2024) evaluated multiple prompting strategies such as In-context Learning (ICL) and Chain-of-Thought (CoT) to improve LLMs' performance. To aid the error correction LLM, they experimented with integrating a relatively smaller language model (i.e. BioLinkBERT) as an error-span predictor. They integrated the predicted error span in two ways; presenting it as a hint for the LLM to correct the error or presenting it as multiple-choice questions for the LLM to select the most likely one.

The knowlab_AIMed team (Wu et al., 2024) used two methods: (i) Dynamic In-Context Learning with RAG, CoT, and manual analysis. In this method, they performed manual analysis on a subset of the dataset. They used the RAG model to implement dynamic ICL, incorporating CoT prompts. They also used ICL-augmented examples from the training dataset. In the second method, the team utilized the training dataset to prompt an LLM to deduce reasons about the correctness or incorrectness of the clinical notes. By leveraging the LLM's capabilities, the constructed reasons provided additional information and insights into the errors present in the notes. These reasons, along with the ICL examples, were used to train the model for error detection, span identification, and error correction tasks.

The IKIM team (Amdada et al., 2024) trained a linear classifier on embeddings from the model pritamdeka/S-PubMedBert-MS-MARCO to predict whether a sentence potentially contains an error in the MS dataset. They also clustered these sentence embeddings. For each cluster, they leveraged GPT-4 to generate a chain of thought that describes the medical reasoning for a sample from the training dataset. For test predictions, they gave GPT-4 the sentence predicted by the linear classifier along with a chain of thought from the cluster to which the sentence belongs, and prompted it to predict whether the sentence was wrong and to provide a correction if needed. They directly prompted GPT-4 with few-shot examples and a chain of thought prompt for UW samples, without clustering or sentence selection.

The MediFact team (Saeed, 2024) employed weakly-supervised SVM and extractive QA for observed errors, alongside pre-trained QA models for unseen errors in clinical text correction. The team achieved the second best score in error flag detection with an accuracy of 0.7373, and an aggregate score of 0.4456 in error sentence correction.

6 Conclusion

The MEDIQA-CORR shared task was tackled by a wide variety of approaches from the participating teams. Ranging from algorithmic reasoning approaches leveraging the LLMs as intermediate extraction tools (e.g., for NER) to approaches that are fully controlled by LLMs and prompting techniques. The best performing methods were datasetdependent, i.e., different methods or parameters were used for each dataset. Generalized, datasetagnostic, approaches fared reasonably well in comparison. A key challenge was in detecting correctly which text and which sentence contained errors, with only two teams reaching an accuracy above 70% in text flagging, and only one team reaching an accuracy greater than 65% in detecting the sentence containing the error. The detection accuracy impacted the quality of the corrected texts (e.g., providing corrections when the sentence contained no errors) but the correction results were less contrasted in general with six teams reaching an

aggregate score greater than 70%.

Moving forward, optimizing the dataset-agnostic approaches is likely to be a key focus as it has the most impact on production-grade models/systems for clinical note generation/validation. The data provided by MEDIQA-CORR can be leveraged for that as they showed to be sufficiently challenging to be used as a benchmark for generalized approaches.

7 Limitations

The paper does not cover all types of possible methods and models for the detection and correction of medical errors. The MS and UW datasets are also limited in terms of size and types of medical errors. Further experiments and evaluations are needed to validate the best performing methods on other datasets and scenarios.

Acknowledgements

We would like to thank Thomas Lin from Microsoft Health AI and the ClinicalNLP organizers for their feedback and support for the shared task. We also thank our annotation team, Erica Labrie, Loren Kimmel, Seanjeet Paul, Thomas Ryan, Brianna L Cowin, Sabrina J Crooks, Karina Lopez, and Kelsi F Nabity, for annotating the MS and UW datasets in time for the challenge and all the participating teams who contributed to the success of the shared task through their interesting approaches and experiments and strong engagement.

References

- Renad M. Alzghoul, Abdulrahman Tabaza, Aya Abdelhaq, and Ahmad Altamimi. 2024. Cld-mec at mediqa- corr 2024 task: Gpt-4 multi-stage clinical chain of thought prompting for medical errors detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Sigall K. Bell, Tom Delbanco, Joann G. Elmore, Patricia S. Fitzgerald, Alan Fossa, Kendall Harcourt, Suzanne G. Leveille, Thomas H. Payne, Rebecca A. Stametz, Jan Walker, and Catherine M. DesRoches. 2020. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. JAMA Netw Open, 3(6).
- 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, Griffin Adams, Neal Snider, and Meliha Yetisgen. 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, ClinicalNLP@ACL 2023, Toronto, Canada, July 14, 2023, pages 503–513. Association for Computational Linguistics.
- Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023a. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.
- Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023b. An investigation of evaluation metrics for automated medical note generation. In *ACL (Findings) 2023*, Toronto, Canada. Association for Computational Linguistics.
- Jean-Philippe Corbeil. 2024. Iryonlp at mediqa-corr 2024: Tackling the medical error detection & correction task on the shoulders of medical agents. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Aryo Pradipta Gema, Chaeeun Lee, Pasquale Minervini, Luke Daines, T. Ian Simpson, and Beatrice Alex. 2024. Edinburgh clinical nlp at mediqa-corr 2024: Guiding large language models with hints. In Proceedings of the 6th Clinical Natural Language Processing Workshop, Mexico City, Mexico. Association for Computational Linguistics.
- Traber Davis Giardina, Shailaja Menon Helen Haskell, Julia Hallisy, Frederick S. Southwick, Urmimala Sarkar, Kathryn E. Royse, and Hardeep Singh. 2018. Learning from patients' experiences related to diagnostic errors is essential for progress in patient safety. *Health Affairs*, 37(11).
- Satya Kesav Gundabathula and Sriram R Kolar. 2024. Promptmind team at mediqa-corr 2024: Improving clinical text correction with error categorization and llm ensembles. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Hyeon Hwang, Taewhoo Lee, Hyunjae Kim, and Jaewoo Kang. 2024. Ku-dmis at mediqa-corr 2024: Exploring the reasoning capabilities of small language models in medical error correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Suramya Jadhav, Abhay Shanbhag, Sumedh Joshi, Atharva Date, and Sheetal S. Sonawane. 2024. Maven at mediqa-corr 2024: Leveraging rag and medical llm for error detection and correction in medical notes. In *Proceedings of the 6th Clinical Natural*

Language Processing Workshop, Mexico City, Mexico. Association for Computational Linguistics.

- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.
- Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: Opportunities and challenges. *Cureus*, 15(5).
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A dataset for commonsense reasoning over entity knowledge. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual.
- Juan Pajaro, Edwin Puertas, David Villate, Laura Estrada, and Laura Tinjaca. 2024. Verbanexai at mediqa-corr: Efficacy of gru with biowordvec and clinicalbert in error correction in clinical notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Swati Rajwal, Eugene Agichtein, and Abeed Sarker. 2024. Em_mixers at mediqa-corr 2024: Knowledgeenhanced few-shot in-context learning for medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Nadia Saeed. 2024. Medifact at mediqa-corr 2024: Why ai needs a human touch. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digit*. *Medicine*, 7(1).
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,*

ACL 2020, Online, July 5-10, 2020, pages 7881–7892. Association for Computational Linguistics.

- Liyan Tang, Zhaoyi Sun, Betina Ross S. Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *npj Digit. Medicine*, 6.
- Writer Engineering Team. 2023. Palmyra-large parameter autoregressive language model. https: //dev.writer.com.
- Augustin Toma, Ronald Xie, Steven Palayew, Gary D. Bader, Patrick Lawler, and BO WANG. 2024.
 Wanglab at mediqa-corr 2024: Optimized llm-based programs for medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Airat Valiev and Elena Tutubalina. 2024. Hse nlp team at mediqa-corr 2024 task: In-prompt ensemble approach with named entity recognition and knowledge graph for medical error checking. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-

2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 307–321, Barcelona (online). International Committee for Computational Linguistics.

- Zhaolong Wu, Abul Hasan, Jinge Wu, Yunsoo Kim, Jason Pui-Yin Cheung, Teng Zhang, and Honghan Wu. 2024. Knowlab_aimed at mediqa-corr 2024: Chainof-though (cot) prompting strategies for medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.
- Wen-wai Yim, Asma Ben Abacha, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations. In Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of CEUR Workshop Proceedings, pages 1347–1360. CEUR-WS.org.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.