MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes

Asma Ben Abacha[1] abenabacha@microsoft.com
Wen-wai Yim[1] yimwenwai@microsoft.com
Yujuan Fu[2] velvinfu@uw.edu
Zhaoyi Sun[2] zhaoyis@uw.edu
Meliha Yetisgen[2] melihay@uw.edu
Fei Xia[2] fxia@uw.edu
Thomas Lin[1] tlin@microsoft.com

[1]Microsoft, Health and Life Sciences AI, Redmond [2]University of Washington, Seattle

Abstract

Several studies have shown that Large Language Models (LLMs) can answer medical questions correctly, even outperforming the average human score in some medical exams. However, to our knowledge, no study has been conducted to assess the ability of language models to validate existing or generated medical text for correctness and consistency. In this paper, we introduce MEDEC[1], the first publicly available benchmark for medical error detection and correction in clinical notes, covering five types of errors (Diagnosis, Management, Treatment, Pharmacotherapy, and Causal Organism). MEDEC consists of 3,848 clinical texts, including 488 clinical notes from three US hospital systems that were not previously seen by any LLM. The dataset has been used in the MEDIQA-CORR 2024 shared task to evaluate seventeen participating systems. In this paper, we describe the data creation methods and we evaluate recent LLMs (e.g., o1-preview, GPT-4, Claude 3.5 Sonnet, Gemini 2.0 Flash, and DeepSeek-R1) for the tasks of detecting and correcting medical errors requiring both medical knowledge and reasoning capabilities. We also conducted a comparative study where two medical doctors performed the same task on the MEDEC test set. The results showed that MEDEC is a sufficiently challenging benchmark to assess the ability of models to validate existing or generated notes and to correct medical errors. We also found that although recent LLMs have a good performance in error detection and correction, they are still outperformed by medical doctors in these tasks. We discuss the potential factors behind this gap, the insights from our experiments, the limitations of current evaluation metrics, and share potential pointers for future research.

1 Introduction

A survey study from US health care organizations showed that one in five patients who read clinical notes reported finding mistakes and 40% perceived the mistake as serious, with the most common category of mistakes being related to current or past diagnoses (Bell et al., 2020).

On the other hand, more and more medical documentation tasks (e.g., clinical note generation) are being supported by LLMs. In multiple studies, LLMs have shown the ability to answer accurately questions from medical exams (Gilson et al., 2023; Johnson et al., 2023; Schubert et al., 2023) and to imitate clinical reasoning in providing diagnoses (Savage et al., 2024).

However, one of the main obstacles in adopting LLMs in medical documentation tasks is their potential to generate hallucinations or incorrect information (Tang et al., 2023) and harmful content that might alter clinical decision making (Chen et al., 2024). Rigorous validation methods are essential to mitigate these risks and make LLMs safer to use for medical content generation (Karabacak and Margetis, 2023).

Relevant benchmarks are required to assess whether such validation can be fully automated. A key task in this regard is the ability to detect and correct medical errors in clinical texts.

Most previous studies on (common sense) error detection have focused on the general domain (Wang et al., 2020; Onoe et al., 2021). In this paper, we tackle the problem of identifying and correcting medical errors in clinical texts. From a human perspective, identifying and correcting these errors requires medical expertise, specialized knowledge, and sometimes practical experience. We introduce a new dataset, MEDEC, and experiment with different recent LLMs (e.g., Claude 3.5 Sonnet, o1-preview, Gemini 2.0 Flash, and DeepSeek-R1). To the best of our knowledge, this is the first publicly available benchmark and study on automatic error detection and correction in clinical notes.

[1]https://github.com/abachaa/MEDEC

Findings of the Association for Computational Linguistics: ACL 2025, pages 22539–22550

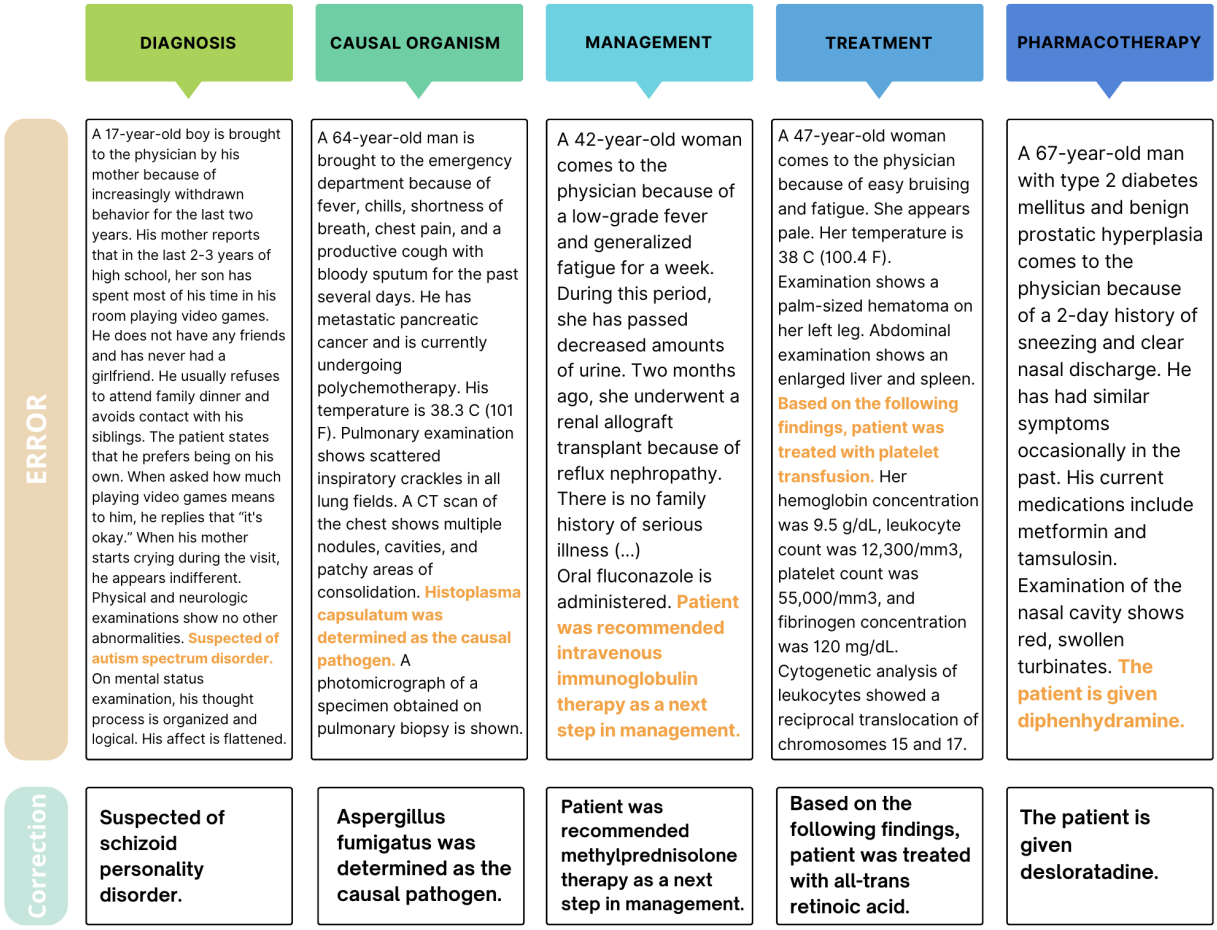

Figure 1: Examples from the MEDEC (*MS*) dataset.

2 Related Work

Jang et al. (2022) introduced a benchmark for consistency evaluation and evaluated pretrained language models (e.g, BERT, T5, and GPT-2) on three main categories: semantic, logical, and factual consistency. They found that those language models do not perform well in every test case and have a high level of inconsistency in many cases. Jang and Lukasiewicz (2023) investigated the trustworthiness of more recent language models, ChatGPT and GPT-4, regarding semantic consistency and found that while both models appear to show an enhanced language understanding and reasoning ability, they often fail at generating logically consistent predictions.

In the medical domain, several recent studies evaluated large language model accuracy and consistency. Johnson et al. (2023) conducted a study to assess the accuracy and reliability of medical responses generated by ChatGPT. Thirty-three physicians across 17 specialties generated 284 medical questions with different levels of difficulty and graded ChatGPT's answers for accuracy and completeness. While most of the generated text was evaluated by physicians as accurate, there were potential limitations in handling complex medical questions.

In two separate studies, Schubert et al. (2023) and Gilson et al. (2023) found that GPT models can answer medical questions correctly in neurology board–style examinations and the United States Medical Licensing Examination (USMLE) Step 1 and Step 2 exams, even outperforming the average human score in some instances.

Chen et al. (2024) assessed the effect and safety of LLM-assisted patient messaging, as one of the earliest applications of LLMs in electronic health records (EHRs). The fact that LLM-assisted responses were more similar to the LLM drafts than to the manual responses, together with the improved interphysician agreement, suggested that doctors might adopt the LLM's responses and assessments. The study also found that a minority of LLM drafts, if left unedited, could lead to severe harm or death.

The safe introduction and use of LLMs in medical documentation tasks requires reliable and automatic validation methods. However, as far as we know, no benchmark was made publicly available to assess the ability of LLMs in validating existing or generated medical text for correctness and consistency.

In this paper, we present MEDEC, the first benchmark for medical error detection and correction in clinical notes. We describe the data creation methods and we evaluate recent state-of-the-art open domain LLMs for these tasks.

The MEDEC dataset has been used in the first shared task on medical error detection and correction, MEDIQA-CORR 2024, to evaluate models and solutions from seventeen participating teams (Ben Abacha et al., 2024).

3 MEDEC Dataset

MEDEC contains 3,848 clinical texts from different specialties. Eight medical annotators participated in the annotation task. The dataset covers five types of errors:

- *Diagnosis*: The provided diagnosis is inaccurate.
- *Management*: The next step provided in management is inaccurate.
- *Pharmacotherapy*: The recommended pharmacotherapy is inaccurate.
- *Treatment*: The recommended treatment is inaccurate.
- *CausalOrganism*: The indicated causal organism or causal pathogen is inaccurate.

These error types were selected after analyzing the most frequent question types identified in official medical board exams. The distribution of error types followed the original distribution of question types found in the analyzed question-answer pairs. Figure 2 presents the distribution of error types (Diagnosis, Management, Treatment, Pharmacotherapy, and Causal Organism) in the MEDEC dataset.

Each clinical text in the dataset is either correct or contains one error introduced using one of two different methods: *MS* (described in Section 3.1) and *UW* (described in Section 3.2). The main motivation for using two different data creation methods was to diversify the errors through different error injection approaches (i.e., leveraging questions and answers from medical board exams in *MS* vs. manual modification of medical entities or spans in original clinical notes in *UW*). By using varied

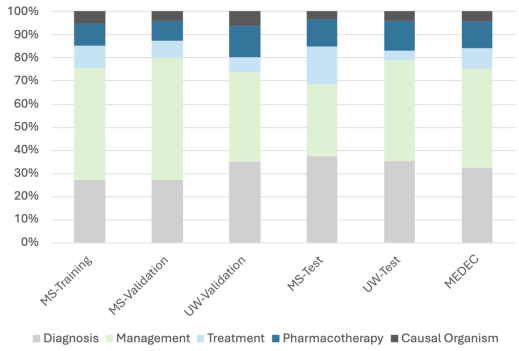

Figure 2: MEDEC - Error Type Distribution.

clinical texts and multiple error injection methods, we aim to enable a more comprehensive evaluation of the models' ability to handle a broader range of scenarios.

Table 1 presents the training, validation, and test splits. The *MS* training set contains 2,189 clinical texts. The *MS* validation set contains 574 clinical texts and the *UW* validation set contains 160 clinical texts. The MEDEC test set consists of 597 clinical texts from the *MS* collection and 328 clinical texts from the *UW* dataset. 51.3% of the test notes contain errors while 48.7% of the notes are correct.

The MEDEC dataset is available at: `https://github.com/abachaa/MEDEC`. The *MS* subset is publicly available. The *UW* subset requires signing a data usage agreement (DUA). Figure 1 presents examples from the MEDEC-*MS* collection.

3.1 Data Creation Method #1 (*MS*)

In this method, we leverage medical board exams from the MedQA collection (Jin et al., 2020). These exams present realistic medical scenarios and provide valuable resource for assessing medical knowledge and identifying gaps in clinical understanding.

Four annotators with medical backgrounds reviewed the medical narratives and multiple-choice questions, first verifying the accuracy of the original question–answer pairs and excluding those with errors, ambiguity, or missing context (e.g., required exam results). They then modified the scenario text by injecting a plausible but incorrect answer, following these guidelines:

- Using medical narrative multiple choice questions, introduce a wrong answer into the scenario text and create two versions with the error injected either in the middle of the text or at the end.
- Using medical narrative multiple choice ques-

Collection		Training	Validation	Test	Total
MS	# texts	2,189	574	597	3,360
UW	# texts	-	160	328	488
MEDEC	# texts	**2,189**	**734**	**925**	**3,848**
	# texts without errors	*970 (44.3%)*	*335 (45.6%)*	*450 (48.7%)*	*1,755 (45.6%)*
	# texts with errors	*1,219 (55.7%)*	*399 (54.4%)*	*475 (51.3%)*	*2,093 (54.4%)*

Table 1: MEDEC Dataset: Training, Validation, and Test Sets

tions, introduce the right answer into the scenario text to create a correct version, as described in Figure 3 (*Generated Text with Correct Answer)*.

- Check manually if the automatically rewritten text is faithful to the original scenario and the included answer.

We randomly selected one correct and one incorrect version for each note from the two different scenarios (error injected in the middle of the text or at the end) in the final dataset.

3.2 Data Creation Method #2 (*UW*)

We used a database of real clinical notes between 2009 and 2021 from three University of Washington (UW) hospital systems[2]: Harborview Medical Center, UW Medical Center, and Seattle Cancer Care Alliance.

From this database, we randomly selected 488 out of 17,453 diagnosis supports, which summarize patients' medical conditions and provide rationales for treatments.

A team of four medical students manually introduced errors into 244 of these notes. Initially, each note was marked with several candidate entities identified as Unified Medical Language System (UMLS)[3] concepts by QuickUMLS[4].

An annotator either selected a concise medical entity from these candidates or created a new span. This span was then labeled with one of the five error types. The annotator then replaced this span with an erroneous version using similar but distinct concepts, crafted by the annotators themselves or provided by a SNOMED- and LLM-based method. This method was used to suggest alternative concepts to the annotators without using the input text. Medical annotators decided on the final concepts/errors to inject manually in the text.

During this process, each error span was required to contradict at least two other parts of the clinical notes (and annotators provided a justification for each error introduced). We de-identified the clinical notes (post error injection) with Philter[5] for automatic de-identification. Each note was then independently reviewed by two annotators to ensure proper de-identification. A third annotator adjudicated any remaining discrepancies.

4 Medical Error Detection & Correction Approaches

In order to evaluate models on medical error detection and correction, we divide the process into three subtasks:

- *Subtask A:* Predicting the error flag (0: if the text has no error; 1: if the text contains an error).
- *Subtask B:* Extracting the sentence that contains the error for flagged texts (-1: if the text has no error; Sentence ID: if the text contains an error).
- *Subtask C:* Generating a corrected sentence for texts flagged as containing errors (NA: if the text has no error; Generated sentence/correction: if the text has an error).

For comparison, we build LLM-based solutions using two different prompts to generate the outputs required to assess the models on the three subtasks:

- **P#1**: *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 one 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 a medical error related to treatment, management, cause, or diagnosis, return the sentence id of the sentence containing the error, followed by a space, and then a corrected version of the sentence. Finding and correcting the error requires medical knowledge and reasoning.*
- **P#2** Similar to the first prompt, but includes an example, randomly selected from the training set: *Here is an example. 0 A 35-year-old woman presents to her physician with a complaint of pain and*

[2]The MEDEC-*UW* subset requires signing a DUA. Examples presented in this paper are selected from the MEDEC-*MS* subset.

[3]https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

[4]https://github.com/Georgetown-IR-Lab/QuickUMLS

[5]https://github.com/BCHSI/philter-deidstable1_mirror

Initial Question and Correct/Incorrect Answers	
Question	A 4670-g (10-lb 5-oz) male newborn is delivered at term to a 26-year-old woman after prolonged labor. Apgar scores are 9 and 9 at 1 and 5 minutes. Examination in the delivery room shows swelling, tenderness, and crepitus over the left clavicle. There is decreased movement of the left upper extremity. Movement of the hands and wrists are normal. A grasping reflex is normal in both hands. An asymmetric Moro reflex is present. The remainder of the examination shows no abnormalities and an anteroposterior x-ray confirms the diagnosis. **Which of the following is the most appropriate next step in management?**
Options	{'A': 'Nerve conduction study', 'B': 'Surgical fixation', 'C': 'Physical therapy', **'D': 'Pin sleeve to the shirt'**, 'E': 'Splinting of the arm', 'F': 'MRI of the clavicle'}
Answer	**D**

Generated Text with Correct Answer	
Scenario with Answer at the End	A 4670-g (10-lb 5-oz) male newborn is delivered at term to a 26-year-old woman after prolonged labor. Apgar scores are 9 and 9 at 1 and 5 minutes. Examination in the delivery room shows swelling, tenderness, and crepitus over the left clavicle. There is decreased movement of the left upper extremity. Movement of the hands and wrists are normal. A grasping reflex is normal in both hands. An asymmetric Moro reflex is present. The remainder of the examination shows no abnormalities and an anteroposterior x-ray confirms the diagnosis. **They pinned the left sleeve to patient's shirt to allow proper healing.**
Scenario with Answer in the Middle	A 4670-g (10-lb 5-oz) male newborn is delivered at term to a 26-year-old woman after prolonged labor. Apgar scores are 9 and 9 at 1 and 5 minutes. Examination in the delivery room shows swelling, tenderness, and crepitus over the left clavicle. There is decreased movement of the left upper extremity. **Left sleeve was pinned to allow proper healing.** Movement of the hands and wrists are normal. A grasping reflex is normal in both hands. An asymmetric Moro reflex is present. The remainder of the examination shows no abnormalities and an anteroposterior x-ray confirms the diagnosis.
Type	Management
Specialty	Pediatrics

Figure 3: Method #1: Correct answer injected in the question text to create the reference note. The same process was used to inject a selected incorrect answer and to create another version of the note containing a medical error.

stiffness in her hands. 1 She says that the pain began 6 weeks ago a few days after she had gotten over a minor upper respiratory infection (...). In this example, the error is in the sentence number 10: Methotrexate is given. The correction is: Prednisone is given. The output is: 10 1 Prednisone is given. End of Example.

5 Experiments & Results

5.1 Language Models

We experiment with several recent small and large language models:

1. Phi-3-7B, a Small Language Model (SLM) (Abdin et al., 2024)
2. Claude 3.5 Sonnet *(2024-10-22)*, the latest model from the Claude 3.5 family offering state-of-the-art performance across several coding, vision, and reasoning tasks (Anthropic, 2024).
3. Gemini 2.0 Flash: the latest/most advanced Gemini model (Google, 2024). Other Google models such as Med-PaLM models (Singhal et al., 2023), designed for medical purposes, were not publicly available.
4. ChatGPT (Brown et al., 2020; OpenAI, 2023a) and GPT-4, a "high-intelligence" model (OpenAI, 2023c,b).
5. GPT-4o providing "GPT-4-level intelligence but faster" (OpenAI, 2024a) and the GPT-4o-mini *(gpt-4o-2024-05-13)* small model for focused tasks (OpenAI, 2024b).
6. The recent o1-mini *(o1-mini-2024-09-12)* (OpenAI, 2024d) and o1-preview *(o1-preview-2024-09-12)* models with "new AI capabilities" for complex reasoning tasks (OpenAI, 2024c).
7. DeepSeek-R1[6], an open-source large language model that uses reinforcement learning to perform reasoning tasks (DeepSeek-AI et al., 2025).

Few models (e.g., Phi-3 and Claude) required minimal automatic post-processing to correct some formatting issues.

5.2 Evaluation Metrics

To evaluate the models' performance in recognizing medical errors in texts, we relied on $Accuracy$ for Error Flag Prediction (subtask A) and Error Sentence Detection (subtask B).

To further analyze the error detection results for each error type, we also computed the $Recall$ using the subset of test examples with errors (i.e., error flag = 1) for each type.

To evaluate the generated corrections (subtask C), we selected lexical, contextual embedding-based, and medical knowledge-graph embedding-based metrics:

- Three open-domain Natural Language Generation (NLG) metrics, that outperformed other standard NLG metrics in terms of correlation scores with medical experts on clinical datasets (Ben Abacha et al., 2023): $ROUGE - 1$ (Lin, 2004), $BLEURT$ (Sellam et al., 2020), and $BERTScore$ (microsoft/deberta-xlarge-mnli) (Zhang et al., 2020), and their Aggregate Score ($AggregateScore$), which is the average of these three NLG metrics.

[6]We used DeepSeek-R1 available in Azure AI Foundry: https://learn.microsoft.com/en-us/azure/ai-studio/how-to/deploy-models-deepseek?pivots=programming-language-python.

- The medical metric MIST (Ben Abacha et al., 2023) that relies on medical knowledge-graph embedding models to compute the similarity between UMLS concepts associated with the medical entities extracted from the reference and automatic texts[7]. The MIST-COMB variant combines MIST, ROUGE-1-R, and BERTScore-R. MIST and MIST-COMB showed positive correlation with medical experts' judgments on clinical datasets.

We computed these error correction scores when both the reference and system corrections are provided (other than NA). Our evaluation scripts are available at: https://github.com/abachaa/MEDIQA-CORR-2024/tree/main/evaluation.

5.3 Comparison with Expert Labeling

Two medical doctors performed the same subtasks on the MEDEC dataset to assess the difficulty of detecting and correcting the errors. The doctors annotated 569 clinical notes from the full test set of 925 texts, with 242 notes annotated by both to compute inter-annotator agreement (IAA).

Given a clinical text from the test set without the ground truth (without the error flag, error sentence, and reference correction), the medical doctors were tasked to (i) judge whether a medical error exists in the text, (ii) if an error exists, write the sentence ID of the sentence where the error occurred, and (iii) provide the most likely error correction and its type (e.g., diagnosis, management, treatment).

The IAA between the two doctors, measured by accuracy, was 69.01% on error flag detection and 57.85% on error sentence detection, which highlights the challenging nature of the task.

5.4 Results

Table 2 presents the results of the manual annotation performed by the medical doctors and the results of several recent LLMs using the two zero-shot and one-shot prompts described above. Claude 3.5 Sonnet outperformed the other LLM-based methods in error flag detection with 70.16% Accuracy and in error sentence detection with 65.62% Accuracy. The o1-mini model achieved the second best error flag detection Accuracy of 69.08%.

In error correction, o1-preview achieved the best Aggregate Score of 0.698, followed by DeepSeek-R1 with 0.675 Aggregate Score. Although DeepSeek-R1 had lower performance in error flag and error sentence detection, the model was able to provide high-quality corrections on the subset of correctly detected errors.

The medical NLG metric MIST highlighted Claude 3.5 Sonnet as the best model in generating corrections that are similar to the references in terms of medical concepts. This result is in alignment with Claude's best accuracy scores in error flag and error sentence detection.

Table 3 presents the error detection Accuracy and error correction scores on each MEDEC collection. The *MS* subset was more challenging for Claude 3.5 Sonnet and Doctor #2, while the *UW* subset was more challenging for o1-preview and Doctor #1.

The results show that recent LLMs have a good performance in error detection and correction, relative to the doctors' scores, but they are still outperformed by the medical doctors in these tasks. This could be explained by the fact that such error detection and correction tasks are relatively rare online and in medical textbooks, which means that these large models are less likely to have encountered such data in their pretraining. This can be seen specifically in the o1-preview results where the model achieved 73% and 69% Accuracy in error and sentence detection on the *MS* subset that was built from publicly available clinical texts, while achieving only 58% and 48% Accuracy on the *UW* collection of private clinical notes.

Another factor is that the task consists in analyzing and fixing an existing text that was not generated by LLMs, which might have a higher level of difficulty than drafting new answers from scratch.

We observed in early experiments that prompting strategies such as in-context learning (P#2) and chain-of-thoughts improved the performance of older LLMs but did not outperform zero-shot prompting with newer LLMs such as o1-preview. This is likely due to larger pre-training data and improved generalization capabilities of the more recent models. Beyond strategies P#1 and P#2, several additional and potentially more effective prompting approaches remain to be explored for the MEDEC tasks, such as retrieval-augmented prompting (which incorporates relevant external knowledge into the prompt) and instruction-based prompting (where the model is given explicit directives).

[7]Medical entities and their UMLS Concept Unique Identifiers (CUIs) are extracted using the scispaCy (en_core_sci_scibert) medical entity linking model (Neumann et al., 2019) with a threshold of 0.7.

Model	Error Detection Accuracy		Error Correction					
	Err Flag	Err Sentence	ROUGE-1	BERTScore	BLEURT	AggScore	MIST	MIST-COMB
Phi-3	0.5276	0.2443	0.2606	0.1514	0.4683	0.2935	0.7506	0.5475
GPT-4o-mini	0.6086	0.4757	0.5148	0.5089	0.5640	0.5292	0.6882	0.6236
o1-mini	0.6908	0.5968	0.6052	0.6275	0.6246	0.6191	0.6277	0.6284
Claude 3.5 Sonnet	0.7016	0.6562	0.2253	0.1033	0.5100	0.2795	0.9325	0.6943
Claude 3.5 Sonnet*	0.6800	0.6508	0.2249	0.1125	0.5081	0.2818	0.9120	0.7074
Gemini 2.0 Flash	0.5805	0.3535	0.3769	0.3127	0.4865	0.3920	0.7774	0.6425
ChatGPT	0.4811	0.4800	0.4198	0.3235	0.5133	0.4189	0.6717	0.5982
GPT-4o	0.6584	0.5665	0.5517	0.5373	0.5852	0.4682	0.6751	0.6345
GPT-4o*	0.6368	0.5449	0.5805	0.5401	0.6022	0.5743	0.6600	0.6269
o1-preview	0.6746	0.6140	0.6884	0.7095	0.6949	0.6976	0.7027	0.7198
GPT-4	0.6573	0.5568	0.5553	0.5804	0.5896	0.5751	0.6528	0.6245
GPT-4*	0.6519	0.5773	0.6271	0.6522	0.6368	0.6387	0.6507	0.6613
DeepSeek-R1	0.5168	0.4605	0.6630	0.6921	0.6703	0.6751	0.7111	0.7068
	Medical Doctors							
Doctor #1	**0.7961**	0.6588	0.3863	0.4653	0.5066	0.4527	0.6213	0.5165
Doctor #2	0.7161	**0.6677**	**0.7260**	**0.7315**	**0.6780**	**0.7118**	**0.6738**	**0.7141**

Table 2: Accuracy of error (flag & sentence) prediction and error sentence correction scores. * Uses $P\#2$ prompt. Best LLM scores are double underlined. Second best scores are underlined. Best Error Detection Accuracy achieved by Claude followed by o1-mini (but lower than both doctors' accuracy scores). o1-preview and DeepSeek-R1 achieved the best error correction $AggregateScore$ (but lower than Doctor#2 score).

Dataset	Error Detection Accuracy		Error Correction			
	Error Flag	Error Sentence	ROUGE-1	BERTScore	BLEURT	AggregateScore
	Claude 3.5 Sonnet *(2024-10-22)*					
MS Subset	0.6750	0.6348	0.1822	0.0793	0.4996	0.2537
UW Subset	**0.7500**	**0.6951**	**0.3100**	**0.1508**	**0.5305**	**0.3304**
	o1-preview *(2024-09-12)*					
MS Subset	**0.7286**	**0.6884**	0.6857	**0.7227**	**0.7046**	**0.7043**
UW Subset	0.5762	0.4787	**0.6936**	0.6848	0.6767	0.6850
	Medical Doctor #1					
MS Subset	**0.8125**	**0.7670**	**0.4199**	**0.5127**	**0.5394**	**0.4907**
UW Subset	0.7595	0.4177	0.3073	0.3542	0.4298	0.3638
	Medical Doctor #2					
MS Subset	0.6890	0.6459	0.6845	0.6981	0.6503	0.6776
UW Subset	**0.7723**	**0.7129**	**0.8016**	**0.7926**	**0.7284**	**0.7742**

Table 3: Accuracy and error correction scores on each subset: *MS* & *UW* test sets. The MEDEC-*MS* subset was more challenging for Claude and Doctor #2. MEDEC-*UW* was more challenging for o1-preview and Doctor #1.

Table 4 presents the error detection Recall and error correction scores for each error type (Diagnosis, Management, Treatment, Pharmacotherapy, and Causal Organism). The o1-preview model had substantially higher error flag and sentence detection Recall scores across all error types compared to Claude 3.5 Sonnet and both doctors. Combined with the overall Accuracy results (cf. Table 2), where the doctors achieved better Accuracy, these results indicate that the model(s) had a substantial issue on the Precision side and hallucinated error presence in many cases compared to medical doctors.

The results also show that there is a ranking discrepancy between classification performance and error correction generation performance. For instance, Claude 3.5 Sonnet was first in Accuracy of error flag and sentence detection among all the models, but was last in correction generation scores (cf. Table 2). Also, o1-preview was fourth in error detection Accuracy among all the LLMs, but was first and substantially ahead in correction generation. The same pattern could be observed between the two medical doctors.

Error Type	Error Detection Recall		Error Correction			
	Error Flag	Error Sentence	ROUGE-1	BERTScore	BLEURT	AggregateScore
	Claude 3.5 Sonnet *(2024-10-22)*					
Diagnosis	0.5977	0.5344	0.2416	0.1051	0.5390	0.2953
Management	0.6131	0.4881	0.2157	0.0968	0.4877	0.2667
Treatment	0.6034	0.5345	0.1607	0.0831	0.4890	0.2442
Pharmacotherapy	0.7017	0.6316	0.2577	0.1401	0.5089	0.3023
Causal Organism	0.8333	0.7222	0.2422	0.0851	0.5130	0.2801
	o1-preview *(2024-09-12)*					
Diagnosis	**0.9655**	**0.8391**	0.7706	0.7852	**0.7447**	0.7668
Management	**0.9345**	**0.7679**	0.5697	0.6039	0.6125	0.5954
Treatment	**0.9310**	**0.8965**	**0.7034**	**0.7628**	**0.7207**	**0.7289**
Pharmacotherapy	**0.9649**	**0.8947**	0.7536	0.7369	**0.7406**	**0.7437**
Causal Organism	**1.0000**	**1.0000**	**0.7131**	**0.6802**	**0.7318**	**0.7084**
	Medical Doctor #1					
Diagnosis	0.8333	0.6863	0.4810	0.5616	0.5668	0.5365
Management	0.8267	0.6000	0.2788	0.3375	0.4371	0.3511
Treatment	0.7200	0.6800	0.2726	0.4032	0.4316	0.3691
Pharmacotherapy	0.8000	0.7200	0.4377	0.5319	0.5371	0.5022
Causal Organism	0.7273	0.7273	0.3664	0.4309	0.5090	0.4354
	Medical Doctor #2					
Diagnosis	0.7232	0.6786	**0.8121**	**0.8128**	0.7413	**0.7887**
Management	0.6893	0.6311	**0.6763**	**0.6774**	**0.6487**	**0.6675**
Treatment	0.7273	0.6970	0.5594	0.6147	0.5770	0.5837
Pharmacotherapy	0.8182	0.7576	**0.7592**	**0.7464**	0.6774	0.7277
Causal Organism	0.4286	0.2857	0.4474	0.4632	0.4141	0.4415

Table 4: Recall and error correction scores for each error type using the subset of test examples with errors. The size of each reference subset is as follows: Diagnosis (174 texts), Management (168), Treatment (58), Pharmacotherapy (57), and Causal Organism (18).

Part of it could be explained by the difficulty of the correction generation task, but also, the limitations of current SOTA text generation evaluation metrics in capturing synonyms and similarities in medical texts.

Table 5 presents examples from the reference texts, doctors' annotations, and automatically generated corrections by Claude 3.5 Sonnet and GPT models. For instance, the reference correction of the second example indicates that the patient is diagnosed with *Bruton agammaglobulinemia*, while the LLMs provided correct answers mentioning *X-linked agammaglobulinemia* (a synonym of the same rare genetic disease).

Also, some LLMs such as Claude provide long answers/corrections with more explanation. Similar observations can be found within the doctors' annotations, where Doctor #1 provided longer corrections than Doctor #2, and both doctors had different opinions in some examples/cases, reflecting some of the differences in style and content found in clinical notes written by different doctors/specialists.

Our observations suggest that future research on medical error detection and correction could benefit from incorporating in-context learning strategies and retrieval-augmented prompting. In this introductory study, we focused on evaluating state-of-the-art open-domain LLMs; however, future work should also consider specialized medical language models and explore new evaluation metrics tailored to clinical texts.

6 Conclusion

This paper presented the MEDEC benchmark for medical error detection and correction in clinical notes. An empirical evaluation of LLM-based methods showed that, while recent LLMs have a good performance, they are still outperformed by medical doctors. The results of the doctors' annotation showed that the MEDEC dataset is a sufficiently challenging benchmark to assess the ability of models to validate existing or generated notes and to correct medical errors. We hope that this dataset will enable further studies on medical error

	MEDEC Dataset (Text with Error and its Correction)		Doctors' corrections		LLMs' corrections	
	Text with Error highlighted in Bold	**Reference**	**Doctor #1**	**Doctor #2**	**GPT**	**Claude**
Diagnosis (ms-test-141)	A 23-year-old man presents with sudden loss of consciousness while pitching in a baseball game. There is no history of injury. Consciousness is regained after cardiopulmonary resuscitation. Past medical history is negative for any neurological and cardiovascular problems. **Patient is diagnosed with aortic stenosis after physical examination reveals a prominent A wave on the jugular venous pulse and a double apical impulse.** There are no audible murmurs. An S4 is present.	Patient is diagnosed with hypertrophic cardiomyopathy after physical examination reveals a prominent A wave on the jugular venous pulse and a double apical impulse.	The patient s diagnosed with hypertrophic cardiomyopathy afte physical examination reveal a double apical impulse, prominent "a wave" and S4.	Patient is diagnosed with hypertrophic cardiomyopathy after physical examination reveals a prominent A wave on the jugular venous pulse and a double apical impulse.	*[GPT-4 & GPT-4o & o1-preview & o1-mini]* Patient is diagnosed with hypertrophic cardiomyopathy after physical examination reveals a prominent A wave on the jugular venous pulse and a double apical impulse.	A systolic ejection murmur is present at the right upper sternal border. The error is in sentence 5. In aortic stenosis, a characteristic systolic ejection murmur is typically heard best at the right upper sternal border (2nd right intercostal space) and often radiates to the carotid arteries. This is one of the key diagnostic features of aortic stenosis, along with the other findings mentioned in the case (prominent A wave in JVP, double apical impulse, and S4)...
Diagnosis (ms-test-234)	A 4-year-old boy presents with a history of recurrent bacterial infections, including several episodes of pneumococcal sepsis. His 2 maternal uncles died after having had similar complaints. **Patient is diagnosed with common variable immunodeficiency after lab investigations reveal an undetectable level of all serum immunoglobulins.**	Patient is diagnosed with Bruton agammaglobulinemia after lab investigations reveal an undetectable level of all serum immunoglobulins.	The differential diagnosis includes autosomal recessive agammaglobulinemia, but further evaluation is necessary.	Patient is diagnosed with X-linked agammaglobulinemia after lab investigations reveal an undetectable level of all serum immunoglobulins.	*[GPT-4 & GPT-4o & o1-preview & o1-mini]* Patient is diagnosed with X-linked agammaglobulinemia after lab investigations reveal an undetectable level of all serum immunoglobulins.	Patient is diagnosed with X-linked agammaglobulinemia after lab investigations reveal an undetectable level of all serum immunoglobulins. The error is in sentence 2. Given the clinical presentation (young male patient with recurrent bacterial infections), family history (maternal uncles affected, suggesting X-linked inheritance), and laboratory findings...
Management (ms-test-261)	A 50-year-old woman visits her primary care practitioner with the complaints of generalized weakness, lightheadedness, and fatigability for the past month. She also claims to have epigastric pain, heartburn, and a sensation of a lump in her throat. Her family history is irrelevant, but her medical history is significant for a Pap smear that reported atypical squamous cells of undetermined significance (ASCUS), which was followed up with a cervical biopsy negative for malignancy. She occasionally takes over-the-counter medicines to cope with the heartburn, and recently her hemoglobin was found to be 11 g/dL, for which she received iron and vitamin B12 supplementation. **Fecal occult blood test is completed.** Physical examination is unremarkable, except for pale skin, and a pulse of 120/min.	The patient is referred for an endoscopy.	She has not had followup after her cervical biopsy and HPV testing is indicated.	Text annotated as CORRECT	*[GPT-4-P1]* Physical examination is unremarkable, except for pale skin, and a pulse of 100/min. *[GPT-4-P2 & GPT-4o-P1 & GPT-4o-P2]* Text annotated as CORRECT. *[GPT-4o-mini & o1-preview & o1-mini]* She occasionally takes over-the-counter medicines to cope with the heartburn, and recently her hemoglobin was found to be 11 g/dL, for which she received iron supplementation.	Text annotated as CORRECT

Table 5: Examples of manual & automatic corrections. Incorrect annotations/outputs are highlighted in Grey.

detection and correction in clinical notes, enhancing clinical reasoning capabilities of LLMs, and facilitate additional efforts on evaluation metrics for clinical texts and applications.

7 Limitations

The paper does not cover all types of possible methods and models for the detection and correction of medical errors. The dataset is also limited in terms of size and types of medical errors. The *MS* collection, created based on official medical board exam materials, reflects the original content and answer choices as provided in those sources. However it is possible that medical professionals may hold differing opinions or perspectives. The new benchmark is intended for research purposes on evaluating both medical knowledge and reasoning capabilities of language models and should not be used for medical diagnosis or treatment.

8 Ethics Statement

Medical doctors and annotators were paid a fair hourly wage consistent with the practice of the state of hire.

Acknowledgements

We thank the doctors who participated in this study as well as our annotation team (Erica Labrie, Loren Kimmel, Seanjeet Paul, Thomas Ryan, Brianna L Cowin, Sabrina J Crooks, Karina Lopez, and Kelsi F Nabity).

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/claude/sonnet. Accessed: 12/2024.

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, Yujuan 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*, pages 596–603, Mexico City, Mexico. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation metrics for automated medical note generation. In *ACL (Findings) 2023*, Toronto, Canada. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H Kann, Fallon E Chipidza, Jonathan Leeman, Hugo J W L Aerts, Timothy Miller, Guergana K Savova, Jack Gallifant, Leo A Celi, Raymond H Mak, Maryam Lustberg, Majid Afshar, and Danielle S Bitterman. 2024. The effect of using a large language model to respond to patient messages. *Lancet Digit Health 6(6):e379-e381*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan,

Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartas. 2023. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*.

Google. 2024. Gemini 2.0 flash. https://gemini.google.com. Accessed: 12/2024.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. BECEL: benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3680–3696. International Committee on Computational Linguistics.

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15970–15985. 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.

Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, Elizabeth Scoville, Tyler Reese, Debra Friedman, Julie Bastarache, Yuri van der Heijden, Jordan Wright, Nicholas Carter, Matthew Alexander, Jennifer Choe, Cody Chastain, John Zic, Sara Horst, Isik Turker, Rajiv Agarwal, Evan Osmundson, Kamran Idrees, Colleen Kieman, Chandrasekhar Padmanabhan, Christina Bailey, Cameron Schlegel, Lola Chambless, Mike Gibson, Travis Osterman, and Lee Wheles. 2023. Assessing the accuracy and reliability of ai-generated medical responses: An evaluation of the chat-gpt model. *PREPRINT (Version 1) available at Research Square*.

Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: Opportunities and challenges. *Cureus*, 15(5).

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.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. 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*.

OpenAI. 2023a. gpt-3.5-turbo. https://platform.openai.com/docs/models#gpt-3-5-turbo. Accessed: 03/2024.

OpenAI. 2023b. Gpt-4. https://platform.openai.com/docs/models#gpt-4-turbo-and-gpt-4. Accessed: 03/2024.

OpenAI. 2023c. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024a. Gpt-4o. https://platform.openai.com/docs/models#gpt-4o. Accessed: 09/2024.

OpenAI. 2024b. Gpt-4o mini. https://platform.openai.com/docs/models#gpt-4o-mini. Accessed: 09/2024.

OpenAI. 2024c. o1. https://platform.openai.com/docs/models#o1. Accessed: 12/2024.

OpenAI. 2024d. o1-mini. https://platform.openai.com/docs/models#o1. Accessed: 12/2024.

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).

Marc Cicero Schubert, Wolfgang Wick, and Varun Venkataramani. 2023. Performance of Large Language Models on a Neurology Board–Style Examination. *JAMA Network Open*, 6(12):e2346721–e2346721.

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.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas,

Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617.

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.

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.

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.