

PromptMind Team at MEDIQA-CORR 2024: Improving Clinical Text Correction with Error Categorization and LLM Ensembles

Satya K Gundabathula
satyakesav123@gmail.com

Sriram R Kolar
sriramrakshithkolar@gmail.com

Abstract

This paper describes our approach to the MEDIQA-CORR shared task, which involves error detection and correction in clinical notes curated by medical professionals. This task involves handling three subtasks: detecting the presence of errors, identifying the specific sentence containing the error, and correcting it. Through our work, we aim to assess the capabilities of Large Language Models (LLMs) trained on a vast corpora of internet data that contain both factual and unreliable information. We propose to comprehensively address all subtasks together, and suggest employing a unique prompt-based in-context learning strategy. We will evaluate its efficacy in this specialized task demanding a combination of general reasoning and medical knowledge. In medical systems where prediction errors can have grave consequences, we propose leveraging self-consistency and ensemble methods to enhance error correction and error detection performance.

1 Introduction

With rapid advancements in Natural Language Processing (NLP), we are witnessing a surge of its applications across various domains, including healthcare. Incorporating NLP in clinical settings brings about a multitude of advantages. It enhances clinical decision-making through advanced support, making health information more accessible, streamlines documentation, and accelerates research initiatives (Hossain et al., 2023). These developments contribute to improved patient care, reduced healthcare costs, and alleviated physician burnout.

NLP for healthcare applications pose inherent challenges due to the need for medical expertise. However, advancements in LLMs trained on internet data including medical information have significantly enhanced their knowledge and reasoning capabilities, enabling them to tackle more complex problems in the healthcare domain involving text

processing and generation. Few recent applications in healthcare include information extraction, question answering, summarization, and translation, all while comprehending intricate medical knowledge (Nazi and Peng, 2023). Despite these advancements, safety and accuracy remain major concerns as training data may contain unreliable and misleading information that could have adverse effects if not handled appropriately. Nevertheless, the effective utilization of these LLMs has the potential to revolutionize healthcare and bring immense benefits to society (Clusmann et al., 2023).

In the healthcare industry, there is a need for automated systems capable of efficiently analyzing and interpreting clinical texts that improves patient’s safety, quality of care and costs. Processing the clinical texts presents a unique and significant challenge due to the complexities introduced by medical jargon, abbreviations, syntactic variations, and context-specific nuances. The MEDIQA-CORR shared task (Abacha et al., 2024), part of the ClinicalNLP 2024 workshop, seeks to address this issue of identifying and correcting (common sense) medical errors found in clinical notes.

The MEDIQA-CORR shared task involves three subtasks: detecting errors in clinical notes, identifying specific error sentences, and correcting those sentences. Our approach involves tackling all three subtasks simultaneously using a single prompt for LLMs invoking chain-of-thought. By doing this, we seek to assess the complex reasoning capabilities of LLMs in the clinical domain.

First, we analyze the dataset and curate a list of the most common types of errors in clinical notes. We then utilize this information to create task specific instructions for the LLM. We leverage contemporary LLMs using in-context learning (ICL) with a few-shot approach. Since publicly accessible LLMs are instruction-tuned models, considering the approach of directing them towards assessing the clinical note based on specific error

types maximizes the objective while only utilizing a few training examples.

By employing single-prompt ICL approaches to solve end-to-end tasks, we pave the way forward for building more complex healthcare applications using simple yet intuitive strategies leveraging the capabilities of advanced LLMs.

In addition to the task prompt, an LLM’s performance on a particular task is mainly influenced by two factors: training techniques and underlying training data. Consistency, which measures how frequently an LLM produces the same output given the same instructions, can be viewed as another dimension that can affect the quality of an LLM. Leveraging self-consistency can significantly improve the accuracy of an LLM, particularly for complex reasoning tasks (Wang et al., 2023). In healthcare datasets, where hallucinations in LLMs can occur more frequently due to training on factually unverified data, this could lead to serious problems. While self-consistency is one approach to obtaining more accurate results, LLM ensemble, which has not yet been fully explored, presents a promising opportunity. We validate the results of each LLM by using the output of other LLMs that are trained on different corpora. In our approach, we investigate both self-consistency and ensemble concepts.

The remainder of the paper includes related work, the MEDIQA-CORR task and dataset, our approach, results and conclusion.

2 Shared Task and Dataset

The shared task focuses on leveraging LLMs for the following three subtasks:

- **Binary Classification:** Detect if the text from a clinical note includes a medical error.
- **Span Identification:** Identify the text span (in the sentence) associated with the error, if an error exists.
- **Natural Language Generation:** Generate the corrected text, if an error exists.

2.1 Subtasks and Metrics

2.1.1 Error Detection

For each clinical text, the prediction is assigned a value of 1 when an error is detected, and 0 if no error is detected. Given the binary nature of this classification task, accuracy serves as the primary metric for performance evaluation.

2.1.2 Error Span Identification

Each clinical text comprises sentences associated with unique IDs. The subtask involves predicting the error ID, which is an integer between 0 and the highest sentence ID. If no error is detected, the prediction should be -1. The primary evaluation metric is accuracy, calculated based on all samples, including those with and without errors.

2.1.3 Correct Sentence Generation

If a model identifies an error sentence in the previous subtasks, it should also generate a corrected sentence as prediction for this subtask. Here, the full corrected sentence is evaluated against the ground truth sentence for measuring the performance. Clinical note generation tasks are challenging to evaluate automatically due to the large number of possible correct answers. Metric ensembles (Abacha et al., 2023) have been shown to outperform individual state-of-the-art measures, such as ROGUE for such tasks. The evaluation metric for this subtask is computed as an unweighted average of the following three scores:

- **ROUGE-1F** measures the similarity between system-generated and human-written texts by measuring the overlap of unigrams (Lin, 2004). It uses the F-1 score to assess the quality of the generated sentence.
- **BERTScore** leverages contextual word embeddings obtained from BERT models to assess the similarity between a candidate sentence and a reference sentence (Zhang* et al., 2020). In this context, it signifies the F-1 score of the semantic similarity performed using the DeBERTa XL model (He et al., 2021).
- **BLEURT-20** is a learned metric trained on human ratings that aims to better correlate with human judgments for measuring quality compared to traditional BLEU (Sellam et al., 2020).

2.2 Dataset

The train data consists of clinical texts from MS data while the valid and test data contains MS and UW collections. Each entry in the datasets includes a text, its ID and sentences as inputs. Table 1 shows the composition of the dataset.

3 Approach

We propose to tackle all subtasks concurrently within a single prompt for the following reasons:

Dataset Type	# Samples	% of Error Samples
MS Training	2189	55.69
MS Validation	574	55.57
UW Validation	160	50.00
MS + UW Test	925	51.35

Table 1: MEDIQA-CORR Dataset

- **Comprehensive Evaluation:** To enable performance evaluation on a complex task, rather than assessing them on isolated, simpler tasks. This provides a more holistic view of the LLMs’ capabilities.
- **Efficiency Optimization:** To minimize inference costs and developmental efforts by eliminating the need for multiple models (or) sequential processing. It streamlines the process, making it more efficient and cost-effective.
- **Ease of Adoption:** To alleviate the adoption burden in practical applications and facilitate seamless upgrades to more advanced LLMs amidst the rapid technological advancements.

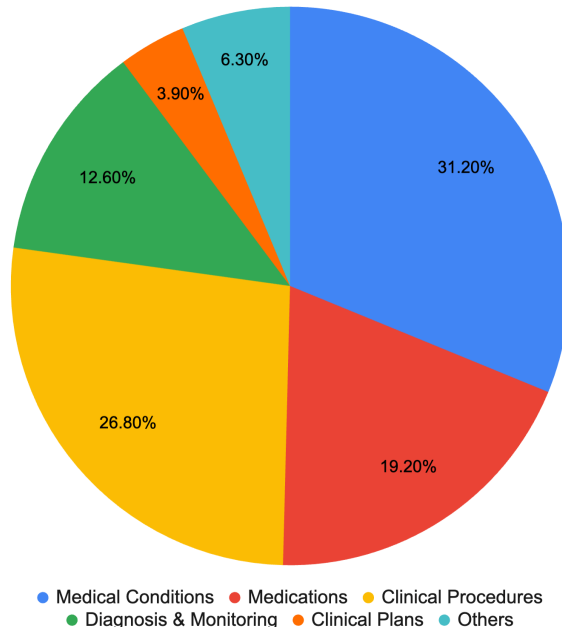
Publicly accessible LLMs are models that are fine-tuned to follow instructions, with the aim of performing user-defined instructions as accurately as possible. The success of the task then depends on the quality of the instructions provided and the model’s ability to follow them effectively. Our approach focuses on refining the instructions for the LLM to facilitate comprehensive learning of all subtasks using ICL. We initiate this process by analyzing error types in the dataset followed by curating the prompt and inferencing with different LLMs.

3.1 Error Analysis

In the MEDIQA-CORR task, the definition of an error is loosely defined and can be interpreted differently by humans or systems without examining the dataset. To address this, we perform error type classification in clinical texts by extracting error sentences and corresponding corrected sentences from the training data. We create an LLM prompting task to broadly categorize the entities modified from the error sentence to the corrected sentence within the clinical domain. We utilize GPT-3.5 for categorizing the errors and cluster these generated free-form categories into a manually defined set. This categorization results in the identification of

various error types as depicted in figure 1. Finally, we use the well-defined error categories for the rest of the task while handling "Others" category discreetly.

Figure 1: Error Types Extracted from Training Data



3.2 Prompt Curation

When prompted to identify errors in clinical texts without being specific, LLMs may introduce biases from their training data and flag non-critical errors adhering to their own standards of composing clinical notes. To mitigate this, we propose converting an abstract definition of a clinical note error into a concrete and approximate one by analyzing and categorizing the errors. Consequently, we expand the original task from identifying the error to include error classification, which facilitates chain-of-thought for the LLM. We conduct ablation studies to demonstrate the effectiveness of these techniques and present in section 4.

To illustrate the task more thoroughly, we incorporate reasoning within the task prompt for the LLM. Through this, we aim to provide more generalizable patterns for detecting and identifying errors. Additionally, it adds explainability to the LLM systems, which is crucial for real-world applications. Finally, we adopt a few-shot approach utilizing random samples from the training data to teach the LLM how to detect, identify, classify errors, provide reasoning and demonstrate the corrected text. We utilize the same samples for few-shot throughout the task as we

need to manually generate the additional fields such as error category and reason. The designed task prompt is provided as below. Note that this prompt is a tailored version for demonstration purposes. Finally, errors that fall into the "Others" category are processed as "No Error" as they are unimportant for this task.

Prompt Template

In this task, you will be given a clinical report presented as sentences while each sentence is associated with a sentence number. Now, you need to go through the report line by line and identify if there is any error in the sentence. The error can fall into one of the following main categories:

1. Medications
2. Medical Conditions, Virus or Bacteria
3. Reports, Diagnosis and Monitoring
4. Clinical Procedures and Treatments
5. Clinical Plans and Recommendations
6. Medical Devices
7. Others including clarity/improper usage of terminology

The error can be identified by looking at the entities present in each sentence and check if these entities fall into one of the aforementioned categories and validate if the entire report is correct with this entity. If there is an error, you need to output details as shown in the examples. Use all your medical knowledge and make the right judgements. Here are a few examples for your understanding:

/ Five random samples from training data with manually curated error category and reason */*

Example Clinical Report:

0 Mr. <Name> is admitted ..
 1 He has a surgical ..
 2 He is also being managed ..

Output:

```
{
  "Error Sentence ID": 1,
  "Error Category": "Medical Devices",
  "Reason": "The device should be .."
  "Corrected Sentence": "He has a surgical .."
}
```

...

...

Test Clinical Report:

0 A 45 year old woman ..
 1 She is experiencing ..
 2 She had prior examination ..

Output:

3.3 Model Selection

Using the designed prompt, we utilize the following LLMs for performing the task:

- **GPT-3.5:** A model from the OpenAI's generative pre-trained transformer (GPT) family that can understand as well as generate natural language or code (Ye et al., 2023).
- **GPT-4:** Latest model from the GPT family with broader general knowledge and advanced reasoning capabilities (OpenAI et al., 2024).
- **Claude-3 Opus:** Anthropic's largest model, released in Feb 2024, which sets new industry benchmarks across a wide range of cognitive tasks and outperforms its peers on most of the common evaluation benchmarks for AI systems (Anthropic and others, 2024).

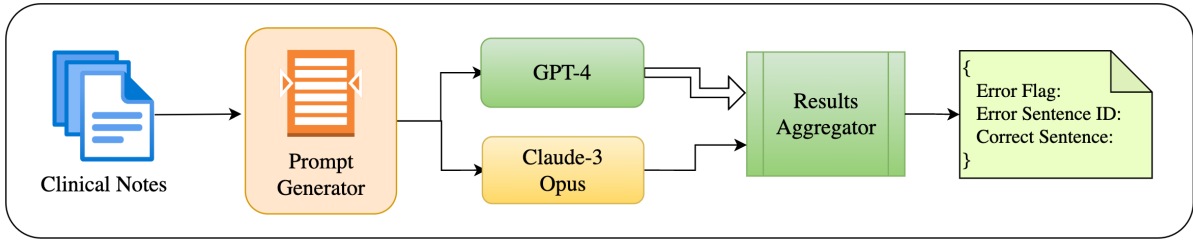
Due to its affordability, speed, and reliability, GPT-3.5 is an excellent choice for experimentation. As a result, we employed GPT-3.5 for error analysis and prompt design, reserving the more advanced GPT-4 and Claude-3 Opus models for the final test runs.

3.4 Enhancing Robustness

Due to the potential limitations such as hallucinations and inconsistent results, which can affect the quality of the LLMs, we investigate two concepts to improve performance on the subtasks: self-consistency and ensemble. The Claude-3 Opus model has slower speed, higher cost, and stricter token limits compared to GPT-4. Therefore, we utilize GPT-4 for self-consistency by generating four outputs per test sample and aggregate them, while only generating one output per test sample for Claude-3 Opus.

To enhance the quality of predictions, we combine the results from both models to predict the outputs for all three subtasks. Figure 2 provides a visual representation of the overall process. The results aggregator module validates and combines the outputs i.e. predicted error flag, predicted error sentence ID and corrected sentence, from GPT-4 and Claude-3 Opus models to generate the final error flag, error sentence ID and corrected sentence.

Figure 2: LLM Ensemble Approach for MEDIQA-CORR task



Model	Prompt	Task-1 Accuracy	Task-2 Accuracy
GPT-3.5	No error categories	48.75%	22.5%
GPT-3.5	Error categories	58.44%	38.55%
GPT-4	Error categories	63.07%	58.17%

Table 2: Performance improvement with error categorization in prompt using GPT-3.5 and assessing GPT-4 performance

4 Ablation Study

We begin by presenting the results obtained from incorporating error categorization into the final prompt, which demonstrates an improvement in performance on both error detection and text span identification tasks. In order to make the comparison, we utilize GPT-3.5 with prompts including and excluding error categorization. Additionally, we assess the performance of GPT-4 to ascertain the extent to which it surpasses GPT-3.5 for the finalized prompt. The results obtained using the combined MS and UW validation sets (during the development phase) are presented in Table 2.

The results indicate that by integrating error categorization which initiates an intermediate chain-of-thought, results in a significant performance boost of nearly 10% and 16% for Task-1 and Task-2, respectively. Additionally, GPT-4 outperforms GPT-3.5, confirming its enhanced reasoning capabilities. These advancements make GPT-4 a preferred choice for final test runs.

5 Results

We present the results of individual models first followed by incorporating self-consistency and ensembles on the test data. The performance of GPT-4 and Claude-3 Opus using the final prompt on all

Model	Task-1 Accuracy	Task-2 Accuracy	Task-3 Aggregate Score
GPT-4	62.05%	56.43%	0.6172
Claude-3 Opus	62.59%	58.48%	0.6669

Table 3: Comparison of GPT-4 and Claude-3 Results

Model	Task-1 Accuracy	Task-2 Accuracy	Task-3 Aggregate Score
GPT-4 with consistency (Majority=3/4)	62.91%	59.45%	0.6390
GPT-4 + Claude-3 Opus (Majority=4/4)	62.16%	60.86%	0.7865
GPT-4 + Claude-3 Opus (Majority=3/4)	63.78%	62.48%	0.7492

Table 4: Self-consistent GPT-4 and its ensemble with Claude-3 Opus Results

three subtasks is presented in Table 3.

Claude-3 Opus surpasses GPT-4 in error detection and significantly in error sentence identification. However, GPT-4 tends to be more verbose during error correction, leading to lower scores in metrics such as ROUGE, BERT, and BLEURT. Although Claude-3 Opus exhibits superior performance, its daily token limit, slower inference, and shorter test phase duration hinder its usability for self-consistency. Therefore, we employ GPT-4 to demonstrate the performance enhancement using self-consistency in large language models (LLMs). Additionally, we ensemble the self-consistent GPT-4 with Claude-3 Opus to showcase further improvement. Table 4 presents the results for all subtasks using the aforementioned methods:

The majority ratio x/y for GPT-4 results is a measure of how often the model produces the same result for a given input. For example, a majority ratio of $3/4$ means that at least three out of the four

results should be the same to qualify the predicted error sentence ID as correct. In the ensemble approach, a prediction is considered correct if the self-consistent GPT-4 result matches the Claude-3 Opus result. Otherwise, the prediction is considered "no error". To select the best corrected sentence from the ensemble, the ROGUE score is used to estimate the distance between each corrected sentence and the error sentence. The sentence with the highest ROGUE score is used for the evaluation because LLMs tend to generate verbose corrected sentences which decreases the aggregate scores. In our experiments, using a majority ratio of 3/4 for GPT-4 in the ensemble resulted in the best Task-1 and Task-2 performances. Using a majority ratio of 4/4 resulted in the best Task-3 aggregate score. Our best subtask-3 score is ranked 2nd among all participants in the competition, and our best subtask-2 performance is among the top 3 according to the reported scores (Ben Abacha et al., 2024).

6 Related Work

In recent years, there has been a surge of research exploring the potential of prompt engineering techniques with large language models (LLMs) in healthcare. These techniques have shown promising results in various healthcare tasks, often achieving state-of-the-art performances (Zhou et al., 2023), (He et al., 2023). One notable study, Med-Prompt, highlighted several research directions demonstrating the power of prompt exploration for LLMs (Nori et al., 2023). LLMs exhibited impressive knowledge and reasoning abilities, tackling various tasks effectively. These advancements showcase the potential of LLMs in healthcare, offering new opportunities for leveraging language models to address healthcare challenges.

Evaluating common sense reasoning is essential for computer systems, as it impacts language comprehension, communication reliability, and general task performance. SemEval-2020 ComVE aims to address general common sense questions and seeks logical justification for correct responses, assessing reasoning abilities. Pretrained language models naturally acquire common sense through training on vast word tokens obtained from the web (Wang et al., 2020). MEDIQA-CORR, specifically tailored to identify and correct errors in clinical notes, offers a valuable resource for evaluating pretrained LLMs in medical common sense reasoning. Inspired by prompt-based explorations, our research

also focuses on utilizing pretrained LLMs to assess reasoning capabilities in medical common sense scenarios.

7 Conclusion

Our research demonstrates that incorporating error categorization into the prompt enhances the performance of large language models (LLMs) in detecting, identifying, and classifying clinical note errors. By initiating an intermediate chain-of-thought, this approach facilitates better reasoning and aids the LLM in providing more accurate and explainable results. Furthermore, our findings suggest that self-consistency and ensembles can further enhance the robustness and performance of LLMs on these tasks. These advancements pave the way for the development of more reliable and interpretable AI systems for clinical documentation analysis, ultimately contributing to improved healthcare outcomes.

8 Limitations and Risks

While promising, our approach has limitations. LLMs trained on general data may lack specific medical knowledge, potentially leading to misinterpretations and inaccurate corrections. Despite mitigation efforts, the risk of hallucinations and inconsistencies in LLM outputs remains a concern. Additionally, the effectiveness of our approach relies heavily on prompt engineering, which requires expertise and may not be easily generalizable.

The black box nature of LLMs also presents challenges in terms of explainability and building trust in medical contexts. To mitigate these limitations, continuous improvements of training data, more robust evaluation metrics, and human oversight are crucial. Further research is needed to explore the full potential and limitations of LLMs in healthcare, ensuring their safe and responsible application for improved patient care.

9 Ethical Considerations

The use of LLMs for medical error detection and correction raises significant ethical concerns. Potential biases in training data and algorithms must be carefully mitigated to prevent propagating existing healthcare disparities and ensure fairness. Transparency in how prompts are designed and how the LLM reaches its decisions is vital for building trust and ensuring accountability. Additionally, robust data security and de-identification practices

are paramount for protecting sensitive patient information.

It is essential to remember that LLMs should serve as tools to augment the expertise of healthcare professionals, not replace them. Clear lines of responsibility, ongoing human oversight, and continuous research and collaboration are necessary to address these ethical challenges. This will ensure the responsible use of LLMs and their positive contribution to improved healthcare outcomes.

References

- Asma Ben Abacha, Wen wai Yim, Velvin Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2024. Medec: A benchmark for medical error detection and correction in clinical notes. *CoRR*.
- Asma Ben Abacha, Wen wai Yim, George Michalopoulos, and Thomas Lin. 2023. [An investigation of evaluation metrics for automated medical note generation](#). *ArXiv*, abs/2305.17364.
- Anthropic and others. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Online; accessed March 2024.
- Asma Ben Abacha, Wen wai Yim, Velvin Fu, Zhaoyi Sun, Fei Xia, and Meliha Yetisgen. 2024. Overview of the mediq-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.
- Jan Clusmann, Fiona R Kolbinger, et al. 2023. The future landscape of large language models in medicine.
- Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. [A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics](#). *Preprint*, arXiv:2310.05694.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). In *International Conference on Learning Representations*.
- Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, and Kathryn Turner. 2023. [Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review](#). *Computers in Biology and Medicine*, 155:106649.
- 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.
- Zabir Al Nazi and Wei Peng. 2023. [Large language models in healthcare and medical domain: A review](#). *Preprint*, arXiv:2401.06775.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoi-fung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. [Can generalist foundation models outcompete special-purpose tuning? case study in medicine](#). *Preprint*, arXiv:2311.16452.
- OpenAI, Josh Achiam, et al. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. 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.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Junjie Ye, Xuantang Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. [A comprehensive capability analysis of gpt-3 and gpt-3.5 series models](#). *Preprint*, arXiv:2303.10420.
- Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Yining Hua Junling Liu, Chengfeng Mao, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.