MediFact at MEDIQA-CORR 2024: Why AI Needs a Human Touch

Nadia Saeed
Computational Biology Research Lab
Department of Computer Science
National University of Computer and Emerging Sciences (NUCES-FAST)
Islamabad, Pakistan
i181606@nu.edu.pk

Abstract

Accurate representation of medical information is crucial for patient safety, yet artificial intelligence (AI) systems, such as Large Language Models (LLMs), encounter challenges in error-free clinical text interpretation. This paper presents a novel approach submitted to the MEDIQA-CORR 2024 shared task (Ben Abacha et al., 2024a), focusing on the automatic correction of single-word errors in clinical notes. Unlike LLMs that rely on extensive generic data, our method emphasizes extracting contextually relevant information from available clinical text data. Leveraging an ensemble of extractive and abstractive question-answering approaches, we construct a supervised learning framework with domain-specific feature engineering. Our methodology incorporates domain expertise to enhance error correction accuracy. By integrating domain expertise and prioritizing meaningful information extraction, our approach underscores the significance of a human-centric strategy in adapting AI for healthcare.[1]

1 Introduction

Accurately identifying pathogens from textual descriptions of symptoms is crucial in effective healthcare management (Qian and Morral, 2022). However, existing datasets often present significant challenges that hinder reliable inferences and accurate pathogen identification, especially for rare diseases with limited data availability (Wang et al., 2021; Qian and Morral, 2022).

One major challenge lies in the inherent linguistic ambiguities present within these descriptions. Synonyms, homonyms, and polysemy (words with multiple meanings) can lead to confusion and misinterpretations (Karabacak and Margetis, 2023). For example, the term "fever" could indicate a wide range of illnesses, making it difficult to pinpoint the specific pathogen without additional context. Additionally, the distribution of diagnostic and pathogen information within the data can be imbalanced, with some diseases being vastly overrepresented compared to others. This imbalance can skew the model's performance and hinder its ability to accurately identify pathogens for less frequently encountered diseases (Thirunavukarasu et al., 2023; Wang et al., 2021).

Furthermore, incorporating sensitive diagnostic data for training LLMs raises significant ethical concerns regarding patient privacy and authorization requirements (Kelly, 2002). Moreover, pretrained LLMs often learn from vast amounts of generic text data, which might not be tailored to the specific domain of pathogenic research (Qian and Morral, 2022). This lack of domain-specific knowledge can hinder their ability to capture the nuances of rare disease entities and the intricate relationships between textual descriptions and underlying pathogens (Thirunavukarasu et al., 2023; Chanda et al., 2022).

Existing approaches to medical text correction have explored various techniques, including rule-based systems like MetaMap (which utilizes predefined rules to map terms to standardized medical concepts) and machine learning algorithms like RNN-based models (trained to identify and correct errors based on patterns learned from training data) (Chanda et al., 2022; Kumar et al., 2021; Minaee et al., 2021). However, these methods often struggle with the complexity of medical terminology, the inherent ambiguities of natural language, and the limitations of rule-based systems in capturing the ever-evolving nuances of medical language (Qian and Morral, 2022).

While recent advancements in LLMs have shown promise in various natural language processing tasks like text correction, their application in medical diagnostics necessitates careful considera-

[1]Code is available: https://github.com/NadiaSaeed/MediFact-MEDIQA-CORR-2024

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 346–352

tion due to the sensitivity of the data and the need for domain-specific knowledge. Existing LLM-based medical text correction approaches primarily address basic issues like typos and grammatical errors (Thirunavukarasu et al., 2023; Lee et al., 2022). However, they often fall short in addressing patient hallucinations, which can introduce factual errors and lead to misdiagnosis (Wang et al., 2023). Additionally, fine-tuning these models on relevant datasets often yields limited improvements, with models producing generic corrections instead of medically accurate ones (Lee et al., 2022).

This paper aims to present a methodology for automatically correcting single-word errors in clinical notes, submitted to the MEDIQA-CORR 2024 shared task (Ben Abacha et al., 2024a). The approach utilizes supervised learning with tailored feature engineering for the medical domain, emphasizing meaningful information extraction from clinical text data. Two distinct strategies are employed: an extractive question-answering (QA) approach for observed error-correction pairs and an abstractive QA approach for unobserved relations. This framework addresses the following important research questions: 1) How can domain expertise be further integrated into the model to improve its accuracy and ability to explain its reasoning? 2) How can this approach be effectively utilized to assist human reviewers in the process of medical record correction, potentially improving efficiency and accuracy? 3) What ethical considerations are involved in using AI for automatic error correction in healthcare settings, such as potential bias, transparency, and accountability?

2 Methodology

This paper introduces MediFact-CORR QA, a data-efficient approach for one-word error correction in clinical text paragraphs. MediFact-CORR QA leverages a two-stage process combining weakly supervised learning with pre-trained models to address labeled medical text data limitations.

2.1 Error Sentence Identification with Weak Supervision Motivation

MediFact-CORR QA, an innovative framework, employs weakly-supervised learning to discern distinctive patterns in clinical errors within textual data. The process involves analyzing paired paragraphs, each comprising an error-laden version and its corrected counterpart, with the error explicitly annotated. Utilizing Support Vector Machines (SVMs) (Jamaluddin and Wibawa, 2021), the framework effectively discriminates between accurate and erroneous sentences within the clinical domain as shown in Figures 1 and 2 respectively.

This methodology capitalizes on the inherent information within error sentences, thereby mitigating the necessity for extensive labeled datasets. Moreover, the model not only indicates the presence of an error but also precisely identifies the erroneous sentence's location when applicable. Initially training separate SVMs for error and correct sentences, the model's efficacy during testing is indirectly enhanced by the utilization of supervised training labels. Consequently, MediFact-CORR QA proficiently tags erroneous sentences based on acquired patterns from the paired training data.

2.2 Error Correction with Extractive QA

Furthermore, in the process of generating correct sentences, MediFact-CORR QA relies on the inherent structure of the training data and adopts an extractive QA methodology. A notable feature of the MEDIQA-CORR dataset is the existence of paragraph pairs, where one contains an error and the other presents the corrected version (Ben Abacha et al., 2024b). Leveraging this characteristic, MediFact-CORR QA focuses on these error-correction pairs. When identifying sentences as erroneous in Step 1, we apply fuzzy matching between them and their corresponding corrected counterparts from the training data. This fuzzy matching helps to annotate the error information and correct information accurately and efficiently. Through this process, we can locate the most probable correct sentence by finding the matched pair of paragraphs, as they closely resemble each other. Extractive QA proves advantageous in scenarios where the answer can be directly extracted from a given text source. In our context, since the corrected sentence is already present within the training data, MediFact-CORR QA efficiently identifies it through similarity matching. This approach stands out for its data efficiency and effectiveness. Figure 3 depicts the framework where matched paragraph pairs are considered, with one containing error information and the other representing the correct information. This behavior of our dataset is crucial for the extractive QA model, as it allows us to utilize the inherent information within the content. This information is then positioned using

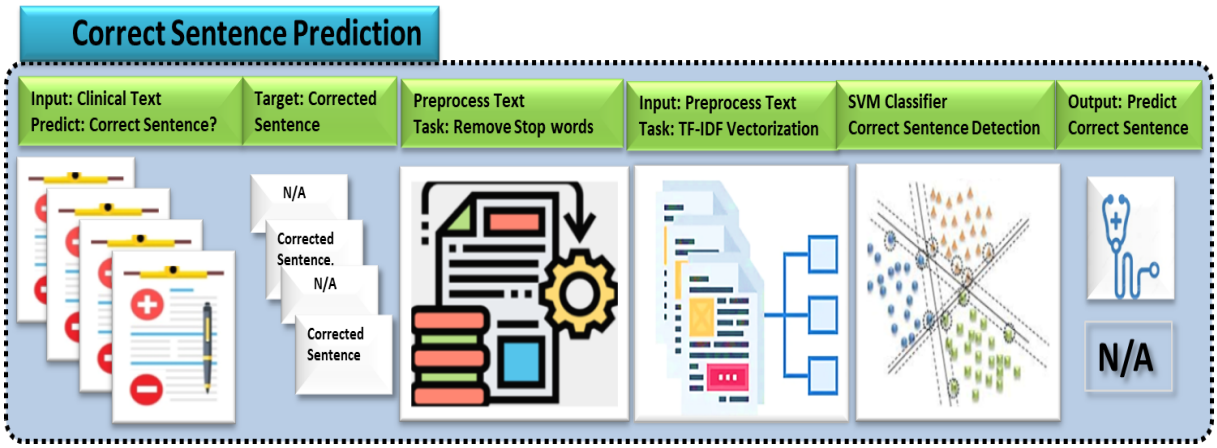

Figure 1: MediFact-CORR: Framework of the Correct SVM model

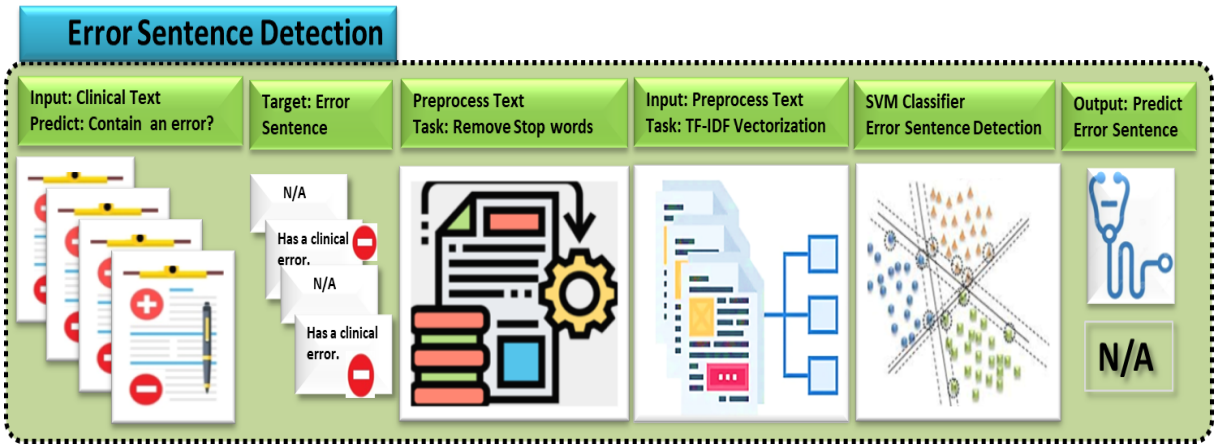

Figure 2: MediFact-CORR: Framework of the Error SVM model

the previously trained SVM models.

2.3 Error Correction with Abstractive QA

Recognizing that not all errors will have corresponding corrected versions in the training data, MediFact-CORR QA employs a pre-trained question-answering (QA) model specifically tailored for unanswerable questions (Lewis et al., 2019). Sentences identified as erroneous in Step 1 but lacking a match in the training data are directed to this pre-trained model. Trained on a vast corpus of text and questions, this model can generate potential corrections for unseen errors by analyzing contextual relationships between words within the erroneous sentence. Pre-trained QA models, having been trained on extensive datasets, excel at handling unseen information and complex language (Cortiz, 2022). Consequently, MediFact-CORR QA can address errors not explicitly present in the training data, thereby enhancing its robustness and generalizability. To illustrate, Figure 4 depicts the framework's step where sentences lacking matched pairs in the training data are passed through the pre-trained QA model for potential corrections (Cortiz, 2022).

By integrating weakly-supervised error detection with extractive QA for observed corrections, and leveraging a pre-trained QA model for unseen errors, MediFact-CORR QA provides a data-efficient solution for error correction in clinical text. This approach is particularly valuable in contexts where access to large labeled medical text data is limited.

3 Experimental Setup and Results

This section details the experimental setup and evaluates the performance of our two-stage model for one-word error correction in clinical text paragraphs.

3.1 Dataset

The MEDIQA-CORR 2024 shared tasks that employed a dataset of clinical texts from the MS and

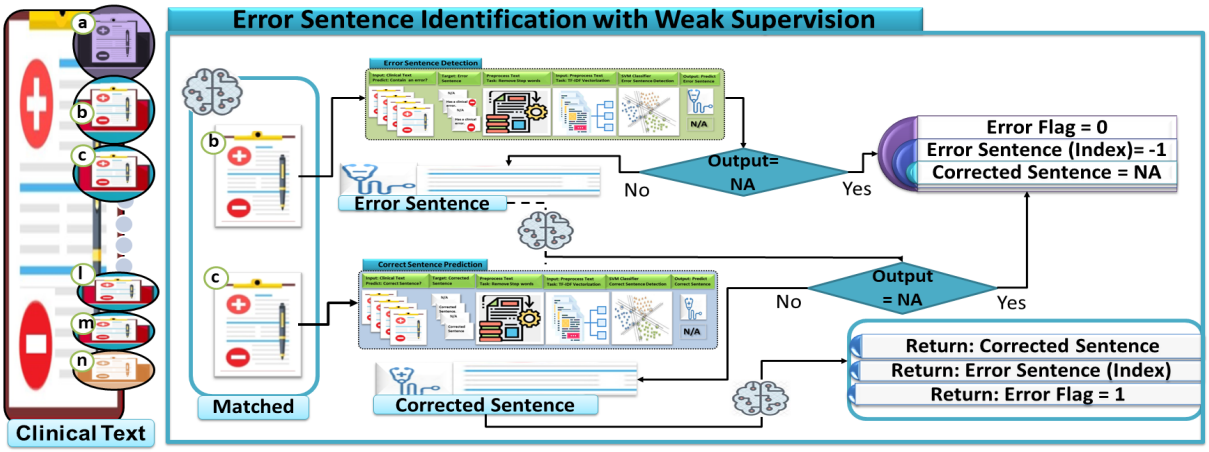

Figure 3: MediFact-CORR: Framework of the Error Correction with Extractive QA

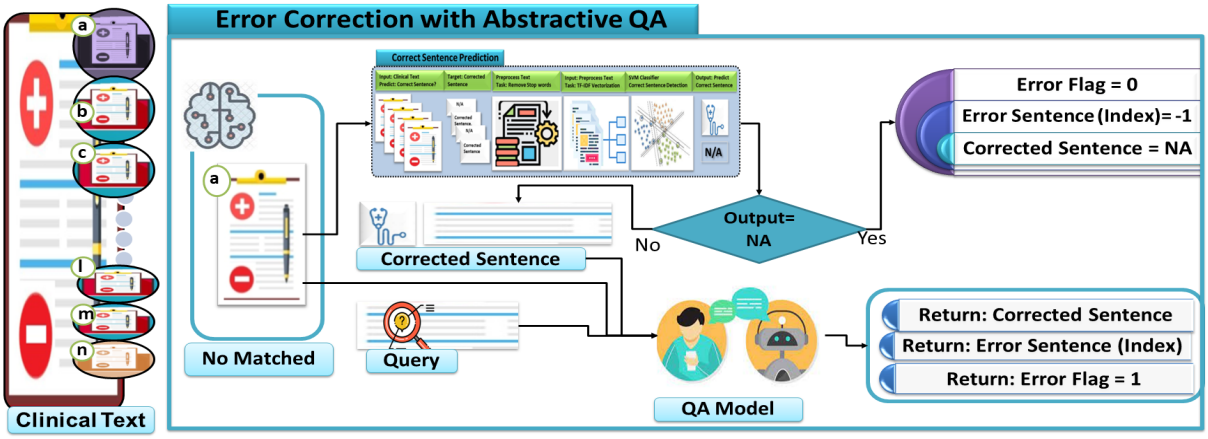

Figure 4: MediFact-CORR: Framework of the Error Correction with Abstractive QA

UW collections (Ben Abacha et al., 2024b). The training set (MS collection) comprised 2,189 texts. Validation sets contained 574 texts from MS and 160 texts from UW. Each text along with the split sentences, Error sentence, and its index, and the corresponding correct sentence, is also given with an error flag. The testing set (MS and UW collection) comprised 925 texts. MEDIQA-CORR 2024 shared tasks comprise three challenging tasks to perform, 1) Error flag prediction, 2) Index of the error sentence detection, and 3) Generate correct sentence.

3.2 Evaluation Metrics

The evaluation has been performed using the available program file by the MEDIQA-CORR 2024 [2]. In performance evaluation following metrics include AggregateScore, R1F score, BERTSCORE, BLEURT, and AggregateC (Yuan et al., 2021; Sellam et al., 2020). *AggregateScore* serves as an overarching metric, consolidating various aspects of model performance, while *R1F* score measures the effectiveness of error correction by considering precision, recall, and F1 score. Additionally, *AggregateC* provides a composite metric summarizing model performance across different dimensions. We also evaluate the model's ability to accurately identify sentences containing errors and pinpoint the precise location of these errors within sentences.

3.3 Results

The models underwent rigorous evaluation across various metrics, including error flag accuracy, error sentence detection accuracy, and Natural Language Generation (NLG) performance. Evaluation was conducted on the validation sets of the MEDIQA_CORR 2024 dataset (Ben Abacha et al., 2024b). Our experimental setup involved training the SVM models using a combination of both train-

[2]MEDIQA-CORR evaluation code: https://github.com/abachaa/MEDIQA-CORR-2024

ing and validation sets. These trained models are now available in our GitHub repository [3].

For the abstraction QA model utilized in the experiment, we leveraged the BART model to answer questions of diagnosing expected medical conditions from provided text (Lewis et al., 2019).

Our performance in the tasks was notably obtained scores out of 106 participants shown in Table 1 (Ben Abacha et al., 2024a). In Task 1 for Error Flags Accuracy, we secured the 2nd rank. For Task 2, which focused on Error Sentence Detection Accuracy, we attained the 8th rank. Task 3 evaluated the Aggregate Score for NLG, where we achieved the 14th rank. Overall, these results underscore the effectiveness of our two-stage model for one-word error correction in clinical text paragraphs, surpassing the performance of the provided baseline model. By integrating error flag prediction, precise sentence extraction, and NLG techniques, we present a promising approach to enhancing the quality and reliability of clinical text data.

4 Discussion

Large Language Models (LLMs) have shown remarkable success in various natural language processing tasks, but their application in medical text correction faces unique challenges (Thirunavukarasu et al., 2023; Wu et al., 2022). Our approach tackles the challenging task of correcting one-word errors in clinical text paragraphs. Unlike LLMs that rely solely on statistical patterns learned from vast amounts of text data, our approach utilizes features specifically tailored to the medical context. This allows the model to leverage domain knowledge and prioritize terms. The example demonstrating the limitations of LLMs and the strengths of SVMs with TF-IDF can be added as a separate paragraph in the same section, following the current paragraph.

Example paragraph: *'A 5-year-old male presents with complaints of a painful mouth/gums, and vesicular lesions on the lips and buccal mucosa for the past 4 days. He is unable to eat or drink due to the pain and reports muscle aches. Vital signs: T 39.1°C, HR 110, BP 90/62 mmHg, RR 18, SpO2 99%. Physical examination reveals vesicular lesions on the tongue, gingiva, and lips, with some ruptured and ulcerated, and palpable cervical and submandibular lymphadenopathy. Patient is diagnosed with an [MASK] infection.'*

While a fine-tuned DistillBERT model predicted a general term like 'goat' or 'Highlander' (Wu et al., 2022). On the other side, our SVM model trained with TF-IDF utilizes domain knowledge through feature weights (Quach et al., 2023). Features like 'vesicular lesions', 'lips', and 'gingiva' receive high weights, guiding the model towards the medically accurate prediction of 'HSV-1' due to its alignment with the clinical context."

Our journey focused on error detection and correction within clinical text data. While Transformer-based models are powerful, their limitations in interpretability, data requirements, and over-fitting prompted us to explore an alternative: SVMs with TF-IDF features. Unlike many models, SVMs offer valuable insights through feature weights (Campbell and Ying, 2022). Features were designed to recognize specific medical terms, abbreviations, and entities like drug names, diagnoses, and anatomical locations. Rules and patterns observed in common errors were translated into features (Quach et al., 2023). Features captured aspects like sentence structure, negation markers, and temporal inconsistencies, which can indicate factual errors like incorrect dates or inconsistent medication names.

The provided dataset posed a unique challenge due to pre-defined sentence indices that deviated from standard newline ("\n") splitting (Ben Abacha et al., 2024b). To address this challenge, we compared detected errors' content with the dataset's available sentences. The index reported in the "Error sentence index" column was predicted as the starting digit of the most similar sentence. Therefore, we must recognize that inherent dataset issues influenced our final score. These challenges underscore the significance of high-quality data for training machine learning models.

In our submission, we investigated three key outcomes in an alternative setting. In the first and second scenarios, utilizing a QA model instead of the static correction model of SVM resulted in an improved R1F score from 0.342 to 0.454. This enhancement underscores the effectiveness of employing a QA model for error correction tasks. Moreover, the accuracy of error sentence detection significantly increased from 0.39 to 0.6 by utilizing the starting digit of the most similar sentence in the pre-defined index of sentences within given samples. This improvement stemmed from addressing an index problem, specifically by selecting the in-

[3]MediFact-SVM models are available: https://github.com/NadiaSaeed/MediFact-MEDIQA-CORR-2024

Model	R1F	BERT	BLEURT	AggScore	AggC	Error Flag	Error Sentence
MediFact_CORR	0.454	0.444	0.439	0.446	0.535	0.737	0.600

Table 1: Performance on error correction tasks, including error flags accuracy and error sentence detection accuracy (submitted at the competition).

dex from the upper part of the sentence. Table 2 provides a summary of these findings.

This research demonstrates the effectiveness of combining human expertise and AI through feature engineering in a supervised learning approach. While SVMs offer interpretability and efficiency, human collaboration remains crucial for optimal performance in complex domains like healthcare (Campbell and Ying, 2022). This collaboration paves the way for improved error detection and correction in clinical text data, ultimately leading to better patient care.

5 Future Work

Our initial success with SVMs for pathogen identification in clinical text data paves the way for further exploration using LLMs. However, LLMs pose unique challenges. Data scarcity, particularly in the specific medical domain, could be a significant hurdle (Wang et al., 2023). Limited data restricts the use of a separate validation set. Future work will explore acquiring more data and data augmentation to enhance model generalizability. Techniques like data augmentation and transfer learning from pre-trained medical LLMs might be crucial to overcome this limitation.

Ethical considerations are paramount, and mitigating biases within the training data is essential. Furthermore, ensuring interpretability through techniques like attention mechanisms is vital for trust and acceptance in healthcare settings.

Finally, for practical implementation, we need to explore computationally efficient LLM architectures or develop task-specific models focused on pathogen identification. Continuous evaluation through techniques like active learning and performance monitoring will be crucial for maintaining a robust, ethical, and interpretable system in real-world clinical text analysis.

References

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*.

Colin Campbell and Yiming Ying. 2022. *Learning with support vector machines*. Springer Nature.

Ashis Kumar Chanda, Tian Bai, Ziyu Yang, and Slobodan Vucetic. 2022. Improving medical term embeddings using umls metathesaurus. *BMC Medical Informatics and Decision Making*, 22(1):114.

Diogo Cortiz. 2022. Exploring transformers models for emotion recognition: A comparision of bert, distilbert, roberta, xlnet and electra. In *Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System*, pages 230–234.

M Jamaluddin and Adhi Dharma Wibawa. 2021. Patient diagnosis classification based on electronic medical record using text mining and support vector machine. In *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pages 243–248. IEEE.

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

Curly Kelly. 2002. Hipaa compliance: Lessons from the repeal of hawaii's patient privacy law. *Journal of Law, Medicine & Ethics*, 30(2):309–312.

A Sampath Kumar, Leta Tesfaye Jule, Krishnaraj Ramaswamy, S Sountharrajan, N Yuuvaraj, and Amir H Gandomi. 2021. Analysis of false data detection rate in generative adversarial networks using recurrent neural network. In *Generative Adversarial Networks for Image-to-Image Translation*, pages 289–312. Elsevier.

Eun Byul Lee, Go Eun Heo, Chang Min Choi, and Min Song. 2022. Mlm-based typographical error correction of unstructured medical texts for named entity recognition. *BMC bioinformatics*, 23(1):1–16.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Model		R1F	BERT	BLEURT	AggScore	AggC	Error Flag	Error Sentence
MediFact_ CORR	Corr+ \n Indexing	0.342	0.355	0.419	0.372	0.508	0.737	0.600
	QA-Model+ \n Indexing	**0.454**	**0.444**	**0.439**	**0.446**	**0.535**	**0.737**	**0.600**
	QA-Model	0.409	0.401	0.418	0.409	0.353	0.507	0.398

Table 2: Performance comparison of different models on error correction tasks, including error flags accuracy and error sentence detection accuracy. The table showcases improvements achieved by employing a QA model and adopting a comprehensive approach to error flag annotation and error sentence detection (results before the competition).

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Gene Qian and Núria Morral. 2022. Role of non-coding rnas on liver metabolism and nafld pathogenesis. *Human Molecular Genetics*, 31(R1):R4–R21.

Luyl-Da Quach, Anh Nguyen Quynh, Nguyen Quoc Khang, and An Nguyen Thi Thu. 2023. Using the term frequency-inverse document frequency for the problem of identifying shrimp diseases with state description text. *International Journal of Advanced Computer Science and Applications*, 14(5).

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

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Haoqing Wang, Huiyu Mai, Zhi-hong Deng, Chao Yang, Luxia Zhang, and Huai-yu Wang. 2021. Distributed representations of diseases based on co-occurrence relationship. *Expert Systems with Applications*, 183:115418.

Hongyan Wang, WeiZhen Wu, Zhi Dou, Liangliang He, and Liqiang Yang. 2023. Performance and exploration of chatgpt in medical examination, records and education in chinese: Pave the way for medical ai. *International Journal of Medical Informatics*, 177:105173.

Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah Goodman. 2022. Causal distillation for language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4295.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.