A Deep Dive into GPT-4's Data Mining Capabilities for Free-Text Spine Radiology Reports

Klaudia Szabó Ledenyi[1] [a], András Kicsi[1] [b] and László Vidács[1,2] [c]

[1] *Department of Software Engineering, University of Szeged, Szeged, Hungary*

[2] *HUN-REN-SZTE Research Group on Artificial Intelligence, Szeged, Hungary*

Keywords: Radiology, Clinical Reports, NLP, LLM, GPT, Prompt Engineering.

Abstract: The significant growth of large language models revolutionized the field of natural language processing. Recent advancements in large language models, particularly generative pretrained transformer (GPT) models, have shown advanced capabilities in natural language understanding and reasoning. These models typically interact with users through prompts rather than providing training data or fine-tuning, which can save a significant amount of time and resources. This paper presents a study evaluating GPT-4's performance in data mining from free-text spine radiology reports using a single prompt. The evaluation includes sentence classification, sentence-level sentiment analysis and two representative biomedical information extraction tasks: named entity recognition and relation extraction. Our research findings indicate that GPT-4 performs effectively in few-shot information extraction from radiology text, even without specific training for the clinical domain. This approach shows potential for more effective information extraction from free-text radiology reports compared to manual annotation.

1 INTRODUCTION

In recent years, there has been a significant growth in the field of natural language processing (NLP), one of the most researched fields of artificial intelligence. The application of NLP in biomedical research has significantly expanded due to the rapid development of NLP models. This is particularly evident in the field of radiology, where a large amount of radiologic reports are generated daily. Typically, these are free-form text reports, a format that includes a large amount of raw data. These data can be effectively extracted using various NLP techniques. Biomedical text mining encompasses various tasks on biomedical text data, including sentence classification, sentiment analysis, information extraction, text summarization, question answering, etc. Sentence classification involves classifying sentences into predefined groups. Sentence classification plays a role in organizing and understanding textual data. Sentiment analysis determines the sentiment of the text data, typically categorized as positive, negative, or neutral. Information extraction is the automated process of extracting structured information from unstructured data and transforming it into a more usable format. In radiology, the two main components of information extraction are the extraction of clinical entities and relations from radiology reports. These tasks are known as named entity recognition (NER) and relation extraction (RE). NER is the most important step in extracting relevant data, which aims to identify specific terms. RE focuses on identifying relations between the detected entities.

1.1 Technical Background

The transformer architecture, a type of neural network architecture that was introduced by (Vaswani et al., 2017), has revolutionized NLP. It uses attention mechanisms to identify relationships between words, effectively capturing long-range dependencies in input sequences. The architecture consist of an encoder-decoder structure, multiple layers of self-attention mechanisms, and feedforward networks. This formed the foundation for both pre-trained language models (PLMs) and large-sized PLMs, known as large language models. PLMs use the transformer architecture to train on a vast corpus of text data before fine-tuning

[a] https://orcid.org/0009-0001-4478-632X
[b] https://orcid.org/0000-0002-3144-9041
[c] https://orcid.org/0000-0002-0319-3915

Szabó Ledenyi, K., Kicsi, A. and Vidács, L.
A Deep Dive into GPT-4's Data Mining Capabilities for Free-Text Spine Radiology Reports.
DOI: 10.5220/0012765100003756
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 82-92
ISBN: 978-989-758-707-8; ISSN: 2184-285X
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

for specific downstream tasks. The currently popular large language models (LLMs) are transformer-based deep learning models too, that integrate the concept of pretrained models with an emphasis on billions of parameters. These models capture contextual information from a wide range of texts, enabling them to understand and generate human-like language. LLMs have demonstrated state-of-the-art (SOTA) results across a wide range of NLP tasks. At the end of 2022, OpenAI developed ChatGPT, a free-to-use AI system. ChatGPT has garnered considerable public attention because it does not require domain expertise for usage, unlike for instance Bidirectional Encoder Representations from Transformers-based (BERT) (Devlin et al., 2019) models which are currently still achieving SOTA results in several domains. The latest version of GPT (GPT-4) (OpenAI, 2023) has 100 trillion parameters, 570 times more than its predecessor, GPT-3. Because of the number of parameters, fine-tuning LLMs for specific tasks is impractical. The primary way to interact with AI systems for different tasks is prompting (Liu et al., 2021). Prompt engineering is a novel paradigm in NLP, which involves designing prompts to guide LLMs, such as GPT-4, to generate specific outputs for downstream tasks. A prompt is a text input that guides the model to generate the desired output. The design of the prompt significantly influences the model's performance on specific tasks. Few-shot and zero-shot prompting are specific techniques within prompt engineering. A well-designed prompt should provide clear guidance to the model, aiding in accomplishing tasks efficiently. Few-shot prompting involves providing demonstrations in the prompt to guide the model's behaviour to solve specific tasks. This technique can serve clear and explicit prompt inputs, guiding the model towards desired outputs. Zero-shot prompting means that the prompt used to interact with the model does not contain any examples or demonstrations. In-context learning (ICL) is a method used to solve complex tasks with LLMs (Brown et al., 2020). An ICL prompt is a task description presented in natural language text with a small number of task examples with the desired input-output pairs. These examples are also known as few-shot demonstrations. Prompt-based learning allows GPT to solve various NLP problems without updating its parameters, resulting in significant time and cost savings.

1.2 Research Objective

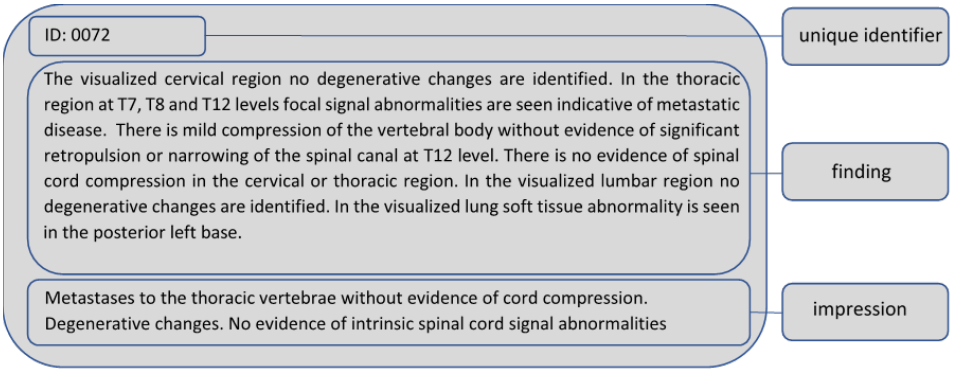
Our research objective is to evaluate the capabilities of GPT-4 in various biomedical text-mining tasks, focusing on spine radiology reports. We evaluate GPT-4's performance in sentence classification, sentence-level sentiment analysis, NER and RE tasks. We provide an optimized prompt that enables GPT-4 to perform these diverse tasks in a single step. Our final prompt, as well as the .csv format of one of the datasets (MTSamples) is available in our online appendix[1]. We note that while we strived to provide reproducible results, GPT-4's results are still not entirely deterministic. We also provided the resulting outputs produced by GPT-4 for this dataset. The other evaluated dataset's text cannot be published as per its terms of use, thus we omitted this dataset from the online appendix. Our experiments used version 0613 of GPT-4. Our evaluation result demonstrates that when given instructions and examples as a prompt, GPT-4 is capable of handling the examined tasks reasonably well.

2 RELATED WORK

The transformer architecture has reformed machine learning models for NLP. The two foundations of this architecture are the previously mentioned BERT and GPT models. Initially, these models were trained on general texts but later appeared domain-specific models, too. The objective of these new models was to outperform the general models in various domain-specific tasks. BioBERT (Lee et al., 2020) and BioGPT (Luo et al., 2022) models were trained on biomedical literature. MedBERT (Rasmy et al., 2021), a German medical natural language processing model, was fine-tuned using medical texts, clinical notes, research papers, and healthcare-related documents. RAD-BERT (Bressem et al., 2020) and RadiologyGPT (Susnjak, 2024) models, trained on radiology reports, outperformed previous, more general biomedical domain models, demonstrating a better understanding of radiology language.

These models have transformed the field of information extraction from unstructured text data. Current SOTA results were achieved by encoder-only models like BERT. These models are typically fine-tuned on annotated data before being used in NER and RE tasks. With the introduction of LLMs, like GPT-3, ChatGPT, and GPT-4, researchers started experimenting with these models to solve biomedical NER and RE tasks. Researchers (Agrawal et al., 2022) found that even without explicit training for the clinical domain, recent language models like InstructGPT and GPT-3 can effectively extract clinical information in a few-shot setting. Research has

[1] https://github.com/anonymusradiology/Data2024-submission-dataset



ID: 0072

The visualized cervical region no degenerative changes are identified. In the thoracic region at T7, T8 and T12 levels focal signal abnormalities are seen indicative of metastatic disease. There is mild compression of the vertebral body without evidence of significant retropulsion or narrowing of the spinal canal at T12 level. There is no evidence of spinal cord compression in the cervical or thoracic region. In the visualized lumbar region no degenerative changes are identified. In the visualized lung soft tissue abnormality is seen in the posterior left base.

Metastases to the thoracic vertebrae without evidence of cord compression. Degenerative changes. No evidence of intrinsic spinal cord signal abnormalities

Figure 1: Structure of a fictitious report with unique identifier, finding and impression sections.

been made to compare the few-shot performance of GPT-3's in-context learning with the fine-tuning of smaller, BERT-sized PLMs on NER and RE tasks (Jimenez Gutierrez et al., 2022). They used various benchmark datasets and found that GPT-3 could not outperform fine-tuned BERT-sized PLMs. Chen et al. (Chen et al., 2023b) employed prompt engineering to evaluate ChatGPT's performance on biomedical NER and RE tasks within the BLURB (Biomedical Language Understanding and Reasoning Benchmark) datasets. They used both zero-shot and few-shot approaches. Experiments have been (Chen et al., 2023a) established benchmarks for GPT-3.5 and GPT-4 in biomedical NER and RE at zero-shot and one-shot settings. They selected examples from BC5CDRchemical, NCBI-disease, ChemProt, and DDI datasets, using consistent prompts for evaluation. Their experiments showed that LLMs like ChatGPT, GPT-3.5, and GPT-4 performed less effectively than fine-tuned pretrained models on common NER and RE datasets. Researchers have also analyzed free-text CT reports on lung cancer (Fink et al., 2023). They compared ChatGPT with GPT-4 on the task of labelling oncologic phenotypes. They found that GPT-4 outperformed ChatGPT in extracting lesion parameters, identifying metastatic disease and generating correct labels for oncologic progression.

GPT models has greatly benefited sentiment analysis. Studies observed that the zero-shot performance of LLMs achieves comparable performance to the fine-tuned BERT model (Qihuang et al., 2023). There was a study using ChatGPT for sentiment analysis task, specifically examining its capacity to manage polarity shifts, open-domain scenarios, and sentiment inference issues (Wang et al., 2023). Researchers investigated the GPT's sentiment analysis capabilities in a prompt-based GPT, a finetuned GPT model, and GPT embedding classification (Kheiri and Karimi, 2023). They demonstrated a significant performance improvement compared to the SOTA models.

Medical professionals and researchers have utilized LLM technologies in a variety of text-generation tasks, leading to significant advancements. These tasks include question answering, automatic impression generation, summarization of medical documents, medical education, etc. Studies presented a comparative evaluation of GPT-4, GPT-3.5 and Flan-PaLM 540B (Nori et al., 2023). Their findings revealed that GPT-4 significantly outperforms GPT-3.5's and Flan-PaLM 540B's performance in generating answers for the US Medical Licensing Exam (USMLE) and on the MultiMedQA dataset. GPT-4's performance was evaluated in generating evidence-based impressions from radiology reports (Sun et al., 2023), they compared the results to human data. Impressions composed by radiologists outperformed GPT-4-generated ones in coherence, comprehensiveness, and factual consistency.

In addition, numerous articles deal with exploring LLMs' opportunities and pitfalls in revolutionizing radiology (Thapa and Adhikari, 2023; Akinci D'Antonoli et al., 2023; Liu et al., 2023)

We dive deep into the analysis of spine reports, examine them at several levels, which is not typical in the radiology research. While multiple research articles deal with tasks such as sentence classification, sentiment analysis, information extraction and summarisation, to our knowledge, none of the above studies has combined all these problems into a single process. While we deal with a thin slice of medical text processing, our analysis and evaluations require a high granularity of data on anatomy levels and specific nomenclature of spine conditions which are, to the best of our knowledge, not included in any benchmark that is currently available. Thus, our results were evaluated by a human linguist annotator.

3 METHOD

3.1 Data

This study utilized two databases, the publicly available MIMIC-III[2] (Johnson et al., 2016) database (which is accessible through a course) and the MT-Samples [3] corpus.

MIMIC-III is a large, de-identified, and publicly available collection of medical reports. The dataset is freely accessible, making it a valuable resource for researchers. The dataset contains approximately 60K medical reports from ICUs. It contains detailed information on medications, laboratory measurements, procedure codes, diagnostic codes, imaging reports, hospital length of stay, and more. After completing a web course offered by the National Institutes of Health (NIH), we gained access to the MIMIC-III database. The database consists of 26 tables. In this article, we work with radiology reports, so we utilize only the NOTEEVENTS table. This table contains unstructured text data, including nursing and physician notes, ECG reports, radiology reports, and discharge summaries. These free-text radiology reports contain the finding, the impression, the date of the study, and additional information depending on the study type. We filtered the spine radiology reports based on the technical area of the study, which is also provided. After this selection process, we structured these reports and kept only the findings and the impressions sections. We note that the MIMIC-III dataset’s website expressly forbids the dataset’s use with online GPT services, with the caveat of using them with one of two services with the appropriate settings. We used the Azure OpenAI service for our experiments, as suggested by the website.

The MTSamples database, a publicly accessible resource of transcribed medical reports, comprises transcribed medical transcription sample reports and examples across 40 medical specialities, such as radiology, neurology, and surgery. These are free text reports with headings, which change according to the speciality. For the purpose of our research, we created a small corpus from the MTSamples website focusing on spinal reports. Our final collection includes 53 spine reports’s findings and impression sections.

Figure 1 illustrates the structure of a report after its extraction from the table and preprocessing. This structure includes three main sections: a unique identifier generated by us, the findings section, and the impression section. The report in Figure 1 is an entirely fictitious example.

3.2 Tasks and Methods

Our project encompassed five tasks: sentence classification, sentence-level sentiment analysis, NER, RE and anatomy level determination.

The **sentence classification** task involved two labels, ”spinal” and ”extraspinal”. During the interpretation of radiologic studies, radiologists often identify abnormalities outside the region of interest. For instance, a lumbar spinal MRI can detect many extraspinal abnormalities, which may carry significant clinical implications and are crucial to recognize (Dilli et al., 2014). Common extraspinal regions include the renal area, uterus, kidney, prostate, and infrarenal aorta. This classification was essential as our focus was only on sentences related to spine anomalies. In the following steps, we focused on sentences labelled ”spinal”.

We determined the **sentiment**, categorizing them as either positive or negative. A sentence was labelled positive if the radiologist found no concerning issues and reported a normal or unremarkable state. Such sentences typically report normal results or reassuring information about a patient’s imaging studies. Negative sentiment indicates concerns or abnormalities in a patient’s imaging results, providing crucial insights for professionals into potential issues or areas requiring further investigation. It’s common to encounter sentences describing both normal and abnormal conditions in different anatomical areas. We labelled these sentences with negative sentiment because, from our point of view, it is better to draw attention to a normal condition than to ignore an abnormal state. Please note that our current nomenclature can potentially be misleading for medical professionals, as they usually use the ”positive” term to indicate the presence of an anomaly (such as positive for hernia), and ”negative” for its absence. We used these terms in the positive and negative sentiment sense, not in the medical sense. In the information extraction task, we worked only with the negative sentiment sentences.

In the **named entity recognition** task, we determined two entity types: anatomy and disorder. A term was considered anatomy, if it described a specific part of the human body in the spine area, like disc, neuroforamen. Disorders are various pathologies observed by the radiologist like hernia, and lysthesis. If there are both normal and abnormal conditions in a sentence, only negative conditions were labelled with the disorder label.

In our **relation extraction** task, we instructed the model to determine the relations between anatomy and disorder entities that were extracted by the pre-

[2] https://physionet.org/

[3] https://mtsamples.com/

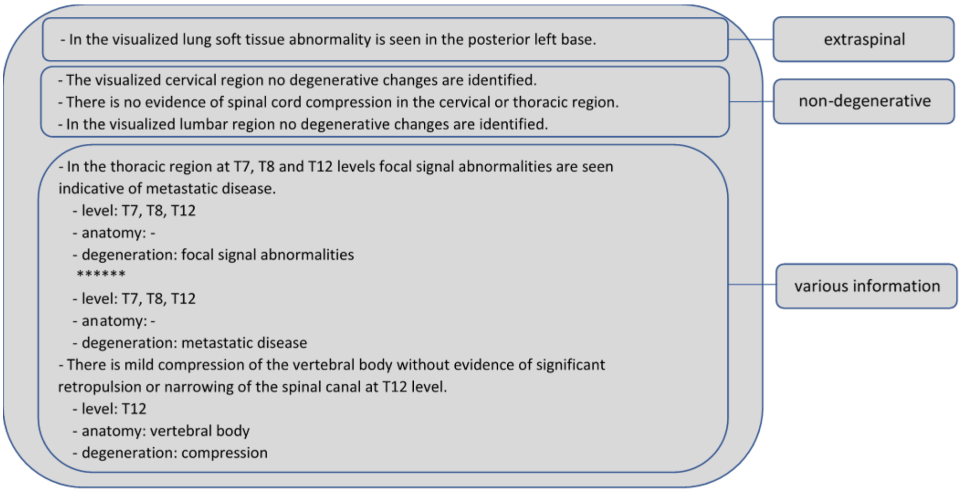

- In the visualized lung soft tissue abnormality is seen in the posterior left base.
extraspinal
- The visualized cervical region no degenerative changes are identified.
- There is no evidence of spinal cord compression in the cervical or thoracic region.
- In the visualized lumbar region no degenerative changes are identified.
non-degenerative
- In the thoracic region at T7, T8 and T12 levels focal signal abnormalities are seen indicative of metastatic disease.
- level: T7, T8, T12
- anatomy: -
- degeneration: focal signal abnormalities
******
- level: T7, T8, T12
- anatomy: -
- degeneration: metastatic disease
- There is mild compression of the vertebral body without evidence of significant retropulsion or narrowing of the spinal canal at T12 level.
- level: T12
- anatomy: vertebral body
- degeneration: compression
various information

Figure 2: Results of a finding processed with the proposed method, categorized into three sections: ”extraspinal” sentences, positive sentiment sentences, and negative sentiment sentences with various extracted information.

vious step. The anatomy-disorder relation connects the disorder with its anatomic locations.

In the case of anatomies, we determined the **sections and the levels** within them. We also determined the intervals (”L3-L5”) and various other referring phrases (”at this level” and ”the two lower discs at the cervical spine”). When there was no level or section in a sentence, we instructed the model to determine it considering the previous sentences, even if it had a positive sentiment. If the report did not contain any level, then we determined just the section of the spine.

Figure 2 illustrates the results of the fictitious finding seen in Figure 1 processed using the proposed method and our final prompt. The results can be categorized into three sections. The first section contains the output of the sentence classifier. This section includes sentences labelled as ”extraspinal”, which were omitted during the further processing. The middle section lists sentences expressing positive sentiments. Lastly, the lower box contains sentences describing negative sentiments, from which various information was extracted.

3.3 Prompt

Initially, we performed experiments to find the most effective prompt for use in subsequent analyses. The development of the final prompt was a multi-step process, with the help of a development set (5 new reports along the lines of the ones in the MTSamples dataset), which was separated from the available reports. We tested the different prompts, each incorporating all five radiology reports from the development set. Upon identifying errors, we tried to correct them by adjusting the prompt. We have repeatedly tested and modified the prompt to optimize output quality. During our experiments to find the optimal prompt, it was revealed that each stage of the prompt significantly influenced the performance and the efficiency of the entire process. The process of creating the final prompt is shown in Figure 3. Initially, we evaluated a starting prompt. Following this, we improved it cyclically, produced the output using GPT-4, and evaluated the outcome manually to measure the effectiveness of prompt variations. Meanwhile, we attempted to refine the prompt along the evaluation output.

During the development of the optimal prompt, we utilized a prompt engineering method, in-context learning. Each prompt included a task description and the expected output for two reports. In the initial attempts, we did the whole process in a single step, providing a brief task description and the expected output for two example reports. In the initial prompt the task description consisted of a few sentences: ”Your task is to analyze radiological reports about the spine. Focus solely on statements made about the spine, extract sentences indicating a negative status, and from these sentences, extract information on disorder, anatomy, and the level of the anatomy.” Upon evaluating the development reports with this prompt, we identified several issues: The results lacked sentences indicating negative status on the spine, included sentences with negated negative statuses such as ”no herniation is seen”, incorporated anatomy and descriptive features within the extracted disorder, contained anatomical terms of location within the extracted anatomy, as well as other, less general mistakes.

The initial approach to fixing the prompt involved formulating a set of rules to avoid the identified errors. We created a list containing ten specific rules, and directed the model to adhere to these during the processing of the reports. Throughout the rule development, we performed multiple tests on the incorrectly processed sentences identified in the prior evaluation. We observed that too many rules adversely impacted the output, leading to new errors. Despite our efforts, the list of rules did not bring the expected improvements; the output still included the previously noted errors.

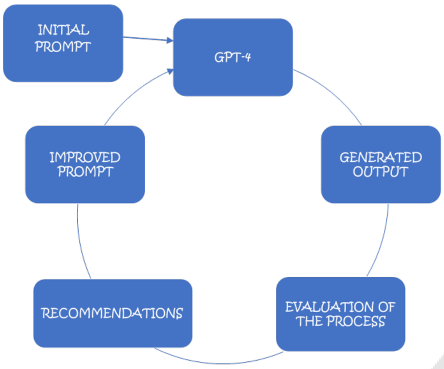


Figure 3: Creation of the final prompt involving iterative output generation using GPT-4, evaluation and improvement.

Our subsequent strategy for enhancing the prompt involved dividing the task into several subtasks, considering the errors we had identified. We experimented with various task divisions until we found the one that appeared optimal based on the development reports. This final prompt comprises five subtasks and a final step, which instructs the model to produce the final output, without displaying the outputs of the intermediate steps. In the first step, we instruct the model to generate a list of sentences from the input report. This step was crucial because it missed the sentence segmentation task, which forms the basis of the subsequent steps. The following step involves sentence classification and sentence-level sentiment analysis. Here, we guided the model to annotate each sentence with an ”extraspinal” label if it contained information only about a non-spine region, like the lung, aorta, and head, otherwise the sentence was labelled as ”spinal”. The sentiment analysis task then assigns a positive or negative sentiment to the remaining sentences. The third step works with sentences labelled as negative, performing NER and RE tasks. Initially, it identifies anatomies and disorders then establishes relations between these detected entities. The next step deals with these anatomy-disorder pairs and assigns the level of the anatomy to each one. If there is no level in the current sentence, the model attempts to determine it, considering the report’s previous sentences. The fifth step comprises rules that verify the absence of irrelevant information in the extracted anatomy or disorder entity. Almost each step includes specified rules for the current task and a template for the output. In the prompt we provided illustrative examples of complex rules. We developed two versions of the final prompt. In the first version, the prompt contained the expected output to each step. The second version of the final prompt excluded the output of the inner phases, only featuring the expected output of the entire process. After testing both versions on the development dataset, we found no significant difference between the two prompts. Consequently, we decided to proceed with the prompt, which only included the final output in the example.

We utilized GPT-4 for the project’s development. We created a Python script to repeatedly evaluate the prompt on different development examples. Within this script, we adjusted the temperature and top_p values to zero, aiming for the most probable word to achieve consistent and deterministic output. We also set the seed value to zero, further enhancing the deterministic response. This adjustment ensures nearly identical output for a given input and seed combination. Despite our attempts to make the model deterministic, some outputs remained unpredictable in certain situations.

4 RESULTS

The evaluation process involved 100 reports from the MIMIC-III database and 53 reports from the MTSamples collection. We selected these reports from the previously structured ones (structured reports have unique identifier, finding and impression sections), focusing solely on the findings section. We measured the response time for each report and analyzed the output. The response time mostly depends on the number of output characters generated by the model. In our case, the number of output characters was influenced by several factors, such as the length of the input report and the presence of extraspinal and positive sentences, which required no further processing. The output length was the longest when the majority of the findings included degenerative statements about the spine, leading to an increase in processing time. The average processing time for the 100 MIMIC-III reports was 35 seconds, the average input contained approximately 685 characters per report and the average output contained approximately 1200 characters per report. Figure 4 illustrates the correlation between processing time and the total number of input characters and generated output characters.

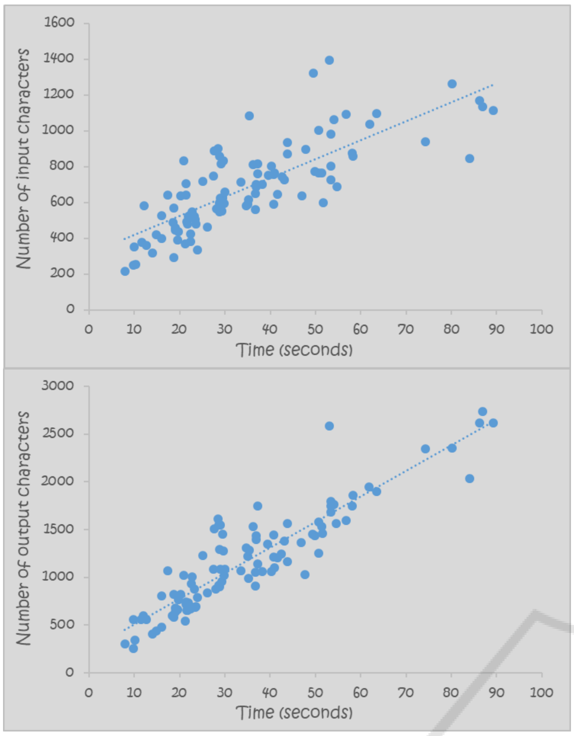


Figure 4: Correlation between processing time and the number of input (top diagram) and output (bottom diagram) characters.

We evaluated the final prompt using a complex evaluation process. We generated the output for each report, then a linguist annotator who had previously worked on tasks related to radiological spinal reports performed the manual evaluation. Finally, one of the authors with years of experience in processing spine radiology reports reviewed the evaluation results once more. The model's performance was evaluated using accuracy, precision, recall, and F1-value, which are based on the concepts of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. Accuracy measures the proportion of correctly classified instances among all instances. Precision describes the proportion of TP predictions made by the model. Recall indicates the number of positives the model rightly predicted. The F1-value is the harmonic mean of precision and recall. Table 1 provides a comprehensive overview with these metrics for the evaluated tasks in the study. The values of accuracy, precision, recall, and F1-value are presented for each task separately.

For the sentence classification (SC), sentence analysis (SA), and anatomy level determination tasks, the manual annotator had to provide a number representing the accurately labelled sentences and correctly determined levels. In the case of anatomy, degeneration, and relation testing, FP and FN values were determined on entity level. The annotator increased the FP value if the model labelled a non-disorder phrase as a disorder. Similarly, if the model did not identify a negative sentiment disorder in a sentence, the FN value was increased. To better distinguish the different types of errors, we divided the report-processing workflow into four categories. The first level of Figure 5 represents sentence segmentation. However, we did not evaluate this capability of GPT-4; it is included only to increase the figure's comprehensibility. Each category had specific rules for calculating different measures. As can be seen in Figure 5, each category is built upon the outcomes of the previous one, treating errors from earlier stages as valid inputs for the subsequent layer. The categories included:

- Sentence classification: We analyzed a list of sentences derived from the original report. We checked whether the sentences were correctly categorized into the ”spinal” and ”extraspine” categories.
- Sentiment analysis: This step involved a detailed examination of every sentence labelled as ”spinal” in the previous step. We checked if each sentence was correctly labelled as ’positive’ or ’negative’ based on the presence of disorder within the sentence.
- Information extraction: This step involved analyzing every sentence labelled as ’negative’ in the sentiment analysis step. We evaluated the labels for anatomy and disorder entities and the relationships between the identified entities. Note that this is the point where entities are considered rather than sentences.
- Level extraction: This step involved analyzing the anatomy-disorder pairs generated in the information extraction step. Not all sentences included the level, so the annotator and the model were required to infer this information from the context.

In the context of anatomy detection, there were 7 instances where the sentence did not contain any anatomical reference. However, the model inferred an anatomical location based on previous sentences and basic knowledge. We classified these instances as FP, given that the objective of the named entity recognition task does not include inferring missing entities, and the model was also not instructed to provide these.

In the case of detected anatomies, there were 46 occurrences in the MTSamples and 10 occurrences in the MIMIC-III reports where the entity included supplementary information beyond the extracted anatomy. This additional information often

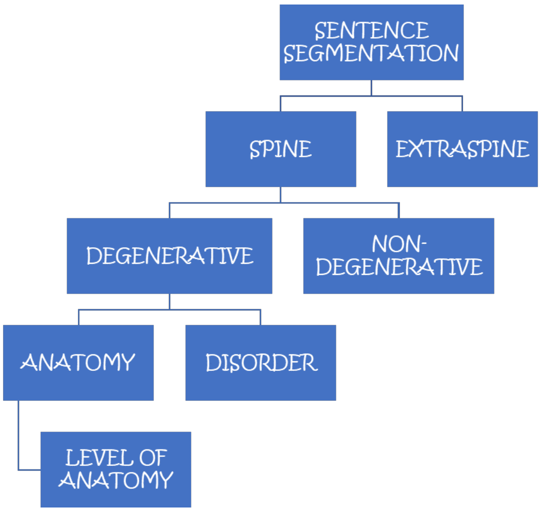


Figure 5: The report-processing workflow is hierarchically divided into five categories to distinguish various types of errors. Each category builds on the results of the previous one.

specified anatomies, using terms like ”posterior”, ”superior”, and ”distal”. In the disorder detection, there were 73 identified entities within MTSamples and 62 instances in the MIMIC-III reports that contained additional information. These terms were used to specify the location of the disorder, using words such as ”anterior”, ”posterior”, and ”congenital”. The extra information also described the characteristics of disorders, like ”small”, ”‘comminuted”, ”partially”, and ”non-displaced”. During the evaluation, we did not classify these instances as errors. This is because, in each case, the model accurately determined the anatomy and disorder terms, along.

5 DISCUSSION

In our study, we evaluated 100 reports from the publicly available MIMIC-III database and 53 reports from our MTSamples set. We aimed to select diverse and representative samples that provide a solid basis for our analyses.

As can be seen in the results section, based on the evaluated reports, the total F1-value of our approach is 98.18% and 98.34% on the two different databases. This F1-value includes the sentence classification, sentence analysis, anatomy and disorder detection, relation extraction and the anatomy level determination tasks. As the complexity of the task increases, there is a slight decrease in result quality. The task of annotating spinal and extraspinal sentences is relatively simple, requiring only a decision on whether the anatomy includes the spine. Sentences referring to extraspinal only mention the body part, such as the lung, without going into its detailed anatomy. The next task, sentiment analysis, remains relatively straightforward. The process takes into account two things. The sentiment is positive if the sentence mentions only positive aspects related to the given anatomy, such as ”normal” or ”patent”, and the mention of disorder in negative sentences, for instance, ”no herniation”. Otherwise, the sentiment is negative. Thus, this task achieved a high level of accuracy. The task of detecting anatomies and disorders is more difficult than the tasks of sentence classification, as evidenced by the results tables. The F1-value of anatomy detection is somewhat higher than the detection of disorders. This can be because anatomical terminologies are more standardized and well-defined across different sources. This consistency enhances the performance of recognition algorithms. Anatomical terms may occur more frequently in text data compared to disorder terms. Disorders can vary widely in representation of their complexity and linguistic expressions, making it harder for models to identify them accurately. The task of relation extraction in our methodology can be relatively simple, so it also reached a high F1-value. The model is designed to link the identified anatomical and degenerative entities in a sentence. In most cases, this is not complex because, the findings’ sentences have various statements about a particular anatomy. Therefore, a clause or sentence typically includes one anatomical and one degenerative entity, making the link between them straightforward.

Each MTSamples report had an average of 11 sentences, with around 5.5% discussing an extraspinal region. About 61% of these spinal sentences were negative, talking about concerns or abnormalities in a patient’s imaging result. The rest of the sentences were positive, reporting a normal or unremarkable state. In comparison, the MIMIC-III reports averaged 8.5 sentences per report in the evaluated 100 reports. Approximately 20% of the sentences talked about an extraspinal region. We also checked the sentiment of these spinal sentences. Surprisingly, just more than half of them, about 51%, were negative, talking about concerns or abnormalities in a patient’s imaging result. The rest of the sentences were positive. A summary of these analyses can be seen in Figure 6.

The other aspect we analysed in the reports is the contained number of anatomy and disorder entities, as well as the relations between them. Our approach found about 515 and 430 mentions related to various parts of the spine and 640 and 550 instances describing different disorders on the spine in the 53 MTSamples and 100 MIMIC-III reports, re-

Table 1: Evaluation results on the various examined subtask of the extraction.

		Accuracy (%)	Precision (%)	Recall (%)	F1-value (%)
Sentence classification	MTSamples	98.65	99.11	99.46	99.28
	MIMIC-III	99.42	99.43	99.86	99.64
Sentiment analysis	MTSamples	98.09	97.43	99.42	98.41
	MIMIC-III	98.00	96.48	99.72	98.07
Named entity recognition	MTSamples	89.89	97.78	91.77	94.68
	MIMIC-III	91.67	97.50	93.91	95.66
Relation extraction	MTSamples	98.22	99.67	98.54	99.10
	MIMIC-III	98.24	99.11	99.11	99.11
Level of anatomy	MTSamples	98.88	98.88	100.00	99.44
	MIMIC-III	98.42	98.42	100.00	99.20

spectively. Additionally, our analysis uncovered 615 and 560 connections between the mentioned anatomy and disorder entities in these reports. This suggests that, on average, each MTSamples report contains approximately 12 negative sentiment disorders and 10 anatomies with disorder and each MIMIC-III report contains approximately 5.5 negative sentiment disorders and 4.3 anatomies with disorder. The method's output helps to measure how often these anatomy and disorder entities appear and how they are linked, leading to a better understanding of the medical information in the reports.

From these analyses, it can be seen that there is a significant difference between the two used databases. Comparing the databases, it is clear that the MTSamples reports contain more sentences, with a majority referring to the spine region. However, only 40% of these sentences do not contain degeneration. In contrast, the MIMIC-III reports have 2.5 fewer sentences per report, and half of the sentences describe normal conditions with positive sentiments.

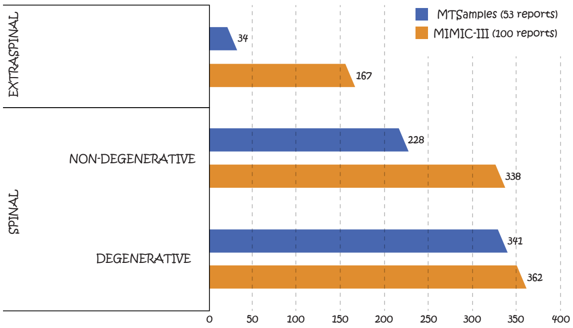


Figure 6: Number of spinal (degenerative, non-degeneraive) and extraspinal sentences in the 53 MTSamples and 100 MIMIC-III reports.

In processing radiology reports, the application of PLMs and LLMs can yield both positive and negative effects. The problem of processing spine radiology reports can also be solved without relying on GPT-4. With PLMs, it could be solved by employing, for instance, four separate BERT-based models, which is a more task-specific approach. A BERT model fine-tuned for sentence classification tasks would be employed to categorize sentences into predefined classes. Another BERT model, fine-tuned for sentiment analysis, would be utilized to determine positive or negative sentiment of sentences. A dedicated BERT model fine-tuned for NER tasks would be applied to recognize and classify anatomy and disorder entities within the text. The last fine-tuned model would be applied to detect relations between the detected entities. The fine-tuning process relies on annotated task-specific data. In the medical field, finding freely available, large and representative annotated datasets can be challenging, and the annotation of these reports is time-consuming and expensive. In addition fine-tuning models requires substantial computational resources, including high-performance GPUs and significant memory. Despite the disadvantages task-specific BERT models might provide clearer interpretability for individual tasks, so it could be easier to further develop the models by analyzing the errors in their outputs. Advantages include that, unlike LLMs, working with such on-premise, fine-tuned models typically involves less privacy concerns when dealing with sensitive information.

Compared to PLMs, GPT's few-shot capability can effectively guide these models to generate the desired outputs, eliminating the need for extensive fine-tuning. This approach can save computational resources. Moreover, GPT models typically require less preprocessing of the input text. This can simplify the data preparation process, especially for tasks involving free-form text like radiology reports. The method's disadvantages include that utilizing cloud-based GPT models may raise privacy concerns, especially when dealing with sensitive medical data such as radiology reports. The models can also have a lack of interpretability and results are harder to reproduce, making it challenging to understand and improve the model's decision-making process. The outputs are also less controllable compared to models like BERT, which is a drawback, particularly in situations where precise output is essential. Similarly to BERT, it is possible to fine-tune a GPT-based model for specific tasks, but this can also require extensive computational resources and data.

Our study has limitations. First, the evaluation was implemented on a relatively small sample size of just 153 spine reports. This may not accurately reflect the processing capabilities of GPT-4 for spine radiology reports in real-world clinical settings. Second, using only the MIMIC-III and MTSamples database restricts the diversity of data. Different databases may include different language uses and terminologies, which might affect the model's performance. Third, each report was evaluated once, yet the output of the model is not deterministic. Repeated evaluations could provide a more accurate assessment.

6 CONCLUSION

In our study, we utilized GPT-4 for processing radiology reports, completing the entire task with a single prompt. We classified the sentences, determined the sentiment of each spine-related sentence and extracted the level of anatomy, anatomy and disorder triplets. Finally, we evaluated the method on two different databases, 100 radiology spine reports from the MIMIC-III database and 53 radiology spine reports from the MTSamples collection. These results highlight how prompt-learning large language models can find information from free-text radiology reports without needing expert knowledge or task-specific fine-tuning. According to our findings, the GPT-4 model performed with over 91% accuracy and F-score values in each of our five subtasks of information extraction of the reports. Our MTSamples input and output data, as well as our final prompt are available in our online appendix.

ACKNOWLEDGEMENTS

The research presented in this paper was supported in part by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. The national project TKP2021-NVA-09 also supported this work. Project no TKP2021-NVA-09 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

REFERENCES

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. (2022). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Akinci D'Antonoli, T., Stanzione, A., Blüthgen, C., Vernuccio, F., Ugga, L., Klontzas, M., Cuocolo, R., Cannella, R., and Koçak, B. (2023). Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagnostic and Interventional Radiology*.

Bressem, K. K., Adams, L. C., Gaudin, R. A., Tröltzsch, D., Hamm, B., Makowski, M. R., Schüle, C.-Y., Vahldiek, J. L., and Niehues, S. M. (2020). Highly accurate classification of chest radiographic reports using a deep learning natural language model pretrained on 3.8 million text reports. *Bioinformatics*, 26(21):5255–5261.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Chen, Q., Du, J., Hu, Y., Keloth, V., Peng, X., Raja, K., Zhang, R., Lu, Z., and Qi, W. (2023a). Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326*.

Chen, Q., Sun, H., Liu, H., Jiang, Y., Ran, T., Jin, X., Xiao, X., Lin, Z., Chen, H., and Niu, Z. (2023b). An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics*, 39(9).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dilli, A., Ayaz, U. Y., Turanlı, S., Saltas, H., Karabacak, O. R., Damar, C., and Hekimoglu, B. (2014). Clinical research incidental extraspinal findings on magnetic resonance imaging of intervertebral discs. *Archives of Medical Science*, 10(4):757–763.

Fink, M. A., Bischoff, A., Fink, C. A., Moll, M., Kroschke, J., Dulz, L., Heussel, C. P., Kauczor, H.-U., and Weber, T. F. (2023). Potential of chatgpt and gpt-4 for data mining of free-text ct reports on lung cancer. *Radiology*, 308 3.

Jimenez Gutierrez, B., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., and Su, Y. (2022). Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.

Johnson, A. E. W., Pollard, T. J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M. M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

Kheiri, K. and Karimi, H. (2023). Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liu, Z., Zhong, T., Li, Y., Zhang, Y., Pan, Y., Zhao, Z., Dong, P., Cao, C., Liu, Y., Shu, P., Wei, Y., Wu, Z., Ma, C., Wang, J., Wang, S., Zhou, M., Jiang, Z., Li, C., Xu, S., and Liu, T. (2023). Evaluating large language models for radiology natural language processing. *arXiv preprint arXiv:2307.13693*.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. (2022). Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23.

Nori, H., King, N., McKinney, S., Carignan, D., and Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Qihuang, Z., Ding, L., Liu, J., Du, B., and Tao, D. (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4:86.

Sun, Z., Ong, H., Kennedy, P., Tang, L., Chen, S., Elias, J., Lucas, E., Shih, G., and Peng, Y. (2023). Evaluating gpt-4 on impressions generation in radiology reports. *Radiology*, 307(5).

Susnjak, T. (2024). *Applying BERT and ChatGPT for Sentiment Analysis of Lyme Disease in Scientific Literature*, pages 173–183.

Thapa, S. and Adhikari, S. (2023). Chatgpt, bard, and large language models for biomedical research: Opportunities and pitfalls. *Annals of Biomedical Engineering*, 51.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Wang, Z., Xie, Q., Ding, Z., Feng, Y., and Xia, R. (2023). Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.