Paragraph Retrieval for Enhanced Question Answering in Clinical Documents

Vojtěch Lanz and **Pavel Pecina**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{lanz,pecina}@ufal.mff.cuni.cz

Abstract

Healthcare professionals often manually extract information from large clinical documents to address patient-related questions. The use of Natural Language Processing (NLP) techniques, particularly Question Answering (QA) models, is a promising direction for improving the efficiency of this process. However, document-level QA from large documents is often impractical or even infeasible (for model training and inference). In this work, we solve the document-level QA from clinical reports in a two-step approach: first, the entire report is split into segments and for a given question the most relevant segment is predicted by a NLP model; second, a QA model is applied to the question and the retrieved segment as context. We investigate the effectiveness of heading-based and naive paragraph segmentation approaches for various paragraph lengths on two subsets of the emrQA dataset (Pampari et al., 2018). Our experiments reveal that an average paragraph length used as a parameter for the segmentation has no significant effect on performance during the whole document-level QA process. That means experiments focusing on segmentation into shorter paragraphs perform similarly to those focusing on entire unsegmented reports. Surprisingly, naive uniform segmentation is sufficient even though it is not based on prior knowledge of the clinical document's characteristics.

1 Introduction

Healthcare professionals spend a lot of time going through extensive clinical documents, such as discharge summaries, to find specific answers to questions about their patients (Demner-Fushman et al., 2009). This process could be aided by Question Answering (QA) models, that search for substrings in the text of a clinical document to provide an evidence for a given question (Pampari et al., 2018).

Currently, encoder-based language models demonstrate strong performance in solving QA tasks (Lan et al., 2020; Zhang et al., 2020), even when we are looking for substrings in a multi-paragraph clinical context (Yue et al., 2020). However, the training process and inference of large language models (LLMs) on document-level QA require significant computational resources that are not always available. In addition, encoder-based and decoder-based models face difficulties in understanding and processing longer documents (Liu et al., 2023). A possible solution might be working with segments (paragraphs) of the document rather than full text.

To achieve this, we must first segment the document into paragraphs, then identify the relevant paragraph for a given question, and then apply an QA model only to the selected paragraph as the context instead of the full document text. This alone significantly facilitates healthcare professionals' work in finding answers in clinical texts, which is another reason why it is worth addressing the paragraph retrieval issue.

Clinical texts often lack structure (Richter-Pechanski et al., 2024; Gallego Donoso and Veredas, 2023) and contain information that is not expressed in natural language (Pampari et al., 2018). Moreover, each clinical text, authored by distinct doctors from various hospitals and even different countries, is arranged uniquely. Therefore, the task of segmenting a document into natural language paragraphs is inherently non-trivial. However, the question arises: is it necessary to segment clinical text into such structured paragraphs? Will a naive uniform segmentation without knowledge of the text itself have a similar performance?

Our work addresses QA on differently-sized paragraphs of clinical documents. First, given clinical document paragraphs and a given question, retrieve the most relevant paragraph. In the second step, we perform QA on that paragraph. In addi-

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 580–590
August 16, 2024. ©2024 Association for Computational Linguistics

tion, we investigate the potential performance of the model on the QA task if the relevant paragraph is always predicted correctly.

We work with two subsets of the emrQA dataset: *Medication* and *Relations* (Pampari et al., 2018). We propose a heading-based segmentation into sections regarding different average paragraph lengths over all training clinical documents. We analyze the optimal average paragraph length to achieve the best performance and ensure that we preserve the context while keeping the paragraph as concise as possible. We then compare the performance of the encoding-based models on these segmentations with their performance on a naive segmentation approach. Finally, we demonstrate how LLMs perform under the same training conditions as encoder-based models. Our main contributions are the following:

- We demonstrate the feasibility of simplifying the document-level Question Answering (QA) challenge into a two-step task combining paragraph retrieval and paragraph-level QA.
- We propose a novel heading-based paragraph segmentation approach of emrQA data and compare its performance with naive segmentation.
- We present a comparative analysis of encoder-based and decoder-based models on the QA task, thus enriching the discussion on the optimal choice of architecture.

2 Related Work

The problem of question answering encompasses several different subtypes. One of them is to return an answer for a given question without any context (Berant et al., 2013). Another subtype involves text comprehension. For a given question and some context (such as a document or a paragraph), the task is to answer the question based on the content of the context, but the actual formulation of an answer is not restricted (Joshi et al., 2017). In our work, however, we focus on finding substrings in a given context that serve as both evidence and answer to a given question.

A significant resource in this field is the SQuAD dataset (Rajpurkar et al., 2016), containing questions, context paragraphs based on Wikipedia articles, and answer substrings. The dataset has been used to train and compare various neural methods, including encoder-based and decoder-based architectures (Lan et al., 2020; Zhang et al., 2020; Schmidt et al., 2024). This dataset was later extended into SQuAD v.2 (Rajpurkar et al., 2018), which also includes questions and corresponding paragraphs that do not contain an answer for a given question. As an alternative to this dataset in the clinical domain, the emrQA dataset (Pampari et al., 2018) was published. The emrQA dataset contains synthetically generated questions and substring answers for clinical reports from the n2c2 dataset (previously called i2b2). The emrQA consists of 5 subsets: *Medication*, *Relations*, *Heart disease*, *Obesity*, and *Smoking*, each focusing on different aspects and different complexity. From the emrQA dataset, the emrqa-msquad dataset (Eladio and Wu, 2024) was derived by summarizing clinical reports into single paragraphs as contexts and providing new manual annotated substring answers. However, this process removes the naturalness of clinical notes written by healthcare professionals. There is also the QA reading comprehension dataset in the medical scientific domain, which is BioASQ (Tsatsaronis et al., 2015). The dataset includes instances consisting of a question, ideal answer, PubMed medical article abstracts containing the answer, and the substring answers of all such related article abstracts.

In our work, we exploit the emrQA dataset (Pampari et al., 2018). Yue et al. (2020) analyzed the two largest subsets from the emrQA dataset in detail: *Medication* and *Relations*. They preprocessed and filtered these two subsets and trained encoder-based models, such as BERT-base (Devlin et al., 2018), BioBERT (Lee et al., 2019), and Clinical-BERT (Alsentzer et al., 2019), and then compared their performance. However, developments in the field have introduced other medically pre-trained encoder-based models, such as MedCPT (Jin et al., 2023), designed specifically for biomedical information retrieval, or BioLORD (Remy et al., 2024). Although the emrQA dataset authors perceive the analysis provided by Yue et al. (2020) as misleading due to the use of only 2 out of 5 subsets, for the purposes of our work, these two subsets with the same preprocessing and filtering are equally suitable. Therefore, our work indirectly follows up on the analysis conducted by Yue et al. (2020).

Another type of QA task involves multiple-choice questions. In the field of medicine, there is a PubMedQA dataset (Jin et al., 2019), which contains questions related to PubMed article abstracts. Furthermore, exam-like multiple-choice question

	Medication	Relations
Number of questions	222,957	904,590
Number of reports	262	426

Table 1: Basic statistics of both *Medication* and *Relations* subsets. Each question contains at least one answer present in the report.

datasets such as MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), MMLU (Hendrycks et al., 2021) were published. These datasets have been used as benchmarks for LLMs, such as MediTron (Chen et al., 2023) and BioMistral-7B (Labrak et al., 2024), which are open-source LLMs pre-trained on medical data. In addition to medical scientific and exam-like multiple-choice question datasets, Richter-Pechanski et al. (2024) focused on multiple-choice questions on German doctors' letters.

3 Setup

We solve the task of document-level QA on the emrQA dataset by a two-step method combining paragraph retrieval and paragraph-level QA. We analyze performance of the two tasks in combination and also separately. We follow the work of Yue et al. (2020) focusing on the *Medication* and *Relations* subsets only and applying the same data preprocessing. Table 1 shows the basic statistics of the two subsets. However, our results are not directly comparable due to the different random split into training, development, and test sets.

Throughout the rest of our study, we refer to these definitions:

- **Paragraph Retrieval (PR)**: Given a question and n paragraphs (i.e. report segmented into n paragraphs) as input, the objective is to rank the paragraphs based on the confidence that they contain relevant information. The task is evaluated using precision at top 1 (P@1), precision at top 2 (P@2), and precision at top 3 (P@3) paragraphs. Ground truth relevant paragraphs are those containing an answer evidence to a given question defined in the emrQA dataset.

- **Oracle Paragraph-driven Question Answering (Oracle-QA)**: Given a question and an Oracle paragraph (guaranteed to contain the answer) the objective is to identify and extract a minimal substring from the paragraph that precisely addresses or answers the given question. The task is evaluated using the official SQuAD metrics (Rajpurkar et al., 2016), which are *F1* and *Exact Match* scores. We compare our predictions with the original form of the testing dataset generated by the filtration of Yue et al. (2020), i.e., with the dataset before the segmentation process.

- **Paragraph Retrieval–Question Answering (PR-QA)**: Given a question and n paragraphs (i.e. report segmented into n paragraphs), the goal is to identify and extract a substring from one of the paragraphs that precisely addresses or answers a given question. Evaluation of the task is based on the *F1* and *Exact Match* scores the same way as in the Oracle-QA task.

Yue et al. (2020) concluded that it is sufficient to use only 20% and 5% of training data to train models of the *Medication* and *Relations* subsets, respectively. Since a larger amount of training data has no effect, we use the same ratio of data samples for training. Their data instances consisted of triples of document+question+answer where the answer was guaranteed to be present in the document. Our data instances were generated as triples paragraph+question+answer where the answer was also guaranteed to be present in the paragraph and pairs paragraph+question where the corresponding answer was not present in the paragraph. For each question in a sampled training subset of a given report, we randomly select a paragraph from the same report that does not contain an answer. Thus, we have a balanced dataset where the number of paragraphs containing an answer matches the number of paragraphs without them.

In our experiments, we train the ClinicalBERT (Alsentzer et al., 2019) and BERT-base (Devlin et al., 2018) models, just as Yue et al. (2020) did. Additionally, we measure the performance of the MedCPT Article Encoder (Jin et al., 2023) model. Since we are working with a balanced dataset, it is necessary to specify how to handle cases where the answer is missing in the context. If there is no response in the dataset sample, the model is trained to predict the CLS token as a response prediction. During the inference, we apply the softmax function to the output logits and use its negative value as confidence that the paragraph contains the answer. Then, for a given question and segmented report into paragraphs, the model solves the QA problem for all paragraphs (we evaluate the Oracle-QA task using the ground truth relevant paragraph), ranks

```
PAST SURGICAL HISTORY: Notable for
the above , as well as debridements
...
DISCHARGE MEDICATIONS: Vancomycin
1250 mg IV q d , Ofloxacin 200
...
LABORATORY DATA :
White count 12.6 , hematocrit 28.9
...
PHYSICAL EXAMINATION :
On admission vital signs were
...
```

Figure 1: Example of report paragraph headings of both *Medication* and *Relations* subsets.

all results by confidence, and selects the substring of the paragraph with the highest confidence as the final answer prediction (and then the PR and PR-QA tasks are evaluated as well).

4 Document Segmentation

Our goal is to design a method for segmenting reports into natural language paragraphs that each contain all necessary context while minimizing unnecessary information. As Pampari et al. (2018) pointed out, the segmentation of clinical reports into sentences is not straightforward. These complications arise from factors such as the frequent use of dots in acronyms, list items, and values and the irregular alternation of uppercase and lowercase letters. Because our goal is to create concise paragraphs without losing context, we must ensure that paragraph boundaries do not disrupt sentence cohesion or, worse, do not appear in the middle of a word. Therefore, our initial step is to split each report into groups of complete sentences, ensuring that no sentence is fragmented across groups and that no substring of a response is split into two paragraphs.

To achieve this, we leverage the structure of the official emrQA dataset (Pampari et al., 2018). In this dataset, each report text is stored as a list of lines, with the answer evidence (in our case, the answer substring) being one of the report lines. Therefore, we set the condition that none of the sentence groups starts or ends in the middle of any line, ensuring that no answer substring is split into two paragraphs. We then split the *Medication* subset into groups of sentences if the following pattern for the end of a line is satisfied: a dot at the end of a line, preceded by five characters that are neither dots nor uppercase letters, and the next line starting with an uppercase letter (eventually this second line can also be an item of a numbered list, which means it could start with a number instead of an uppercase letter). Clinical notes in the subset of *Relations* are structured more clearly. Dots marking the end of a sentence are surrounded by spaces, while dots forming part of abbreviations are not. Thus, such space-surrounded dots at the end of a line indicate a sentence group boundary in the context of the *Relations* subset.

Another pattern we utilize as a criterion for segmentation contains a sequence of characters ending with a colon, indicating headings followed by corresponding content, as shown in Figure 1. Using the end of the previous line of such heading lines as a sentence group separator makes sense. To decrease the risk of detecting not-heading lines, we only consider uppercase titles for *Medication* when determining sentence group boundaries. In the case of the *Relations* subset, only lines ending with a colon preceded by space are considered, similar to the situation with dots.

Finally, we need to determine how we will group sentence groups together to create a final paragraph segmentation. By using the following regex pattern

```
(^([0-9]+[\s]*[\.\)][\s]*)?[A-Z][a-zA-Z\s\(\)]*:)
```

we identify all potential headings at the beginning of all sentence groups. Subsequently, we calculate how often these headings appear in the training data. We assume that frequently used titles signify sections generally discussed in clinical notes by healthcare professionals that do not need any additional context. Therefore, the question arises: what is the minimum number of occurrences of headings in the training data that we want to use for paragraph separations?

We call such segmentation as *heading-based segmentation*. As the range of possible headings serving as paragraph boundaries increases, the average length of paragraphs decreases. As shown in Table 2, segmenting reports using all detected headings yields PR-QA results comparable to those from unsegmented reports. Therefore, as part of our analysis, where we evaluate how frequently headings should be used as boundaries in segmentation, we assess our three tasks (PR, Oracle-QA, PR-QA) across different segmentations based on varying heading frequencies, resulting in different average segment lengths. This helps us understand

	Medication		**Relations**	
	F1	EM	F1	EM
MedCPT - *unsegmented reports*	68.33	27.48	94.69	87.68
MedCPT - *heading-based PR-QA*	64.79	26.63	94.05	88.44
BERT-base - *unsegmented reports*	70.09	30.07	95.04	89.32
BERT-base - *heading-based PR-QA*	68.19	30.23	95.15	91.28
ClinicalBERT - *unsegmented reports*	72.24	31.13	96.45	90.93
ClinicalBERT - *heading-based PR-QA*	70.80	31.19	96.44	92.69
Doc Reader (Yue et al. (2020))	*70.45*	*25.68*	*94.85*	*86.94*
Human-labled (Yue et al. (2020))	*74.70*	*26.0*	*95.40*	*92.00*

Table 2: Comparison of the results of pre-trained BERT models for QA applied to unsegmented reports and PR-QA applied to heading-based segmentations with the shortest possible average segment lengths. We also include the best results by Yue et al. (2020) evaluated on a test set sampled with a different random seed and their human-labeled analysis evaluated on a sampled subset of the test set.

the challenges involved in distinguishing relevant paragraphs from finding exact substring answers.

Despite the structured nature of the emrQA dataset, the rules for splitting the *Medication* and *Relations* subsets into paragraphs can be generalized to other clinical datasets with caution. Although different countries, hospitals, and doctors may structure their reports differently, there are often similar paragraphs and even common headings across various discharge summaries. This observation allows us to take the list of headings collected from the segmentation process of emrQA and use it when segmenting other discharge summaries. However, some level of preprocessing and postprocessing will always be necessary, as this method is not a one-size-fits-all solution for all clinical reports.

4.1 Medication

The newly created segmented datasets derived from the *Medication* subset need to be analyzed first. When segmenting reports into shorter paragraphs, more paragraph+question+answer triples are generated. This is because some questions have multiple possible answers in different document parts. By breaking the text into paragraphs, these question+answer pairs can be split into two or more. Figure 3 shows this expansion is minimal, only about $3-5\%$. However, this phenomenon does not affect the results since we compare our predictions with the original unsegmented reports. For questions with answers in multiple paragraphs, only answer, the most confident one, is selected for the evaluation. Figure 4 displays a list of 542 discovered headings sorted by their frequency of occurrence. We can see that the first third of the headings appear more frequently in all training reports. In contrast, two-thirds of the headings found do not appear to refer to traditional clinical sections. The average lengths of the segmented paragraphs are shown in Figure 5. Even though we collected headings only from training reports, it did not significantly impact the development and test sets. Anyway, it is still interesting to observe the wide range of segmented paragraph lengths.

We segmented the *Medication* subset into paragraphs with varying average lengths from hundreds to thousands of characters and evaluated the performance of ClinicalBERT (Alsentzer et al., 2019) model on all 3 tasks: PR, Oracle-QA, and PR-QA. We sampled the training dataset and trained the model with three different seeds. The results can be seen in the first row of Figure 2. Results are shown for segmentations with different average paragraph lengths corresponding to the x-axis.

The shorter the paragraphs, the easier the Oracle-QA task, but at the same time, the more paragraphs correspond to one report, making the PR task more challenging. Although the Oracle-QA task tends to perform better in both F1 and Exact Match scores for shorter paragraphs, the difference is not that significant. For an average paragraph length of 2500 characters and less, the model is not always confident in its top selection for the PR task. On the other hand, considering two or three top predictions, the correct paragraph is almost certainly included. After combining the predictions into the PR-QA chart, the resulting curve for the Exact Match remains constant for all possible average paragraph lengths. The curve of the F1 score is also constant, except for the shortest paragraphs. However, overall, the challenging part of the PR-QA is the Oracle-QA prediction.

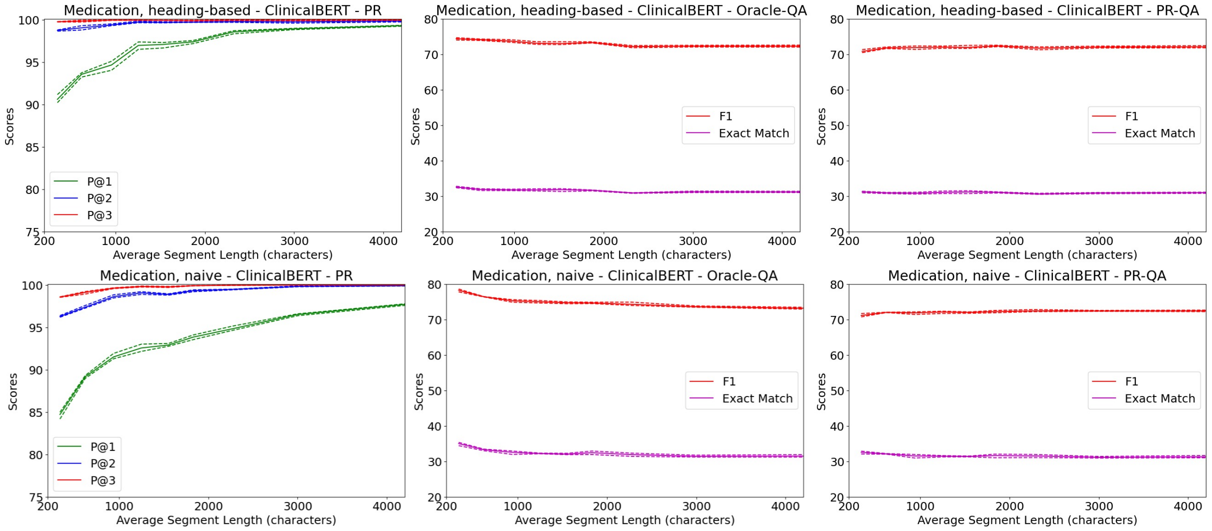


Figure 2: The comparison of heading-based and naive segmentation approaches for different average paragraph lengths using the ClinicalBERT (Alsentzer et al., 2019) model regarding all three tasks (PR, Oracle-QA, PR-QA) on the *Medication* subset. All values are computed as an average of three experiments based on different training seeds. The dashed lines visualize the range of score values.

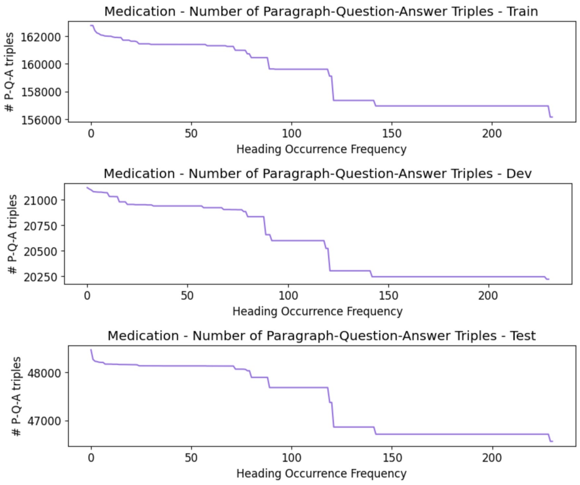


Figure 3: Number of paragraph+question+answer triples of the *Medication* subset in terms of the minimal occurrence frequency of headings we consider for segmentation for the training, development, and test sets.

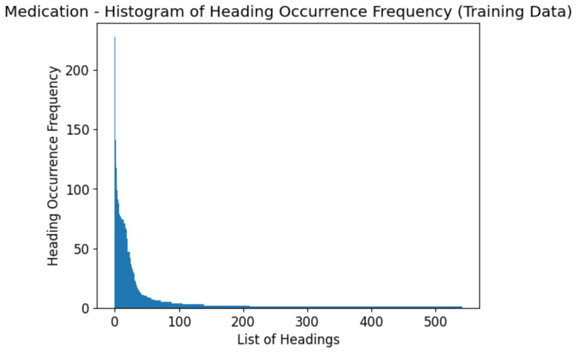


Figure 4: Histogram of heading occurrence frequencies for the *Medication* training set. The x-axis represents the heading positions within the sorted list of headings based on their occurrence frequency. Each point on the x-axis corresponds to a specific heading, with the y-value indicating its occurrence frequency. In total, there are 542 headings, with the most common appearing over 200 times. Many headings appear only once in the training data.

4.2 Relations

Although *Medication* and *Relations* are different subsets with different complexities, Figures 7, 8, and 9 indicate that the header-based segmentation approach behaves similarly for both. However, in this case, we found 953 headings, which we use for segmentation.

We conducted the same experiments on the *Relations* subset as we did in the case of the Medication section. The results, illustrated in the first row of Figure 6, cover the performance of the ClinicalBERT model (Alsentzer et al., 2019) on segmentations of *Relations* subset with varying average paragraph lengths. Given the lower complexity of the *Relations* subset compared to the *Medication*, the model performed better in all three tasks. The PR task achieved better than 98% of $P@1$, even for the shortest paragraphs. The Oracle-QA task indicates that the model performs notably better on shorter paragraphs so that PR-QA results could be improved. Following the combination, i.e., PR-QA task, a constant F1 curve was observed. Further-

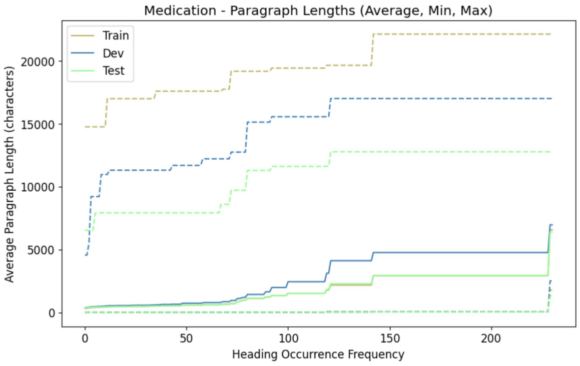


Figure 5: Average paragraph lengths regarding minimal occurrence frequency of headings we consider for *Medication* subset segmentation. The dashed lines show the minimum and maximum lengths of the segmented paragraphs.

more, there is a slight improvement in the Exact Match score by 1–2% when the shortest paragraphs are taken into account.

5 Is Segmentation into Paragraphs Necessary?

We have shown that heading-based paragraph segmentation has no significant effect on the PR-QA task except for improving the Exact Match score of the *Relations* when segmenting into shorter paragraphs. However, most clinical texts are unstructured and use unique text formatting; sometimes, finding a segmentation into coherent sections in the sense of meaning as well as syntax is not easy or even possible. To determine its necessity, we conduct experiments with *naive segmentation*.

We choose a target average segment length t to create the naive segmentation. Then, we calculate each report's length n and determine the rounded number of segments in the report as $p = \texttt{round}(n/t)$. Subsequently, we compute the actual average segment length of the report for the value of p as $r = n/p$. Finally, the report is divided into segments of r characters. Postprocessing is then applied across all segments and all answers in the report. In cases where an answer substring is part of two separate segments, we adjust the segment boundaries so that the entire answer is in one segment only.

The target average segment length t for naive segmentation is chosen to match the average segment lengths of the headings-based segmentation experiments. Specifically, for each measured segmentation level of the headings-based approach, we also measure the naive segmentation using a target average segment length equal to the average segment length of the given headings-based segmentation.

In Figures 2 and 6, we visualize the comparison between ClinicalBERT (Alsentzer et al., 2019) using heading-based and naive approaches. Except for segmentation with the shortest paragraphs, the choice of segmentation method has no noticeable effect on the PR-QA task. In the case of the shortest paragraphs of the *Relations* subset, the naive approach begins to decline in PR-QA performance, while the heading-based approach becomes more accurate. The reason for that is worse performance on the PR task as well as the Exact Match on the Oracle-QA. The performance of naive segmentation on the PR task is significantly worse. On the other hand, the Oracle-QA naive segmentation experiments show better results. The most confident segment contains less relevant content compared to heading-based segmentation, making it easier to find the correct substring as an answer (fewer relevant and potential words in the segment) if the segment itself is predicted correctly. Overall, the PR-QA performance of both heading-based and naive approaches is similar.

6 Paragraphs and LLMs

Considering the impact of segmentation into shorter paragraphs on the scores, it is noteworthy that it does not significantly affect them and may even enhance them. This observation suggests the potential for leveraging LLMs without the necessity for unlimited computational resources in future applications. In this study, we evaluate the performance of BioMistral-7B (Labrak et al., 2024) in the Oracle-QA task and compare it with MedCPT (Jin et al., 2023), BERT-base (Devlin et al., 2018), and ClinicalBERT (Alsentzer et al., 2019) models. BioMistral-7B (Labrak et al., 2024) is trained on question+paragraph+answer triplets where each paragraph contains an answer. Negative examples are omitted to focus solely on the Oracle-QA task. The model prompt is shown in Figure 10. For evaluation, the model's response is parsed into a JSON object, and the value of the "answer" field is extracted.

Table 3 presents the F1 and Exact Match results of the Oracle-QA task using heading-based segmentation with the shortest possible paragraphs, categorized into *Medication* and *Relations* subsets. The results demonstrate that BioMistral-7B

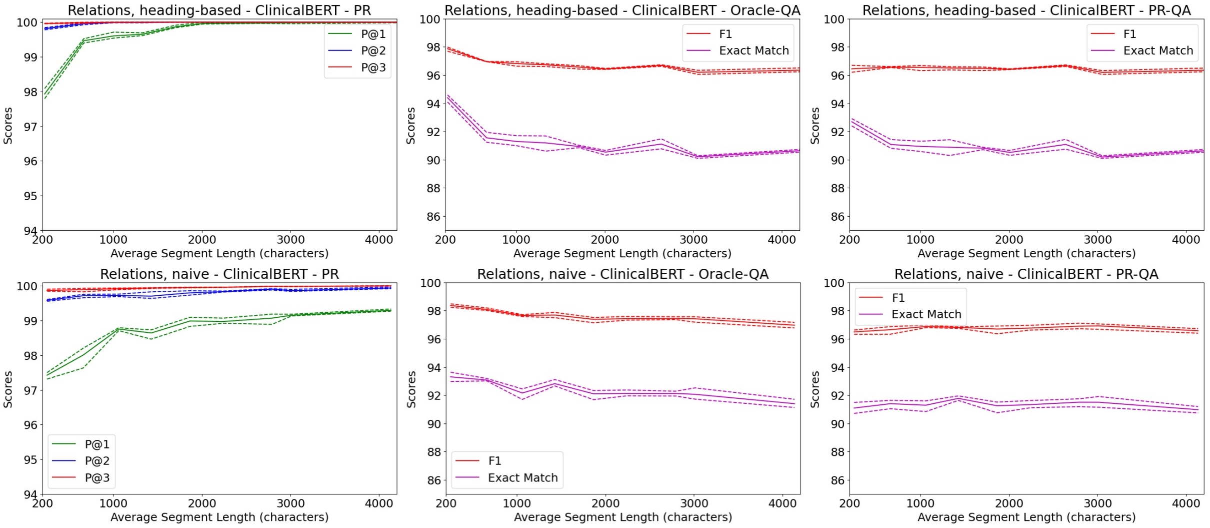


Figure 6: The comparison of heading-based and naive segmentation approaches for different average paragraph lengths using the ClinicalBERT (Alsentzer et al., 2019) model regarding all three tasks (PR, Oracle-QA, PR-QA) on the *Relations* subset. All values are computed as an average of three experiments based on different training seeds. The dashed lines visualize the range of score values.

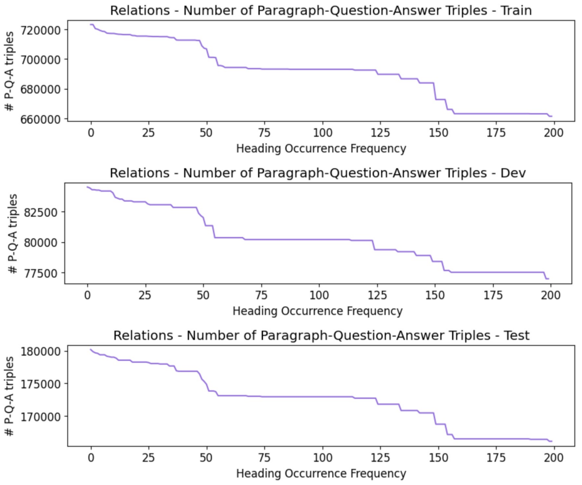


Figure 7: Number of paragraph+question+answer triples of the *Relations* subset in terms of the minimal frequency of occurrence of headings we consider for segmentation for the training, development, and test sets.

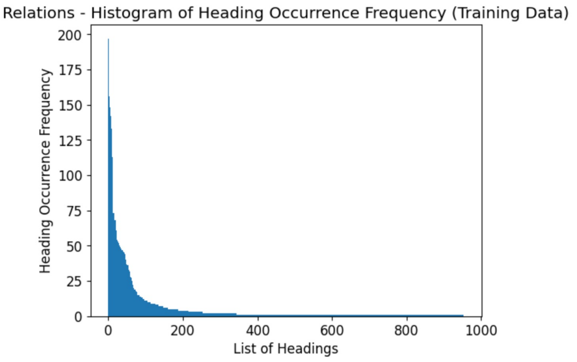


Figure 8: The histogram of heading occurrence frequencies for the *Relations* training set. The x-axis represents the heading positions within the sorted list of headings based on their occurrence frequency. Each point on the x-axis corresponds to a specific heading, with the y-value indicating its occurrence frequency. In total, there are 953 headings, with the most common appearing almost 200 times.

(Labrak et al., 2024) achieves competitive performance but still lags behind encoder-based models such as ClinicalBERT (Alsentzer et al., 2019) and BERT-base (Devlin et al., 2018). BioMistral-7B (Labrak et al., 2024) shows not only promising Exact Match scores compared to MedCPT (Jin et al., 2023), highlighting its potential in clinical QA tasks. However, further exploration is needed to optimize prompts and explore larger models to enhance performance.

7 Conclusions

Our study explores the efficiency of language models in addressing clinical document-level QA. We described an approach to perform heading-based segmentation and extract clinical report headings and found that segmenting documents into shorter sections through heading-based or naive approaches does not decline the performance of ClinicalBERT (Alsentzer et al., 2019), BERT-base (Devlin et al., 2018), or MedCPT (Jin et al., 2023) models. Paragraph length has no significant impact

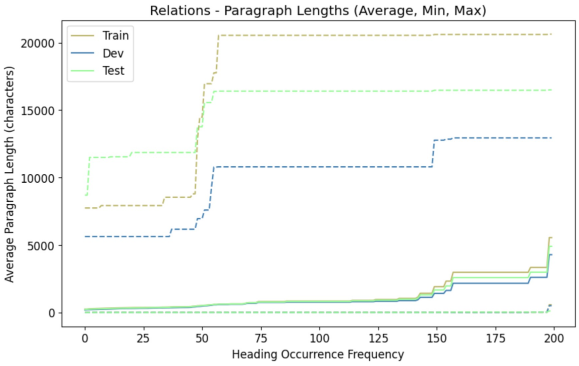
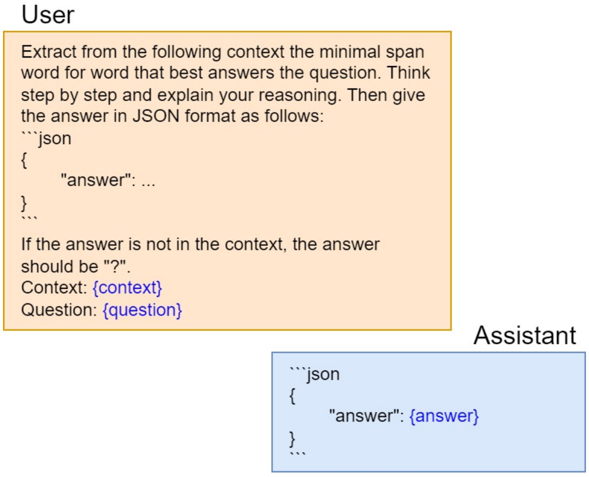


Figure 9: Average paragraph lengths regarding minimal occurrence frequency of headings we consider for *Relations* subset segmentation. The dashed lines show the minimum and maximum lengths of the segmented paragraphs.



User

~~~
Extract from the following context the minimal span
word for word that best answers the question. Think
step by step and explain your reasoning. Then give
the answer in JSON format as follows:
```json
{
    "answer": ...
}
```
If the answer is not in the context, the answer
should be "?".
Context: {context}
Question: {question}
~~~

Assistant

~~~
```json
{
    "answer": {answer}
}
```
~~~

Figure 10: The prompt used for BioMistral-7B training and inference in the Oracle-QA task for extracting answers from a context given a particular question.

on the QA task. Furthermore, knowledge of clinical document characteristics is unnecessary since naive segmentation performs similarly to heading-based segmentation. The main difference is that naive segmentation is more challenging for paragraph retrieval but easier for question answering. In both cases, however, we observe that the correct segment containing the answer is almost always found within the three most confident paragraph retrieval predictions.

Leveraging LLMs like BioMistral-7B (Labrak et al., 2024) shows potential for document-level clinical QA tasks even when computational resources are limited. However, there is still room for improvement and it is necessary to explore other pre-trained LLMs with different training approaches. It remains an open question how the segmented paragraph approach would affect results and behavior on more complex tasks or datasets. Further research is needed to evaluate these methods in more challenging QA scenarios to fully understand their impact and potential.

	m-F1	**m-EM**	**r-F1**	**r-EM**
MedCPT	70.7	28.3	96.7	91.5
BERT-base	73.0	31.9	97.5	94.0
ClinicalBERT	74.4	32.5	97.9	94.4
BioMistral	66.6	29.8	94.4	89.0

Table 3: F1 (**F1**) and Exact Match (**EM**) Oracle-QA results using the heading-based segmentation of the shortest possible paragraphs for both *Medication* (**m**) and *Relations* (**r**) subsets.

Acknowledgments

This work has received support and funding from the European Union's Horizon Europe research and innovation programme project *RES-Q plus* (Grant Agreement No. 101057603).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772. Biomedical Natural Language Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jimenez Eladio and Hao Wu. 2024. emrqa-msquad: A medical dataset structured with the squad v2.0 framework, enriched with emrqa medical information. *Preprint*, arXiv:2404.12050.

Fernando Gallego Donoso and Francisco Veredas. 2023. Icb-uma at biocreative viii @ amia 2023 task 2 symptemist (symptom text mining shared task). In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *Preprint*, arXiv:1909.11942.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *CoRR*, abs/1809.00732.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, page ocae029.

Phillip Richter-Pechanski, Philipp Wiesenbach, Dominic M. Schwab, Christina Kiriakou, Nicolas Geis, Christoph Dieterich, and Anette Frank. 2024. Clinical information extraction for low-resource languages with few-shot learning using pre-trained language models and prompting. *Preprint*, arXiv:2403.13369.

Maximilian Schmidt, Andrea Bartezzaghi, and Ngoc Thang Vu. 2024. Prompting-based synthetic data generation for few-shot question answering. *Preprint*, arXiv:2405.09335.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weißenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and Georgios

Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrQA dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020. Retrospective reader for machine reading comprehension. *Preprint*, arXiv:2001.09694.