Drilling Down into the Discourse Structure with LLMs for Long Document
Question Answering

Inderjeet Nair*1, Shwetha Somasundaram*2, Apoorv Saxena2, Koustava Goswami2
1University of Michigan, Ann Arbor, MI
2Adobe Research, India
inair@umich.edu
{shsomasu,apoorvs,koustavag}@adobe.com

Abstract

We address the task of evidence retrieval for long document question answering, which involves locating relevant paragraphs within a document to answer a question. We aim to assess the applicability of large language models (LLMs) in the task of zero-shot long document evidence retrieval, owing to their unprecedented performance across various NLP tasks. However, currently the LLMs can consume limited context lengths as input, thus providing document chunks as inputs might overlook the global context while missing out on capturing the inter-segment dependencies. Moreover, directly feeding the large input sets can incur significant computational costs, particularly when processing the entire document (and potentially incurring monetary expenses with enterprise APIs like OpenAI’s GPT variants). To address these challenges, we propose a suite of techniques that exploit the discourse structure commonly found in documents. By utilizing this structure, we create a condensed representation of the document, enabling a more comprehensive understanding and analysis of relationships between different parts. We retain 99.6% of the best zero-shot approach’s performance, while processing only 26% of the total tokens used by the best approach in the information seeking evidence retrieval setup. We also show how our approach can be combined with self-ask reasoning agent to achieve best zero-shot performance in complex multi-hop question answering, just \(\approx 4\%\) short of zero-shot performance using gold evidence.

1 Introduction

Long Document Question Answering (LDQA) is a complex task that involves locating relevant evidence from lengthy documents to provide accurate answers to specific questions (Dasigi et al., 2021). LDQA is challenging for the following reasons:

a) Long documents often exceed the maximum token limit of existing transformer-based Pretrained Language Models (PLMs) (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020; Raffel et al., 2020), posing a challenge in directly processing their content to extract pertinent information (Dong et al., 2023). b) The information required to answer a question is often dispersed across different sections or paragraphs within the document which may require sophisticated reasoning process to identify and extract the relevant information (Nie et al., 2022). c) Processing the entire document to find answers can be computationally expensive and inefficient (Dong et al., 2023).

One popular approach for LDQA is the retrieve-then-read method (Zheng et al., 2020; Gong et al., 2020; Nie et al., 2022; Ainslie et al., 2020, 2023), where relevant paragraphs are retrieved from the document to provide the answer. A major drawback of existing works is reliance on supervised fine-tuning for the evidence selection phase, exhibiting poor generalization on out-of-distribution data (Thakur et al., 2021).

Given the remarkable few-shot/zero-shot performance and enhanced generalization capabilities demonstrated by Large Language Models (LLMs) across various Natural Language Generation and Understanding tasks (Brown et al., 2020; Chen et al., 2021; Rae et al., 2022; Hoffmann et al., 2022; Chowdhery et al., 2022), we investigate the potential of leveraging these LLMs for zero-shot evidence retrieval. Notably, LLMs that have been instruction fine-tuned (Wei et al., 2022a; Chung et al., 2022) or trained using Reinforcement Learning with Human Feedback (Bai et al., 2022; Ouyang et al., 2022) exhibit exceptional generalization performance even on unseen tasks (Ouyang et al., 2022; Min et al., 2022; OpenAI, 2023). Thus, we explore the feasibility of utilizing LLMs for zero-shot evidence retrieval. However, LLMs, which are based on transformer architecture (Vaswani
et al., 2017), are limited by their context length and suffer from expensive inference times that increase quadratically with the number of tokens in the input. Additionally, utilizing enterprise LLM solutions such as OpenAI’s gpt-3.5-turbo, text-davinci-003, gpt-4, etc.\textsuperscript{1} to process an entire long document without optimizations would incur significant monetary costs. This highlights the need for an LLM-based evidence retrieval solution that can achieve faster and more cost-effective inference by selectively processing relevant portions of the document, without compromising downstream performance.

To overcome these challenges, we harness the inherent discourse structure commonly present in long documents. This structure encompasses the organization of topics, semantic segments, and information flow, enabling effective information search and knowledge acquisition for question answering. (Guthrie et al., 1991; Meyer et al., 1980; Taylor and Beach, 1984; Cao and Wang, 2022; Dong et al., 2023). Utilizing this valuable structure, we construct a condensed representation of the document by replacing the content within each section with a corresponding summary. This condensed representation is then fed to the LLM, enabling efficient processing of tokens while allowing the model to comprehensively analyze the entire input context for identifying relevant sections. Thereafter, the content within each relevant section is further processed by the LLM for fine-grained evidence retrieval. We call our proposed approach \textbf{D}\textsuperscript{3} (Drilling Down into the Discourse) due to the nature of the solution described above.

Our approach undergoes evaluation in two distinct settings: Information Seeking and Multi-hop Reasoning in Question Answering. In the information seeking experiments, our approach retains the best zero-shot state-of-the-art (SoTA) results, while only utilizing 26\% of the tokens employed by the SoTA approach. Additionally, we examine the robustness of our model across various document lengths and analyze the number of tokens required and latency for different zero-shot approaches. Moreover, we explore the integration of our approach with other zero-shot techniques within an agent framework designed to break down intricate queries into a sequence of simpler follow-up queries.

\section{Related Work}

\subsection{LLMs in Retrieve-Then-Read Approaches}

The retrieve-then-read (Green Jr et al., 1961; Chen et al., 2017; Wang et al., 2018; Das et al., 2019; Guu et al., 2020) approach is a widely adopted technique in open-domain (Voorhees et al., 1999; Dunn et al., 2017; Joshi et al., 2017; Zhu et al., 2021), multi-document question answering (Yang et al., 2018; Perez et al., 2020; Ferguson et al., 2020) and long-document question answering (Pereira et al., 2023). In this approach, LLMs are utilized specifically for the reader component, which generates responses based on the relevant fragments retrieved by the retriever (Pereira et al., 2023). Although LLMs have been utilized as decision-making agents in browser interactions for document retrieval (Nakano et al., 2022), their direct application for fine-grained evidence retrieval has not been extensively explored to the best of our knowledge. On that front, our paper is the first to evaluate the applicability of LLMs for evidence retrieval.

\subsection{Chaining LLMs Runs for Question Answering}

Chaining in LLMs refers to the task of breaking complex overarching task into a sequence of fine-grained targetted sub-tasks where the information generated by a particular run of the sequence is passed to the subsequent runs (Wu et al., 2022a,b). This allows for realizing powerful machine learning applications without requiring any changes to the model architecture (Tan et al., 2021; Betz et al., 2021; Reynolds and McDonell, 2021). LangChain has implemented procedures using chaining for evidence retrieval and question answering in long documents. They employ three chaining variants (map-reduce, map-rerank, and refine)\textsuperscript{2}, which processes document chunks individually and aggregate the information from each chunk to derive the final answer. This implementation, however, processes the entire document input resulting in significant compute and monetary cost.

\subsection{Evidence Retrieval for LDQA}

Prior evidence retrieval approaches typically employ following two mechanisms which are trained by supervised fine-tuning - local processing to handle individual document chunks with occasional information flow between them (Gong et al., 2020).

\footnotesize\textsuperscript{1}https://openai.com/pricing

\footnotesize\textsuperscript{2}https://python.langchain.com/docs/modules/chains/document/
3 D³: Drilling Down into the Discourse

3.1 Problem Formulation

In LDQA, a question q is asked for a document D = [p₁, p₂, . . . , pₙ], where pᵢ (1 ≤ i ≤ n) is the ith paragraph in the natural reading order of D. The task of LDQA is to retrieve a set of relevant paragraphs E_q ⊆ D and generate a free-form answer a based on q and D (Dasigi et al., 2021; Nie et al., 2022). Due to the length of the documents, often exceeding 5K tokens, we employ the retrieve-then-read strategy. This approach involves first determining E_q and subsequently generating a using only q and E_q.

3.2 Motivation

The cognitive strategy employed by humans to search for relevant information from a document entails a systematic approach of first categorizing the information within the document to determine relevant coarse segments and then conducting a deeper analysis of the relevant categories to extract fine-grained segments (Guthrie and Kirsch, 1987; Guthrie et al., 1991; Guthrie and Mosenthal, 1987). Long documents often possess a well-structured discourse that categorizes information coherently based on topical similarity (Cao and Wang, 2022; Nair et al., 2023; Dong et al., 2023). This inherent discourse structure serves as a valuable framework, enabling effective categorization of information and facilitating a clear and logical flow of ideas within the document.

Drawing from these insights, we posit that encapsulating the essence of a section through its name and content summary would yield valuable cues in determining its relevance for answering specific questions. Thereafter, we emulate the above-described cognitive process by fine-grained analysis of relevant sections to extract evidence paragraphs. This methodology offers three key advantages:

1. By condensing each section with its name and summary, we can effectively reduce the document’s token count, enabling LLMs to analyze the entire context and make accurate inferences.
2. By efficiently filtering out irrelevant sections in the initial stage, our method reduces the number of tokens processed.
3. Our method is applicable for any instruction-following LLMs (Ouyang et al., 2022; Min et al., 2022; OpenAI, 2023), enabling zero-shot application without the need for architectural modifications.

3.3 Methodology

Instead of representing a document D as a ordered set of constituent paragraphs, we repre-
sent \( D = [S_1, S_2, \ldots, S_k] \), where \( S_i (1 \leq i \leq k) \) denotes the \( i \)th section, such that, \( \text{name}(S_i) \) and \( \text{paragraphs}(S_i) \) denotes its name / heading and the list of constituent paragraphs respectively. \( \text{paragraphs}(S_i) = [p_{i,j}]_{j=1}^{S_i} \) where \( |S_i| \) denotes number of constituent paragraphs). Note that, \( \sum_{i=1}^{k} |S_i| = n \). Inspired by the cognitive process of knowledge acquisition / information search for question answering, our approach first finds the relevant sections that may answer the question and then, analyses the paragraphs from the relevant sections for fine-grained evidence paragraph retrieval. (Figure 1)

3.3.1 Finding Relevant Sections

The crux of this step is to represent the content in each section \( S_i \) by the summary of \( \text{paragraphs}(S_i) \). Summarization (El-Kassas et al., 2021) refers to the task of generating a concise summary for a given input that captures its main idea within a limited number of tokens, effectively conveying its topical essence. We denote the summarization operation by \( S \), for which we have used \text{bart-large} (Lewis et al., 2020) fine-tuned over CNN/Daily-Mail Corpus (Nallapati et al., 2016). Thereafter, we represent the entire document as follows:

\[ \star \text{Section}: \text{name}(S_i) \\
S(\text{paragraphs}(S_i)) \]

\[ \star \text{Section}: \text{name}(S_2) \\
S(\text{paragraphs}(S_2)) \]

\[ \ldots \]

\[ \star \text{Section}: \text{name}(S_k) \\
S(\text{paragraphs}(S_k)) \]

An instruction is passed to an LLM involving the above representation to identify all the sections that are relevant to \( q \). Due to this condensed representation, the LLM can process the entire document context enabling comprehensive analysis of long range dependencies for accurate inference. Let the set of sections identified as relevant be denoted by \( R_q \subseteq D \).

3.3.2 Fine-Grained Evidence Retrieval

The objective of this step is to infer the set of relevant paragraphs from \( R_q \). Here, we explain multiple zero-shot strategies to achieve this step. We, first, obtain a set of all paragraphs \( P_q \) associated with \( R_q \).

\[ P_q = \bigcup_{S \in R_q} \text{paragraphs}(S) \]

Thereafter, one of the following strategy can be employed for fine-grained retrieval:

1. \text{MONOT5:} This employs MonoT5 (Nogueira et al., 2020), which a sequence-to-sequence model trained over the task of Document Re-ranking (Nguyen et al., 2016), to select the most relevant paragraphs from \( P_q \).

2. \text{BASE:} Each paragraph from \( P_q \) is marked with an identifier and then, these identifier annotated paragraphs are concatenated with a newline separator. Thereafter, we prompt the LLM in a zero-shot manner to generate all paragraph identifiers whose corresponding paragraph is relevant to \( q \). If the number of paragraphs in \( P_q \) exceeds the maximum context length of LLM, we make multiple LLM calls. In each call, we fit the maximum number of paragraphs that can fit into the context length, ensuring that paragraphs are not ‘chopped’.

3. \text{HiERBASE:} In our approach, we adopt a two-step process to capture the essence of each paragraph. Firstly, we represent paragraphs using their corresponding summaries obtained through \( S \). Following that, we employ the \text{BASE} strategy to identify potentially relevant candidates. In the next stage, we apply the \text{BASE} technique once again, this time considering the original content of the paragraphs, to pinpoint the most relevant ones.

In our experimental evaluation, we also explore the effectiveness of chaining these strategies in succession. One such method, called \text{BASE + MONOT5}, combines the \text{BASE} strategy with \text{MONOT5}. This approach initially identifies a relevant set of paragraphs using the \text{BASE} strategy and subsequently employs \text{MONOT5} to refine the selection further, retaining only the most relevant ones from the initially identified set.

For most of the experiments presented in the upcoming sections, we use \text{bart-large} (Lewis et al., 2020) trained over the CNN/Daily-Mail Corpus (Nallapati et al., 2016) for \( S \). For retrieval
We consider the following zero-shot approaches for performance comparison:

(1) MONOT5: In this approach, MONOT5 is directly applied to re-rank the paragraphs of $D = [p_1, p_2, \ldots, p_n]$ based on $q$.

(2) DPR: Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) is a retrieval approach that leverages dense representations. This method utilizes a bi-encoder architecture to generate embeddings for both the documents and the query independently which are used for finding the most relevant documents.

(3) cross-encoder-ms-marco-MiniLM-L-12-v2: This model is a cross-encoder reranker which employs Siamese learning and BERT-like architecture. This model is offered in the sentence-transformers (Reimers and Gurevych, 2019) library. While the library provides many different model checkpoints, we chose cross-encoder-ms-marco-MiniLM-L-12-v2 as it yielded highest $F_1$ score for evidence retrieval.

(4) PARAGRAPH: Every paragraph in $D$ is processed by the LLM independently to assess its relevance to $q$ through a boolean prompt.

(5) CHUNK: The document is partitioned into consecutive fragments, aiming to accommodate as many paragraphs as possible within a predefined token limit called as chunk size (3500). Subsequently, BASE is applied for each fragment.

(6) MAP-REDUCE: This approach, widely adopted in the literature, involves evaluating the relevance of each paragraph to $q$ and subsequently processing the relevant paragraphs together in a single call to the LLM. However, we observed that using LangChain’s implementation directly led to subpar performance in terms of evidence retrieval $F_1$-score and inference cost. This can be attributed to the incompatibility of the prompt with the target domain, resulting in significant performance degradation due to gpt-3.5-turbo’s high sensitivity to the prompt (Ye et al., 2023).

(7) MAP-REDUCE OPTIMIZED: For better alignment with our target task, we made crucial modifications to the original implementation. Building upon the observation that CHUNK outperforms PARAGRAPH in terms of performance, we decided to process document chunks instead of individual paragraphs using the BASE tech-
nique. Following the initial stage, where relevant paragraphs are identified, we concatenate them and subject them to the same strategy (BASE) for further processing.

5 Experiments and Results

We mainly assess the applicability of our method in two scenarios: (1) §5.1: Information-seeking setting (Dasigi et al., 2021) and (2) §5.1: Multi-hop Reasoning in Question Answering (Yang et al., 2018). Thereafter, we conduct an extensive analysis of our approach’s performance across various configurations, as discussed in §5.3.1, and examine its effectiveness in different document length categories, as outlined in §5.3.2. Thereafter, we justify the need for the inclusion of global context modeling in §5.3.3, and highlight the significance of incorporating discourse information in §5.3.4. Finally, §5.3.5 we compare our methods performance with the best performing zero-shot approach for different categories and identify scope for improvement. In our analyses, we will also include insights into the inference cost and latency associated with utilizing large language models (LLMs) for the evidence retrieval stage. This will encompass the monetary cost and the processing efficiency measured in terms of the number of tokens processed and the number of LLM inferences. It is important to note that these measurements pertain exclusively to LLMs and do not encompass smaller fine-tuned models. Unless otherwise specified, we would be using BASE approach for fine-grained retrieval and the overall approach would be called D³-BASE. Due to the monetary cost associated with gpt-based LLMs, we experiment with 150 randomly sampled documents for all the experiments.

5.1 Performance for Information-Seeking Setting

We assess the performance of the models on the QASPER dataset (Dasigi et al., 2021), which comprises information-seeking questions designed for lengthy research papers. The dataset includes a set of ground truth evidence paragraphs and answers. The questions are categorized as extractive, abstractive, yes/no, and unanswerable, and our proposed method must accurately discern the question’s intent to generate a relevant response.

Table 1 presents the results of various approaches, including a fine-tuned state-of-the-art (SoTA) model (Nie et al., 2022), fine-tuned LED model evaluated on gold evidence, and zero-shot gpt-3.5-turbo model evaluated on gold evidence. Among them, our simplest approach, D³-BASE, achieves competitive performance in terms of Evidence $F_1$ score. Notably, it retains 99.6% of the performance of the best zero-shot approach, MAP-REDUCE OPTIMIZED, while processing only 26% of the tokens required by the latter.

The original implementation of MAP-REDUCE suffers from two main limitations. Firstly, it processes each paragraph independently, overlooking the effectiveness of chunk-level processing over paragraph-level processing. Secondly, it employs suboptimal few-shot prompting to retrieve relevant sources, resulting in increased processing costs and poor performance when the few-shot prompting is not well-aligned with the domain. Due to these significant processing costs and the underperformance of PARAGRAPH and MAP-REDUCE, we exclude their evaluations from subsequent analyses. Similarly, among the non-LLM baselines, we exclude the evaluations of DPR and cross-encoder-ms-marco-MiniLM-L-12-v2 due to their poor performance.

While it is possible to enhance the performance of our approach, D³, with alternative configurations (as shown in Section 5.3.1), it remains an excellent choice for rapid testing in different domains. Its cost-effectiveness and minimal latency in terms of API calls make it highly suitable for such purposes.
Table 2: Comparison of various zero-shot approaches for HOTPOTQA-Doc Dataset. While directly applying $D^3$−BASE leads to poor performance, combining this with self-ask prompting methodology yields best performance while processing least number of tokens when compared against other zero-shot LLM-based methods. *: Inference obtained using gold evidence.

5.2 Performance for Questions Requiring Multi-Hop Reasoning

Here, we use HOTPOTQA-Doc dataset (Yang et al., 2018; Nie et al., 2022), where the objective is to answer a complex query involving multi-hop reasoning given two long documents. We have investigated the performance of different zero-shot approaches using two schemes: (a) Direct Processing: Queries are directly fed to the Zero-Shot retrievers to get the relevant evidences. (b) self-ask based Processing: By leveraging the power of elicitive prompting (Yao et al., 2022; Press et al., 2022; Wei et al., 2022b), we employ the technique of self-ask (Press et al., 2022). This approach entails decomposing a complex query into a series of simpler questions, which collectively form the basis for the final answer. Through iterative questioning, the agent analyzes prior answers and previously posed questions to generate subsequent inquiries. Leveraging the zero-shot retrieval approach, the agent obtains relevant answers for each question.

The results for this experiment are tabulated at Table 2. We note the following observations:

- Evidence $F_1$ poorly correlated with Answer $F_1$: Considering same question answering model were used (gpt-3.5-turbo) for each of the zero-shot approaches, we find that the answer performance of MAP-REDUCE OPTIMIZED aligns with that of MONOT5, albeit with noticeably lower evidence retrieval efficacy. It’s worth noting that during dataset construction, only the paragraphs utilized for generating answers were preserved, which does not imply the irrelevance of other paragraphs in addressing the question. This observation highlights the presence of this artifact.

- Augmenting with self-ask boosts performance: This highlights the fact that zero-shot retrievers are better positioned to retrieve fragments for simpler queries and self-ask effectively uses them to get better performance. In fact, our approach $D^3$−BASE is very close in performance to zero-shot question answering with gold evidence.

5.3 Ablations & Analyses

5.3.1 Performance of different configurations of $D^3$

We investigated several configurations to determine possible directions to improve the performance. Following explorations were performed (Table 3):

- Variations in fine-grained retrieval: Although employing MonoT5 can reduce inference costs, it also adversely affects evidence retrieval performance, as evidenced by the table. Conversely, $D^3$−HEIRBASE demonstrates a enhancement in performance with only a marginal increase in inference cost.

- Using LLMs for summarization: We replaced fine-tuned summarizer with enterprise LLMs such as gpt-3.5-turbo and text-davinci-003 (Ouyang et al., 2022) and an open-source LLM, vicuna-13b (Chiang et al., 2023). While the performance increased along several metrics, there is additional processing cost to get the condensed representation of the document (very minimal when smaller bart-large based summarizer was used).

- Exploring alternative LLMs for retrieval: In this analysis, we observe a decline in performance when utilizing other LLMs, highlighting the superiority of gpt-3.5-turbo as the optimal choice for retrieval tasks.

- Investigating alternative LLMs for question answering: We observe a performance boost when employing text-davinci-003. However, it is important to consider the higher monetary cost associated with using this API compared to gpt-3.5-turbo.

https://openai.com/blog/chatgpt
5.3.2 Performance across Different Document Length categories

We divided the test set into different categories based on their respective lengths. We notice the advantage of our method along three aspects:

- **Evidence $F_1$**: Our approach consistently achieves competitive performance in evidence retrieval across various document length categories (Figure 3 (a)).

- **Evidence Retrieval Cost**: Our approach significantly reduces the number of processed tokens compared to other methods in all document length categories (Figure 3 (b)). This cost-efficient characteristic makes it an excellent choice for minimizing both inference and monetary costs, regardless of the document’s length.

- **Latency**: Irrespective of the document’s length, our approach maintains a minimal latency by making approximately 2 API calls to the LLM (Figure 3 (c)). This efficient performance further highlights its desirability and suitability for various applications.

5.3.3 Need for Global Context Modeling

In this section, we investigate the effect of changing the chunk size of the baseline CHiNK in terms of evidence precision, recall and $F_1$-score. As we see from Figure 4, the precision and $F_1$-score increases at the cost of modest decrease in recall as the chunk-size increasing. This observation underscores the fact that larger chunk sizes enable the model to capture longer-range relationships and contextual information, resulting in improved performance. By processing the entire document in a single pass, D$^3$ benefits from accessing the global context and long-range relationships, leading to enhanced performance.

5.3.4 Role of Discourse Headings and Sub-headings

To assess the importance of section headings / sub-headings in the discourse structure for retrieval, we replaced them with randomly initialized Universally Unique Identifiers (UUIDs) and test evidence retrieval performance over the QASPER dataset (Table 4). The significant decrease in the performance shows that section headings / sub-headings are crucial in conveying the topical essence of a section, which is needed for accurate retrieval.

---

**Table 3**: Performance for different D$^3$ configurations for QASPER dataset.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Extractive</th>
<th>Answering Performance</th>
<th>Abstractive</th>
<th>Yes/No</th>
<th>Unanswerable</th>
<th>Overall</th>
<th>Evidence $F_1$</th>
<th>Tokens Processed</th>
<th>API Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUR PRIMARY APPROACH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D$^3$-BASE</td>
<td>42.90</td>
<td>23.65</td>
<td>74.35</td>
<td>79.61</td>
<td>47.45</td>
<td>49.92</td>
<td>1980.94</td>
<td>1.99</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4**: Performance of different methods with and without section information (denoted by ×) for QASPER dataset.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Evidence $F_1$</th>
<th>Tokens Processed</th>
<th>API Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>D$^3$-BASE</td>
<td>49.92</td>
<td>1980.94</td>
<td>1.99</td>
</tr>
<tr>
<td>D$^3$-BASE ×</td>
<td>43.11</td>
<td>2183.55</td>
<td>1.88</td>
</tr>
</tbody>
</table>
Evidence $F_1$ across different document length categories

Tokens Processed for retrieval across different document length categories

API calls for retrieval across different document length categories

Figure 3: Analysing the performance of different approaches across different length categories along three metrics for QASPER Dataset

Figure 4: Evidence Precision, Recall and $F_1$ over different chunk lengths for QASPER Dataset

5.3.5 Evidence Retrieval Performance over Question Categories

In this section, we compare $D^3$-BASE with the best performing zero-shot baseline MAP-REDUCE OPTIMIZED (MRO) over different question categories in QASPER to identify differences in model performance (Table 5). While our approach is precise in identifying the evidence paragraphs, it occasionally falls short in identifying all pertinent evidence (i.e. lower recall). This indicates that representing a section with a summary leads to loss of information and the LLM may miscategorize its relevancy to a question due to this loss of information. Designing an effective strategy to prevent loss of vital information for a particular question may be a future research direction.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Extractive</td>
<td>52.63 / 52.19</td>
<td>71.9 / 81.49</td>
<td>56.39 / 57.46</td>
</tr>
<tr>
<td>Abstractive</td>
<td>41.25 / 45.1</td>
<td>52.51 / 71.57</td>
<td>41.78 / 50.12</td>
</tr>
<tr>
<td>Yes/No</td>
<td>43.6 / 38.79</td>
<td>59.6 / 55.31</td>
<td>45.53 / 40.18</td>
</tr>
<tr>
<td>Unanswerable</td>
<td>42.94 / 25.74</td>
<td>44.4 / 25.46</td>
<td>43.43 / 25.10</td>
</tr>
</tbody>
</table>

Table 5: Comparison of evidence retrieval performance of $D^3$-BASE with the best performing zero-shot baseline MAP-REDUCE OPTIMIZED (MRO)

6 Conclusion

We demonstrated a zero-shot approach for evidence retrieval which leverages the discourse structure of the document for information categorization which not only yielded competitive performance for information seeking and question answering with multi-hop reasoning setting, but also processed lowest number of tokens resulting in significant compute and cost savings. This approach demonstrates several desirable characteristics such as robustness in evidence retrieval performance, lower latency, etc. all across different document length ranges.

7 Limitations

- Although our approach demonstrated competitive performance in the information seeking setup, there is room for improvement, particularly when confronted with questions that require intricate multi-hop reasoning. Since we represent the document by summarizing each section, there is a potential loss of critical information that is essential for addressing complex queries necessitating multi-hop rea-
soning. Moving forward, we aim to explore methods that allow for accurate section selection while minimizing inference costs and mitigating information loss.

- Our experimental analyses primarily focus on enterprise-level language models, which require enterprise compute credits. In the future, we plan to explore the capabilities of more advanced open-source models as they become available, which may offer enhanced performance and accessibility.

- While our experiments have primarily centered around single-document use cases, we have yet to delve into the realm of retrieval involving multiple documents or collections. This area remains unexplored, and we anticipate investigating strategies and techniques to effectively handle such scenarios.

- Although our evaluations provided insight into how various summarizers impacted the final downstream performance, the current study did not inherently assess the quality of summarization. In future work, we aim to assess the summarization’s faithfulness to the original content and its impact on end-to-end performance.

References


LangChain. Langchain.


Retrieving and reading: A comprehensive survey on open-domain question answering. 