# Graph Representation of Narrative Context: Coherence Dependency via Retrospective Questions

Anonymous ACL submission

## Abstract

This work introduces a novel and practical paradigm for narrative comprehension, stemming from the observation that individual passages within narratives are often cohesively related than being isolated. We therefore propose to formulate a graph upon narratives dubbed NARCO that depicts a task-agnostic coherence dependency of the entire context. Especially, edges in NARCO encompass retrospective free-form questions between two context snippets reflecting high-level coherent relations, inspired by the cognitive perception of humans who constantly reinstate relevant events from prior context. Importantly, our graph is instantiated through our designed two-stage LLM prompting, thereby without reliance on human annotations. We present three unique studies on its practical utility, examining the edge efficacy via recap identification, local context augmentation via plot retrieval, and broader applications exemplified by long document QA. Experiments suggest that our approaches leveraging NARCO yield performance boost across all three tasks.

## 1 Introduction

Text comprehension has been advanced significantly ascribed to Large Language Models (LLMs), especially with long context window enabled via techniques such as positional scaling (Xiong et al., 2023; Peng et al., 2024) and efficient attention (Wang et al., 2023; Chen et al., 2024). Nevertheless, though extending context window could resolve certain long context tasks end-to-end, e.g. question answering, more fine-grained tasks that require explicit global dependency beyond local evidence still remain a challenge.

In book-level narrative understanding particularly, such as retrieving relevant plot passages of queries (Xu et al., 2023), or identifying recap passages of a given plot (Li et al., 2024), each local passage in a novel rather serves specific purposes to other parts than being isolated, which may be easily neglected in the end-to-end process without explicit modeling these global dependency relations.

Traditionally, discourse parsing is established to capture those coherence relations between sentences, characterizing how each proposition relate to others to reflect high-level understanding of the global content (Grosz and Sidner, 1986). Past works have materialized various discourse frameworks, such as Rhetorical Structure Theory (Mann and Thompson, 1988) and Penn Discourse Treebank (Prasad et al., 2008). However, despite its adoption in certain applications (Bhatia et al., 2015; Ji and Smith, 2017; Xu et al., 2020; Pu et al., 2023), they have not attracted appreciable focus in a wider spectrum of tasks; the underlying reasons may be twofold. First, popular discourse formalisms pose finite relation space with fixed taxonomies, offering trivial auxiliary signals especially upon LLMs. Second, they require trained experts to perform annotation for training proper parsers, hindering the overall utility due to inevitably limited resources.

In this work, we propose an alternative path to handle the aforementioned challenges in long narrative understanding, which can be deemed as a new paradigm of quasi-discourse parsing. To overcome previous limitations so to promote practical values for narrative tasks, our approach is designed to obtain flexible coherence relations without tying to formal linguistics or human annotations, thus being directly applicable as an off-the-shelf option.

Drawing inspiration from the human cognitive process on narrative perception, whereas humans can constantly reinstate relevant causal events from past context during reading (Trabasso and Sperry, 1985; Graesser et al., 1994), our proposed idea is simple and intuitive: a NARrative COgnition graph is built, dubbed NARCO, where the entire context is split into small chunks that serve as graph nodes, and edges are connected that represent the relations between node pairs. Particularly, edges are constituted by free-form questions regarding the two con-

necting nodes, aligned with recent discourse works on Questions Under Discussion (Kuppevelt, 1995) such as DCQA (Ko et al., 2022, 2023). As humans could relate to past context in retrospect, accordingly, each question in NARCO arises from the succeeding node, asking necessary background or causes that can be clarified by the preceding node. Hence, graph edges consist of inquisitive questions that naturally reflect retrospection. Overall, the resulting graph explicitly depicts task-agnostic understanding of fine-grained coherence flow that could be flexibly utilized by downstream tasks.

The key difficulty of the graph lies in the edge realization, which itself requires capable context understanding, to determine which aspects to inquire upon the context and distinguish whether they are salient for the comprehension. Such process is especially strenuous due to the large hypothesis space compared to conventional discourse formalisms, which may only become feasible recently with assistance by LLMs. To this end, we construct edges automatically through our proposed LLM prompting scheme without human annotation constraints, of which consists a question generation stage and a back verification stage (Section 3).

NARCO primarily addresses challenges for narratives on two perspectives. First, the graph edges provide a view of explicit information flow, enabling task-specific guidance towards the narrative development. Second, each chunk is now enriched with dependency of global coherence relations, thus provided augmentation of local context to deepen the digest of independent passages.

To empirically demonstrate the practical utility of NARCO, we present studies on the following narrative understanding tasks from three angles.

• Our first study examines the **edge efficacy** on *whether it expresses competent coherence relations* (Section 4). We conduct experiments on recap passage identification (Li et al., 2024), where NARCO boosts up to 4.7 F1 over the GPT-4 baseline.

• Our second study concerns the exploitation of **enriched local embeddings**, by *injecting edges with global dependencies into node representation* (Section 5). Evaluated on the plot retrieval task (Xu et al., 2023), our proposed approach with NARCO outperforms the zero-shot baseline by 3% and the supervised baseline by 2.2%.

• Lastly, we utilize NARCO in long document question answering as a broader application (Section 6). Experiments on QuALITY that requires global evidence (Pang et al., 2022) suggest that, NARCO consistently raises zero-shot accuracy by 2-5% upon retrieval-based baselines with various LLMs, able to recognize more relevant context.

Overall, our contributions can be listed as follows:

- A new paradigm for narrative understanding is proposed, parsing the context into a graph of high-level coherence relations, named NARCO.
- The graph is practically realized with our two-stage LLM prompting w/o human annotations.
- We present three studies effectively utilizing NARCO on narratives, focusing on edge efficacy, node augmentation, and broader applications.

## 2 Related Work

**Questions Under Discussion** QUD is a linguistic framework with rich history that approaches discourse and pragmatics analysis by repeatedly resolving queries triggered by prior context (Kuppevelt, 1995; Roberts, 1996; Benz and Jasinskaja, 2017). Recent works have begun adapting QUD for discourse coherence (De Kuthy et al., 2018, 2020; Ko et al., 2020, 2022, 2023) or other applications (Wu et al., 2023b; Newman et al., 2023). Our proposed NARCO can also be perceived as a unique form of QUD variant, though it is principally rooted upon narrative comprehension rather than formal linguistics. Therefore, NARCO differs from QUD works considerably on the following design choices.

• **Coarse Granularity** While QUD tends to employ sentences as the basic discourse unit, NARCO opts for a coarser granularity, adopting passages (or short chunks) as graph nodes. It is driven by the fact that in narratives, complex events or interactions may often be conveyed beyond sentence-level, thus relations in NARCO could target higher-level understanding between nodes.

• **Retrospection-Oriented** Unlike conventional QUD that inquires from prior context to be addressed by subsequent context (forward direction), which could yield unanswerable questions (Westera et al., 2020; Ko et al., 2020), NARCO takes the backward direction, by asking retrospective questions from succeeding context, such that all generated questions in NARCO are naturally grounded by corresponding prior context.

• **Precision-Focused** Unlike previous QUD works that require dedicated human annotations, NARCO is formulated accomplishable by LLMs. Accordingly, we prioritize precision over recall for practical instantiation of graph edges, and do not

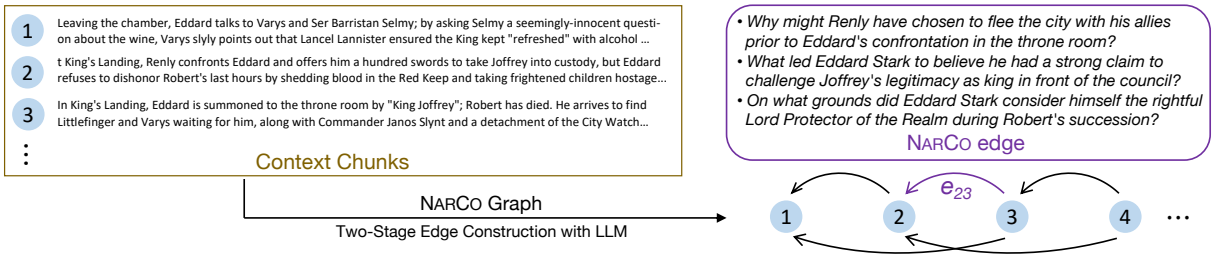


Figure 1: Our proposed NARCO graph described in Section 3, with retrospective questions connecting two nodes.

necessitate strict linguistic criteria, as long as it contributes positively for narrative understanding.

**Long Context Understanding** One of the major research directions of LLMs is the extension of context window, which can be seamlessly applied for long context understanding tasks, including scaling positional embeddings (Chen et al., 2023b; Xiong et al., 2023; Peng et al., 2024), efficient attention (Zaheer et al., 2020; Chen et al., 2024), cached attention (Wu et al., 2022; Wang et al., 2023), recurrent attention (Dai et al., 2019), etc. Though effective, certain narrative tasks demand beyond the end-to-end solution. Recently, new paradigms have been proposed for fine-grained processing, such as compressing context as soft prompts (Chevalier et al., 2023), and MEMWALKER that reads long context interactively via iterative prompting (Chen et al., 2023a). Nevertheless, our proposed approach takes parsing as an alternative paradigm, which is orthogonal to the existing directions and could be even further combined in the future.

## 3 NARCO: Narrative Cognition Graph

In this section, we start by delineating our graph formulation, which is itself not tied to any particular implementation. Subsequently, we elaborate our graph realization using LLMs, as our endeavor to remove dependence on human annotations.

### 3.1 Graph Formulation

**Nodes** For a narrative, the entire context is split into short consecutive chunks (or passages), such that each is within a maximum word limit and constituted by sentences or paragraphs. Graph nodes are then all the chunks adhering the left-to-right sequential order, denoted by $\mathcal{V} = \{v_1, v_2, .., v_N\}$, with $N$ being the total number of chunks.

**Edges** An edge connecting two nodes signifies the relationships they convey. These relations are articulated as free-form inquisitive questions that are not constrained by fixed taxonomies. All edges follow the backward direction, such that for an edge $e_{ij}$ $(i < j)$, the expressed questions always arise from the succeeding node $v_j$, asking clarification regarding specific events or situations appeared in $v_j$, which could be addressed by the preceding $v_i$. For narratives, questions primarily target causal and temporal relations as the coherence dependency.

Functionally speaking, these backward edges resemble the human cognitive process for narrative perception: when reading a certain passage, humans are able to reinstate previous relevant parts in retrospect that lay out the build-up or causes, so to achieve a causally coherent comprehension of the global context (Trabasso and Sperry, 1985; Graesser et al., 1994; Song et al., 2020). Unlike QUD that originally features curiosity-driven questions in a forward direction, which could yield unanswerable questions, all edges in NARCO are fully grounded by the context, such that all questions are addressable by prior nodes, thereby serving as the bridge for global coherent dependency.

Derived upon the above formulation, an edge $e_{ij}$ in NARCO has the following features:

- It may have zero-to-many questions. An empty edge without questions indicates the succeeding node $v_j$ is independent from $v_i$ in terms of causal or temporal relations.
- Each question should be salient towards the comprehension of narrative development, rather than being trivial details. Hence, the number of questions of an edge should reflect how coherently related between the two nodes.
- As we adopt coarse granularity for nodes, questions could inquire higher-level relations based on the extrapolation over multiple sentences, which may be useful for downstream tasks.

### 3.2 Graph Realization

To obtain graph nodes, the full context is split by paragraphs and sentences. We impose each node within 240 words in this work, though the exact limit can be task-specific. For a graph characterized

by $N$ total nodes, there are $O(N^2)$ edges available, which can become cumbersome and excessive. It is also task-dependent to determine which pairs of nodes should be gathered edges upon, e.g. for enriching local representation, it may be enough to obtain proximal coherence relations by edges from neighboring nodes within a context window.

Despite the daunting task of edge construction, the emergence of LLMs presents an opportunity: through LLM prompting, it becomes conceivable to actualize the entire graph without any human annotations involved. To this end, we introduce a two-stage prompting scheme as follows for the challenging edge formulation in NARCO.

**Question Generation** For an edge $e_{ij}$ to be instantiated, LLM needs to determine important aspects to ask upon $v_j$ that reflect the retrospective coherence towards the prior context in $v_i$. Similar utilization of LLMs for question generation (QG) has been explored in other applications, such as performing QG for QUD (Wu et al., 2023a) and passage decontextualization (Newman et al., 2023), where LLM is prompted to generate questions directly based on task-specific criteria. For our case, such direct generation can be briefly outlined as:

> Given a current context $v_j$ and its prior context $v_i$, generate questions upon $v_j$, such that each question asks about the cause or background of specific events or situations in $v_j$, which can be clarified by $v_i$, so to reflect their causal or temporal relations between the two context.

However, our preliminary experiments found that, though LLM could follow the instructions to generate plausible questions, their quality is often unsatisfactory for NARCO, with common errors as follows (examples in Appx A.2):

- LLM often asks questions upon $v_j$ but also answerable by $v_j$ as well. Such patterns align with the more conventional QG setting (Du et al., 2017) that may exist plentifully during supervised finetuning of LLM. However, they are not desirable for NARCO as they cannot effectively indicate coherence with $v_i$.
- LLM could hallucinate their relations by guessing and inferring extra underlying connections not grounded by the provided context, resulting in questions not directly answerable by $v_i$.

In essence, QG for NARCO requires LLM simultaneously aware of questions being: 1) arising from $v_j$; 2) not answerable by $v_j$; 3) answerable by $v_i$. As this process is empirically challenging even for strong LLM (e.g. GPT-4), we perform QG with two heuristic turns that could be viewed as Chain-of-Thoughts (Wei et al., 2022) guided by prompts:

1. List concrete parts in $v_i$ that contribute as the preceding background or cause for specific events or situations mentioned in $v_j$, along with brief explanations.
2. Convert each listed connection to a question, such that it asks about the cause or background upon $v_j$ and can be clarified by the corresponding concrete part in $v_i$, helpful to comprehend their causal or temporal relations.

**Question Filtering** Our pilot study suggests that the above two-turn QG could yield higher-quality questions than rudimentary generation, especially reducing self-answerable questions. However, it is still inevitable to produce noisy questions of the two identified error types. In light of remaining noises, we propose a second stage to filter noisy questions through back verification, akin to the concept of back translation (Sennrich et al., 2016):

> Given a context $\mathcal{C}_{ij}$ and a related question, determine whether it can be answered. If so, reason the answer and provide original sentences of key supporting evidence.

Particularly, $\mathcal{C}_{ij}$ is the concatenated context from $v_i$ and $v_j$ without disclosing their boundary. If the question is answerable, we then parse the response and identify whether the supporting sentences are from $v_i$. If not, the question becomes invalid and discarded, as it does not offer to bridge two nodes.

Overall, all generated questions are back verified; only questions that could be answered by prior nodes are finally retained in NARCO, being a precision-focused approach. In this work, we adopt GPT-4 for strong question generation, and ChatGPT for the easier verification. Our full prompts and more details are provided in Appx A.1.

### 3.3 Graph Analysis

Upon examination of preliminary results on the English version of the novel *Notre-Dame de Paris*, edges in NARCO mostly encompass *what* and *why* types of questions, approximately constituting 61% and 26%. It is worth noting that many questions reflect high-level understanding of the context (examples in Appx A.2), in contrast to conventional discourse relations, e.g. *purpose*, *condition* in RST (Mann and Thompson, 1988). With the two-stage prompting scheme, the averaged node degree reaches 1.9. Particularly, the verification stage identifies 47.4% questions to be filtered out.

As our graph formulation primarily aims at practical values, we demonstrate its effective utility in Sections 4-6, addressing three distinct perspectives.

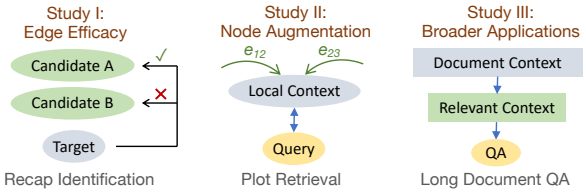


Figure 2: Three presented studies leveraging NARCO.

## 4 Study I: Edge Efficacy

Our first study examines the graph edges on whether they express useful relations. Ideally, a non-empty edge should bridge coherence between two nodes through its retrospective questions. For appropriate assessment, we adopt the recap identification on RECIDENT dataset (Li et al., 2024), a task on narratives that identifies whether certain preceding snippets can function as a recap or prelude to the audience in regards to a current context.

Concretely, the input takes a short snippet from a novel or show script, along with a provided list of its preceding snippets; this task resolves which preceding snippets are directly related with the current one in terms of plot progression, requiring contextual understanding of narrative development. As NARCO is proposed to capture the inter-node coherence relations, edges of retrospective questions could be leveraged to link the current snippet to related preceding ones. Therefore, RECIDENT serves as a natural testbed for comprehensive evaluation of edge efficacy.

### 4.1 Approach

For this study, our proposed approach targets upon the zero-shot baseline with LLMs in (Li et al., 2024), where ChatGPT is originally asked to select the related recap snippets from the list of preceding candidates based on their context.

With NARCO, we regard each current snippet as a **t**arget graph node $v_t$, and the list of its $N$ preceding snippets $\{v_c|c = 1, .., N\}$ as the **c**andidates. For $v_t$ and each of its candidate $v_c$, the edge is realized denoted by $e_{ct}$. As each question in $e_{ct}$ should reflect their causal or temporal relations, we utilize these questions directly from two distinct aspects.

**Edge Relations** Normally, each snippet is represented by its context as in the baseline. To evaluate the coherence depicted by edges, we instead propose to represent each snippet solely based on the edge relations: for a candidate node $v_c$, we use its concatenated questions in $e_{ct}$ for representation, denoted by $\{q_c|c = 1, .., N\}$, and completely neglect the original context, so to ensure an entirely isolated assessment of edge relations.

Specifically, given the context of the target snippet $v_t$, we now ask LLM to select which $q_c$ addresses important questions asking recap information significant to comprehend the current context. As each $q_c$ contains multiple questions, we ask LLM to score each $q_c$ in $[0, 5]$, with higher scores indicating better overall questions. Candidates with empty edges are directly assigned 0 score.

**Edge Degrees** Alternatively, as mentioned in Section 3.1, the edge degree (number of questions) could suggest how cohesively related between two nodes. To this end, we further propose to simply deem the edge degree as the score for ranking candidates, without any inference on the node context or edge relations at all. Though being rather unconventional, ranking recaps by edge degrees could approximately reflect the edge quality.

For either the relation score or degree score, it could be used standalone or interpolated with the baseline selection. More formally, we obtain the rank $\in [1, N]$ of each candidate $i$ by relation scores, denoted as $r_i^{rel}$, and the rank by degree scores as $r_i^{deg}$, along with the binary selection $b_i$ from baseline. The final score $s$ of each candidate is:

$$s_i = \alpha \cdot r_i^{rel} + \beta \cdot r_i^{deg} - \lambda \cdot \mathbb{I}(b_i) \qquad (1)$$

$\mathbb{I}$ is the indicator function that boosts the baseline selection by $\lambda$ rank; relation and degree ranks are interpolated by $\alpha$ and $\beta$. The final score is then ranked to select top candidates with recap information (lower is better). Setting these to 0 accordingly can thereby evaluate each method standalone.

### 4.2 Experiments

**Data** As RECIDENT includes multiple novels and show scripts, we pick one classic novel *Notre-Dame de Paris* (NDP) in English and one TV show *Game of Thrones* (GOT) to reduce the evaluation API cost from OpenAI. The test set of each source consists of 169 / 204 target snippets respectively. Each target is provided 60 candidate snippets, with 5.6 / 4.9 candidates being positive on average.

**Evaluation Metric** We follow Li et al. (2024) and adopt F1@5 (F1 on top-5 selected candidates) as the main evaluation metric.

**Methods** We conduct zero-shot LLM experiments with both ChatGPT (*gpt-3.5-turbo-1106*) and GPT-4 (*gpt-4-1106-preview*) from OpenAI.

- **BL**: the original ChatGPT baseline (*Listwise + Char-Filter* from Li et al. (2024).) We additionally run GPT-4 for comprehensive evaluation.
- **Rel**: standalone ranking by edge relations, without using any candidate context itself.
- **Full**: full interpolation by Eq (1) with both edge relations and degrees. Coefficients are set through a holdout set from another novel.

### 4.3 Results and Analysis

Table 1 shows the zero-shot evaluation results on the test set of RECIDENT. Notably, the interpolation with NARCO edges (Full) consistently brings significant improvement upon the baseline (BL), by 4.9 / 2.4 F1 on NDP / GOT respectively with ChatGPT, up to a 21.7% relative improvement. The stronger GPT-4 boosts performance for all methods as expected, and still advancing 3.5 / 4.7 F1 upon BL on NDP / GOT as well.

Moreover, selection solely based on edge relations without disclosing the context (Rel) could obtain comparable or better performance than the baseline, with the only exception of ChatGPT on GOT. Overall, Table 1 demonstrates the effective utility of NARCO leveraging its edge efficacy, offering a complementary enhancement.

| | NDP | | | GOT | | |
|---|---|---|---|---|---|---|
| | P@5 | R@5 | F@5 | P@5 | R@5 | F@5 |
| | *ChatGPT* | | | | | |
| BL | 22.22 | 22.97 | 22.59 | 31.94 | 38.87 | 35.07 |
| Rel | 22.84 | 23.34 | 23.09 | 28.63 | 37.09 | 32.31 |
| Full | 26.86 | 28.16 | **27.50** | 33.04 | 43.27 | **37.47** |
| | *GPT-4* | | | | | |
| BL | 25.34 | 25.53 | 25.44 | 31.49 | 40.38 | 35.38 |
| Rel | 26.39 | 27.23 | 26.80 | 31.18 | 42.05 | 35.81 |
| Full | 29.11 | 28.74 | **28.92** | 34.90 | 46.93 | **40.03** |

Table 1: Zero-shot evaluation on the test set of RECIDENT for recap identification (Section 4.2). Our approaches with NARCO achieve significant improvement upon the baseline (BL) for both ChatGPT and GPT-4.

For more in-depth insights, we further perform two additional evaluation with ChatGPT:

- **Deg**: standalone ranking by edge degrees; for tied degrees, closer candidates are prioritized.
- **Full$^{-F}$**: the Full setting with all generated questions, without the back verification stage.

Corresponding results are shown in Table 2, where ranking by edge degrees of NARCO exhibits decent performance. It even surpasses the baseline on NDP by 1+%, which is impressive for the fact that it does not undergo any task-specific inference. Understandably, it indeed lags behind the baseline on GOT by a noticeable margin. As for Full$^{-F}$, the trivially degraded performance suggests that, our proposed approach can be quite robust against noisy edge questions, as LLM assigns high scores as long as under the presence of good questions.

| | NDP | | | GOT | | |
|---|---|---|---|---|---|---|
| | P@5 | R@5 | F@5 | P@5 | R@5 | F@5 |
| BL | 22.22 | 22.97 | 22.59 | 31.94 | 38.87 | 35.07 |
| Full | 26.86 | 28.16 | **27.50** | 33.04 | 43.27 | **37.47** |
| Deg | 23.31 | 24.44 | 23.86 | 27.45 | 37.67 | 31.76 |
| Full$^{-F}$ | 26.39 | 27.06 | 26.72 | 33.24 | 42.57 | 37.33 |

Table 2: Zero-shot evaluation with ChatGPT, using NARCO edge degrees (Deg) and all questions (Full$^{-F}$).

## 5 Study II: Node Augmentation

Our second study underscores the NARCO utility of local context augmentation, examining whether the graph typology could enrich the node representation with global contextual information.

Specifically, for a node $v_j$, a preceding node $v_i$ and succeeding node $v_k$ such that $i < j < k$, $e_{ij}$ depicts *outgoing* questions arising from $v_j$ to $v_i$, and $e_{jk}$ specifies *incoming* questions from $v_k$ that can be clarified by $e_j$. These questions either highlight important aspects of events or situations in the current context, or provide implication of subsequent development. Such auxiliary information from neighboring nodes is especially useful for retrieval on narratives, as each passage is not independent and rather being related with others.

We hence investigate if an embedding function on top of NARCO could lead to enriched local representation. Towards this objective, we consider the plot retrieval task defined in (Xu et al., 2023), which aims to find the most relevant story snippets given a query of short plot description. It is challenging as queries are often abstract based on readers' overall understanding of the stories, requiring essential background information clarified on candidates, similar to the concept of *decontexualization* (Choi et al., 2021). Retrieval on narratives thereby fits our evaluation purpose well.

### 5.1 Approach

For this task, candidate snippets from stories are retrieved upon a given query. We build the graph for the full narrative, e.g. an entire novel, according

to Section 3 and regard all candidate snippets as graph nodes to be retrieved from. Our proposed approach focuses on fusing edge questions into the node representation for enhanced retrieval.

Xu et al. (2023) follows the classic paradigm of contrastive learning that learns a BERT-based encoder (Devlin et al., 2019) on queries and candidates. As its trained model is not released yet, our approach adopts the public BGE encoder (Xiao et al., 2023) in this work that ranks top on the MTEB leaderboard[1]. For comprehensive evaluation, we propose methods with NARCO for both zero-shot inference and supervised training.

#### 5.1.1 Zero-Shot Retrieval

Since edge questions are available to provide auxiliary information, edges can be directly integrated in the zero-shot retrieval process. Our motivation is straightforward: if there can be improvement with zero-shot retrieval, it ensures that these questions bring positive information gain, thus confirming the efficacy for augmenting local context.

Concretely, the hidden states (embeddings) for the query, nodes and edges are obtained by the encoder. Let $\mathbf{h}_i^v$ be the L2-normalized hidden state for the *i*th node, $\mathbf{h}_{ij}^e$ for its *j*th outgoing questions, $\mathbf{h}^q$ for the query. The interpolated similarity $\mathcal{S}_i$ between the query and $i$th candidate is defined as:

$$\mathcal{S} = \mathbf{h}^q \cdot \mathbf{h}_i^v + \lambda \cdot \max(\mathbf{h}^q \cdot \mathbf{h}_{ij}^e)|_{j=1}^M \quad (2)$$

The final similarity $\mathcal{S}$ is the typical query-node similarity interpolated with the query-edge similarity by $\lambda$, which is then the max query-question similarity out of total $M$ questions. $\mathcal{S}$ among all nodes are then sorted for retrieval ranking, being a zero-shot approach without task-specific training.

#### 5.1.2 Supervised Learning

We then introduce our proposed supervised approach that reranks candidates with augmented node embeddings. Specifically, the enrichment is formulated as an attention, with the user query as *query*, edge questions as both *key* and *value*, such that a new embedding is obtained upon all edge questions conditioned on the query. Let $\mathcal{A}_i$ be the attention scores of the *i*th node, the augmented node embedding $\mathbf{h}_i^a$ is denoted as:

$$\mathcal{A}_i = \text{softmax}\Big(\frac{(\mathbf{h}^q W_Q)(\mathbf{h}_{ij}^e W_K)^T)}{\sqrt{d}}\Big)\ |_{j=1}^M \quad (3)$$

$$\mathbf{h}_i^a = \mathbf{h}_i^v + \mathcal{A}_i\ (\mathbf{h}_{ij}^e W_V)|_{j=1}^M \quad (4)$$

$W_{Q/K/V}$ is the parameter for *query*/*key*/*value* in attention, and $d$ is the *query* dimension size. For a node $v_i$, we provide both outgoing and incoming questions to/from its direct neighbor node for bidirectional contextual information.

With the augmented local embedding for the *i*th node $\mathbf{h}_i^a$, the model is trained with contrastive loss to maximize the similarity between each query $q$ and its positive nodes $P(q)$ among $N$ candidates:

$$\mathcal{L} = \frac{-1}{|P(q)|} \sum_{x \in P(q)} \log \frac{\exp(\mathbf{h}^q \cdot \mathbf{h}_x^a)}{\sum_{y=1}^N \exp(\mathbf{h}^q \cdot \mathbf{h}_y^a)} \quad (5)$$

For inference, the model simply reranks top retrieved candidates from a baseline system, using the new contextualized embedding function.

### 5.2 Experiments

**Data** For experiments situating our purpose, we adapt the data from (Xu et al., 2023) with slight modification. First, we use the available data of *Notre-Dame de Paris* in Chinese for training and evaluation, instead of using all available novels to avoid large-scale graph realization. Second, the original task operates retrieval on sentence-level. Similar to Section 4, we take short snippets as graph nodes, and label positive snippets converted from the original positive sentences. The resulting dataset has 1288 candidate snippets in total, with 29484/1000/510 queries for the train/dev/test split.

**Evaluation Metric** A query may have one to many positive snippets (up to 7). We take the typical information retrieval metric Normalized Discounted Cumulative Gain (NDCG), assigning the same relevance for each positive snippet equally.

**Methods** Four methods are evaluated as follows; all methods adopt BGE-Large encoder[2].

- Zero Shot (ZS): the baseline method that ranks candidates based on query-node similarity.
- ZS+NARCO: our proposed interpolation with query-edge similarity; $\lambda$ is tuned on the dev set.
- Supervised (SU): the baseline model trained supervisedly on queries and candidates only.
- SU+NARCO: our proposed rerank model that utilizes global-contextualized embeddings; the inference reranks top 50 candidates by SU.

### 5.3 Results

Table 3 shows the evaluation results of our settings. Notably, our proposed zero-shot interpolation with

[1] https://huggingface.co/spaces/mteb/leaderboard

[2] https://huggingface.co/BAAI/bge-large-zh-v1.5

| | NDCG | | |
|---|---|---|---|
| | Top-1 | Top-5 | Top-10 |
| Zero Shot | 17.06 | 20.83 | 23.97 |
| +NARCO | **18.82** | **23.83** | **27.37** |
| Supervised | 37.84 | 46.78 | 49.61 |
| +NARCO | **40.20** | **49.00** | **51.33** |

Table 3: Evaluation results on our test set of the plot retrieval task. NDCG is evaluated on the top-1/5/10 retrieved candidates.

query-edge similarity improves upon its baseline on all NDCG metrics, leading 3.4% on Top-10 NDCG ($\lambda = 0.1$), which corroborates the positive information gain from edges for direct node augmentation. The same trend still holds up for the supervised model, improving by a large margin, especially by 2.4% on Top-1 NDCG. Overall, NARCO is shown beneficial towards the acquisition of better local embeddings, demonstrated useful for narrative retrieval with our proposed utilization.

## 6 Study III: Application in QA

Our last study sheds light on graph utility in broader applications, moving beyond the focus of graph edges and nodes themselves. We choose QuALITY (Pang et al., 2022), a multi-choice question answering (QA) dataset on long documents, mostly on fiction stories from Project Gutenberg. With an averaged length of 5k+ tokens per document, we investigate the potentials for retrieval-based approaches, where NARCO may assist to recognize more relevant context, leading to better QA performance benefited from enhanced retrieval.

Specifically, questions in QuALITY were constructed with global evidence in mind, demanding multiple parts in the document to reason upon. In this work, we target the zero-shot QA evaluation, leveraging NARCO to obtain more accurate context from the retrieval process.

**Methods** Retrieval-based approaches are commonly adopted for tackling long context. As experimented by (Pang et al., 2022; Xu et al., 2024), we also split the full document by short snippets and retrieve relevant snippets with regard to the question. We apply the same retrieval process described in Section 5.1.1, where the query-edge similarity is interpolated as in Eq (2) using BGE-Large encoder. The retrieved snippets are then concatenated as the shortened relevant context for subsequent QA.

| | R | ER |
|---|---|---|
| Llama2-7B | 40.97 (± 0.67) | 45.97 (± 0.63) |
| Llama2-70B | 61.56 (± 0.06) | 63.98 (± 0.23) |
| ChatGPT | 63.66 (± 0.06) | **65.92** (± 0.34) |

Table 4: Evaluation results on the dev set of QuALITY: accuracy with standard deviation (from three runs).

| | | | |
|---|---|---|---|
| ChatGPT* | 66.6 | ChatGPT (R) | 70.8 |
| Llama2-70B (R)* | 70.3 | ChatGPT (ER) | **72.8** |

Table 5: Evaluation results on the test set of QuALITY submitted to the ZeroSCROLLS leaderboard. Accuracy of ChatGPT* is provided by the ZeroSCROLLS organizers; Llama2-70B (+R)* is reported by Xu et al. (2024). Settings with +R or +ER are within 1.5k context limit.

**Experiments** We employ Llama2 (Touvron et al., 2023) and ChatGPT for zero-shot inference. As evaluation on the test set requires submission to the leaderboard, we first perform fine-grained analysis on the dev set with short retrieved context (<1k), then submit the final test results using ChatGPT with 1.5k context limit, aligned with Xu et al. (2024) for comparison. Baseline retrieval and our Enhanced retrieval are denoted by **R** and **ER**.

Table 4 & 5 present the evaluation results on the dev set and test set respectively. Results on the dev set suggest that ER enhanced by NARCO can boost QA performance with all LLMs, especially with the smaller 7B model by 5% accuracy, fulfilling our initiative to utilize NARCO in broader applications. The improvement from superior retrieved context is consistent, further confirmed by the 2% margin with ChatGPT on both dev and test set.

## 7 Conclusion

We introduce NARCO, a novel paradigm of narrative representations using a graph structure composed of snippet nodes connected by their coherence dependencies. The edges are formulated as retrospective questions that find background information from prior snippets to enhance comprehension of the current snippet. To realize this concept without human annotations, we propose a two-stage LLM prompting approach to generate these questions. NARCO facilitates narrative understanding by offering informative coherence relationships between snippets and enriched snippet embeddings with global context, validated by positive results on recap identification and plot retrieval tasks, as well as a downstream question answering task.

## Limitations

While we have demonstrated the usefulness of our proposed NARCO, upon manually verifying the generated edge questions, deficiencies do exist in the current graph generation approach:

- The generated questions are not free from noises, as mentioned in Section 3. One common scenario occurs when pairs of context chunks are irrelevant to each other. GPT-4 struggles to accurately identify irrelevancy, leading it to ask questions that lack informativeness.
- Our approach does not handle the scenario where there is joint dependency among three or more chunks. As we generate questions upon pairs, sometimes the key connecting information exists in the third chunk and is missing, preventing the recognition and formulation of useful questions.

Despite the aforementioned issues, our graph still proves beneficial in various applications. This is partly due to the fact that Large Language Models (LLMs) and our learned models possess the capability to automatically discern which information to utilize. Still, enhancing the quality of questions could further augment the benefits derived from our graph, highlighting the potentials of our proposed representation of narrative context.

An additional limitation lies in our filtering algorithm. For LLMs that struggle with following instructions accurately, the current filtering strategy may prove inadequate. For instance, if an LLM repeatedly poses questions that could be understood and answered solely by referring to prior texts, our filtering process is inefficiency to rule out these questions. One potential solution to mitigate this issue could involve implementing a matching model between the questions and the target texts. However, since our work employs GPT-4 alongside Chain-of-Thought, which effectively reduces such instances of shortcut-taking, we have opted to retain the current strategy. We acknowledge the possibility of exploring alternative LLMs with more sophisticated filtering strategies in future work.

## References


Anton Benz and Katja Jasinskaja. 2017. Questions under discussion: From sentence to discourse. *Discourse Processes*, 54:177–186.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023a. Walking down the memory maze: Beyond context limit through interactive reading.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. LongloRA: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3829–3846, Singapore. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Kordula De Kuthy, Madeeswaran Kannan, Haemanth Santhi Ponnusamy, and Detmar Meurers. 2020. Towards automatically generating questions under discussion to link information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5786–5798, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. QUD-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Arthur Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review*, 101:371–95.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada. Association for Computational Linguistics.

Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6544–6555, Online. Association for Computational Linguistics.

Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022. Discourse comprehension: A question answering framework to represent sentence connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11752–11764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2023. Discourse analysis via questions and answers: Parsing dependency structures of questions under discussion. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11181–11195, Toronto, Canada. Association for Computational Linguistics.

Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of Linguistics*, 31(1):109–147.

Jiangnan Li, Qiujing Wang, Liyan Xu, Wenjie Pang, Mo Yu, Zheng Lin, Weiping Wang, and Jie Zhou. 2024. Previously on the stories: Recap snippet identification for story reading.

William Mann and Sandra Thompson. 1988. Rethorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A question answering framework for decontextualizing user-facing snippets from scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3194–3212, Singapore. Association for Computational Linguistics.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Dongqi Pu, Yifan Wang, and Vera Demberg. 2023. Incorporating distributions of discourse structure for long document abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, Toronto, Canada. Association for Computational Linguistics.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Journal of Heuristics - HEURISTICS*, 49.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Hayoung Song, Bo-Yong Park, Hyunjin Park, and Won Shim. 2020. Cognitive and neural state dynamics of story comprehension. *Journal of Neuroscience*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-

bog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Tom Trabasso and Linda L Sperry. 1985. Causal relatedness and importance of story events. *Journal of Memory and Language*, 24(5):595–611.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: TED talks and the questions they evoke. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1118–1127, Marseille, France. European Language Resources Association.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023a. QUDeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363, Singapore. Association for Computational Linguistics.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023b. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing transformers. In *International Conference on Learning Representations*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. Effective long-context scaling of foundation models.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*.

Shicheng Xu, Liang Pang, Jiangnan Li, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2023. Plot retrieval as an assessment of abstract semantic association.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: transformers for longer sequences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

# A Graph Realization

## A.1 Full Prompts and Details

Full prompts of the two-stage LLM prompting (Section 3) are provided in Figure 3-5. We specify the maximum number of generated questions for a node pair as 4 in the prompt.

For the task of plot retrieval (Section 5) and long context QA (Section 6), we construct edges within a neighboring window of 4 preceding nodes, such that the graph realization is proportional to the input instead of being quadratic. For recap identification (Li et al., 2024), edges are obtained on the provided preceding snippets.

For a context with $T$k tokens, it takes approximately $6T$k tokens to obtain all edge questions of NARCO using GPT-4, which costs $\$0.06T$ as of this writing.

## A.2 Quantitative Examples

Examples of generated questions on *Game of Thrones* from recap identification (Li et al., 2024).

### A.2.1 Case1

**Current Context:**

*However, Oberyn stands too close to his seemingly defeated opponent, and Gregor manages to trip and seize him. Berserk with fury, Gregor grabs Oberyn by the throat and lifts him off the ground, smashing out most of his teeth with a single devastating punch. Climbing on top of Oberyn, Gregor finally admits for all to hear that he raped and killed Elia as he gouges out Oberyn's eyes with his thumbs before crushing the Viper's skull between his hands, which he proclaims having done the same to his sister. As Ellaria screams in horror, a stunned silence sweeps over the crowd. The short joyful moments for Tyrion and Jaime are shattered, as Tywin stands and proclaims the will of the gods is clear: Tyrion is guilty and sentenced to death. Tyrion cannot even reply, shockingly staring in catatonic astonishment at Oberyn's skull-crushed corpse, as does Jaime; the only different reaction is from Cersei, who stares at Oberyn's slaughtered body, listening to Tyrion's death sentence while smirking in vindication.*

**Prior Context:**

*Having received word of the wildlings' raids down south, the Lord Commander states that they do not have the manpower to afford venturing away from the Wall. They are interrupted when Edd and Grenn return to Castle Black after escaping Craster's Keep. Jon reveals he told Mance Rayder that a thousand men armed Castle Black and therefore points out that when Mance reaches Craster's Keep, Rast and Karl Tanner will not hesitate in revealing the truth. Jon then insists the Night's Watch send a party to Craster's Keep to kill their traitor brothers before Mance gets to them first.*

**Generated Question (Valid):**

*What prompted the Night's Watch to act with urgency in sending a party to Craster's Keep to eliminate the traitors?*

**Generated Question (Invalid):**

*What was the reason behind Jon Snow's insistence on a strategic assault to silence the traitors before a specific event could occur?*

(Note: it is a question asked upon the prior context and can be answered by it directly, as addressed in the Limitations Section, not bridging two context.)

### A.2.2 Case2

**Current Context:**

*In what becomes known as the infamous Red Wedding, Lothar draws a knife and repeatedly stabs the pregnant Talisa in the stomach, killing her unborn child. Talisa collapses to the ground as chaos surrounds. Before he can react, Robb is shot by the musicians with crossbows several times and falls to the floor. Numerous other Stark men are killed by the crossbow bolts or set upon by Frey soldiers. Catelyn is shot by one of the musicians in the back and falls to the floor.*

**Prior Context:**

*In Gendry's quarters, Melisandre seduces Gendry long enough to distract him, then promptly ties him to the bed and places leeches on his body. She explains as Stannis and Davos enter the room that Davos wanted a demonstration of the power in king's blood, then removes the leeches and lights a fire in a nearby brazier. As part of the magical ritual that follows, Stannis throws the leeches into the flames at Melisandre's direction, and recites the names of three people he wants dead as they burn: "The usurper Robb Stark, the usurper Balon Greyjoy, the usurper Joffrey Baratheon."*

**Generated Question (Valid):**

*What ritual was performed prior to the Red Wedding that sought the death of Robb Stark and might have influenced his fate?*

(Note: it is an open question whether Melisandre's ritual really worked and is widely discussed among fans. The question uses *might* which adds its accuracy.)

### A.2.3 Case3

**Current Context:**

*In King's Landing, Eddard is summoned to the throne room by "King Joffrey"; Robert has died. He arrives to find Littlefinger and Varys waiting for him, along with Commander Janos Slynt and a detachment of the City Watch. Varys tells him that Renly has fled the city, along with Ser Loras Tyrell and a number of retainers. They were last seen heading south. The party enters the throne room, where Joffrey sits on the Iron Throne. He demands oaths of fealty from his councilors and subjects. Instead, Eddard gives Ser Barristan Selmy the proclamation naming him as Lord Protector of the Realm. To Barristan's shock, Cersei takes the "paper shield" and tears it up. Instead, she suggests that Eddard bend the knee and swear allegiance.*

**Prior Context:**

*Lord Eddard Stark meets with Cersei Lannister. He tells her that he knows the secret that Jon Arryn died for: that Cersei's three children are not Robert's, but the product of incest between her and Jaime. Cersei does not deny the charge and in fact is proud of it, comparing their love to the old Targaryen practice of marrying brother to sister; she also admits to having despised Robert ever since their wedding night, when Robert drunkenly stumbled into Cersei's bed and called her "Lyanna". Eddard angrily tells her to take her children and leave the city immediately. When Robert returns from his hunt, he will tell him the truth of the matter and Cersei*

*should run as far as she can before that happens, lest Robert's wrath find her.*

**Generated Question (Valid):**

*What is the reason behind Eddard Stark's refusal to swear fealty to Joffrey and his decision to present a proclamation in the throne room?*

**Generated Question (Invalid):**

*What prevented Eddard Stark from informing King Robert about the illegitimacy of Cersei's children, which could have significantly altered the succession to the Iron Throne?*

#### A.2.4 Case4

**Current Context:**

*In what becomes known as the infamous Red Wedding, Lothar draws a knife and repeatedly stabs the pregnant Talisa in the stomach, killing her unborn child. Talisa collapses to the ground as chaos surrounds. Before he can react, Robb is shot by the musicians with crossbows several times and falls to the floor. Numerous other Stark men are killed by the crossbow bolts or set upon by Frey soldiers. Catelyn is shot by one of the musicians in the back and falls to the floor.*

**Prior Context:**

*At Harrenhal, Jaime speaks one last time to Brienne before he leaves. Jaime remarks that he owes Brienne a debt for both keeping him alive on their journey and for giving him a reason to live to rouse him from his suicidal depression after losing his hand. Brienne tells Jaime to repay his debt by keeping his pledge. Jaime promises that he will keep his word to return Catelyn Stark's daughters to her.*

**Generated Question (Invalid):**

*What prior commitment made by Jaime Lannister could influence the fate of the Stark family following the Red Wedding, where Catelyn Stark is among those attacked?*

(Note: the question is rather irrelevant in regards to the two context snippets.)

#### A.2.5 Case5

**Current Context:**

*Tormund and Beric Dondarrion review the defenses atop the Wall at Eastwatch-by-the-Sea. Tormund remarks that the crows say he'll get used to the height, but he admits it'll probably be a while. Suddenly, the pair sees movement at the edge of the Haunted Forest. A White Walker emerges atop an undead horse, followed shortly by a horde of wights. More and more White Walkers emerge as the Night Watch's horns sound three times. However, the army of the dead stops some distance from the foot of the Wall and Tormund looks relieved; despite their numbers, the dead don't have anything that could possibly get them past the barrier. But then all on the Wall stop in horror as they hear a very familiar sound; a screeching roar mixed with the heavy thumping of huge wings beating the air.*

**Prior Context:**

*At Eastwatch, Sandor carries the struggling Wight into a boat. Tormund and Beric tell him they will meet again but Sandor retorts he hopes not. Daenerys sends Drogon and Rhaegal to scour the surrounding mountains for Jon. Jorah tells Daenerys that it is time to leave but she insists on waiting a bit longer. Before she can leave, they hear a horn blowing signaling a rider approaching. Looking down from the battlements, Dany sees a wounded Jon Snow approaching on horseback. Aboard their ship, Davos and Gendry remove the frozen-stiff garments and tend to Jon Snow, who has suffered severe hypothermia and several minor injuries. Daenerys also notes the massive scars on his chest from his previous fatal wounds.*

**Geneated Question (Invalid):**

*What was Daenerys waiting for at Eastwatch before Jon Snow's wounded arrival on horseback?*

(Note: this is another example of asking upon the prior context, which could happen more often than irrelevant questions.)

### A.3 Experiments

The usage of ChatGPT and GPT-4 is through OpenAI's paid API service. For inference with open-sourced LLMs such as Llama2 (Touvron et al., 2023), we conduct experiments on Nvidia A100 GPUs. For training a rerank model in Section 5, we perform training on one A100 GPU, which takes around 6 hours to finish, with 20 epochs, learning rate $2 \times 10^{-5}$, and a warmup ratio of $5 \times 10^{-2}$.

You are an expert on reading and analyzing a wide variety of books. Given the following two snippets [snippet_a] and [snippet_b] from a book, where [snippet_a] happens before [snippet_b], you need to find concrete parts in both snippets that reflect this temporal relation, such that certain parts in [snippet_a] contribute as the preceding background or cause for specific events or situations in [snippet_b].

[snippet_a]
...

[snippet_b]
...

Please try your best to provide a brief markdown list of each important point that contains those specific parts from both snippets and briefly explains how one serves as the background or cause for the other so to reflect their temporal or causal relation (no more than four points in total). Note that only list evident and important points without much guessing; it is ok to find only one, or even no such points.

Figure 3: Prompt for Question Generation (turn 1).

Please convert each of your listed point to the form of question, such that each question asks about the cause or background (rather than outcome or consequence) of specific events or situations mentioned in [snippet_b], which can be answered or clarified by the corresponding part in the preceding [snippet_a]. Hence, these questions should be helpful to reflect their temporal or causal or other important relations between the two snippets. Note that the question should ask upon specific things from [snippet_b] that cannot be answered by [snippet_b] itself, and should be answerable by concrete parts from [snippet_a] without disclosing those parts directly in the question.

Please try your best to think of one such question for each listed point; for your response, return each question starting with "Q:". Questions should be asked directly without mentioning "snippet" or any other explanation; questions should be concise but also provide necessary context to avoid ambiguity.

Figure 4: Prompt for Question Generation (turn 2).

You are an expert on reading and analyzing a wide variety of books. Given the following snippet [snippet] from a book, and a related question [question], you need to determine whether the provided snippet could answer this question.

[snippet]
...

[question]
...

Please first reason the question very briefly, then give the judgement. If the provided snippet does not present useful information to answer the question, print [UNANSWERABLE] after the reasoning and terminate your response. Otherwise, if the question is indeed answerable, print [ANSWERABLE] after the reasoning, immediately followed by a concise markdown list of the most crucial original sentences from the snippet that could serve as the key supporting evidence for the answer of the question; directly show each sentence per line, without any extra explanation.

Figure 5: Prompt for Question Filtering via back verification.