# Fine-Grained Modeling of Narrative Context: A Coherence Perspective via Retrospective Questions

**Liyan Xu** **Jiangnan Li** **Mo Yu*** **Jie Zhou**
Pattern Recognition Center, WeChat AI
{liyanlxu,jiangnanli,withtomzhou}@tencent.com moyumyu@global.tencent.com

## Abstract

This work introduces an original and practical paradigm for narrative comprehension, stemming from the characteristics that individual passages within narratives tend to be more cohesively related than isolated. Complementary to the common end-to-end paradigm, we propose a fine-grained modeling of narrative context, by formulating a graph dubbed NARCO, which explicitly depicts task-agnostic coherence dependencies that are ready to be consumed by various downstream tasks. In particular, edges in NARCO encompass free-form retrospective questions between context snippets, inspired by human cognitive perception that constantly reinstates relevant events from prior context. Importantly, our graph formalism is practically instantiated by LLMs without human annotations, through our designed two-stage prompting scheme. To examine the graph properties and its utility, we conduct three studies in narratives, each from a unique angle: edge relation efficacy, local context enrichment, and broader application in QA. All tasks could benefit from the explicit coherence captured by NARCO.

## 1 Introduction

Since the advent of Large Language Models (LLMs), document comprehension has been improved significantly by simply employing the end-to-end generative paradigm. Especially, with long context window enabled via techniques such as position interpolation (Xiong et al., 2023; Peng et al., 2024), cached or efficient attention (Wang et al., 2023; Ge et al., 2024a; Munkhdalai et al., 2024), context compression or pruning (Chevalier et al., 2023; Anagnostidis et al., 2023; Ge et al., 2024b), the end-to-end paradigm is deemed undoubtedly simple and effective for comprehension tasks (e.g. question answering) on various documents.

However, while the typical benchmarks are continually enhanced by more advanced LLMs (Touvron et al., 2023; Jiang et al., 2024; OpenAI et al., 2024), the end-to-end paradigm may not suffice for all comprehension scenarios. In this work, we focus around the narrative context, i.e. stories or novels, and propose a conceptually original framework of fine-grained context modeling: a graph is formulated that depicts the relations between context snippets, abstracting over the context to reflect a high-level understanding of the narrative. The graph itself is practically realized by LLMs to harness their rapidly evolving strengths, and the resulting graph could serve to facilitate various downstream narrative comprehension tasks.

Our motivation arises from the distinctive nature of narratives: multiple development of characters or events in a story could be entangled over long context ranges, where each local passage usually serves specific purposes for others. Thus, individual passages tend to be cohesively interconnected than being isolated. As the end-to-end paradigm implicitly grasps these context connections through sequence modeling, our approach explicitly models these dependency relations to capture coherence, offering a directly-applicable alternative path orthogonal to the end-to-end paradigm.

Concretely, drawing inspiration from the cognitive process on narrative perception, whereas humans can constantly reinstate relevant or causal events from past context during reading (Trabasso and Sperry, 1985; Graesser et al., 1994), our formalism, termed NARrative COgnition graph (NARCO), splits the entire context into chunks that act as graph nodes, with edges representing the relations between node pairs. In particular, edge relations are constituted by free-form questions. As humans could relate to past context in retrospect, accordingly, each question in NARCO edges arises from the succeeding node (*latter* context), asking necessary background or causes that can be clarified by the preceding node (*prior* context). Hence, graph edges consist of inquisitive questions that naturally

*Corresponding author.



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5822–5838

reflect retrospection. Overall, the resulting graph explicitly depicts task-agnostic understanding of fine-grained coherence flow that could be flexibly utilized by downstream tasks.

Though our graph formulation partially shares motivations with discourse parsing that characterizes how each proposition relates to others within a close context (Grosz and Sidner, 1986), our method targets on practical utility for narrative comprehension, where edges in NARCO are designed to be easily obtained and effectively consumed by downstream tasks. Consequently, NARCO is formulated in a different scope from discourse parsing by two main perspectives. First, as most discourse frameworks, such as Rhetorical Structure Theory (Mann and Thompson, 1988), Penn Discourse Treebank (Prasad et al., 2008), or the recent Questions Under Discussion (QUD) (Ko et al., 2022, 2023) are rooted upon linguistic principles, their relation types are oriented for formal discourse analysis, requiring trained experts to annotate edges according to a defined linguistic taxonomy. Whilst for NARCO, the relation space is larger without taxonomy constraints, offering diverse high-level semantic signals for narrative tasks. Second, NARCO practically leverages LLMs to derive edge relations, without reliance on human annotations. Thus, the edge quality is not restricted by annotation resources, and shall be continuously enhanced along with the ongoing LLM advancement.

The key difficulty of NARCO lies in the edge realization between two nodes, which itself demands strong context understanding to determine which aspects to inquire upon the context, and to assess their saliency for comprehension. Such process is especially strenuous due to the large hypothesis space compared to conventional discourse formalisms. To this end, we pose soft semantic constraints on relations, and employ LLM's capabilities to construct edges automatically through our proposed prompting scheme, of which consists a question generation stage and a self verification stage (Section 3). The obtained edges could then be utilized by downstream tasks in two primary ways. First, edges themselves directly provide information flow to guide various comprehension tasks. Second, they offer global coherence view for each node, thereby augmenting the local context to deepen the digest of independent passages.

To empirically demonstrate the practical utility of NARCO, we present three studies on narrative comprehension tasks, each from a unique angle:

• Our first study examines the **edge efficacy** on *whether the relation questions capture capable retrospective coherence* (Section 4). We conduct experiments on the recap identification task (Li et al., 2024), where NARCO is shown to recognize coherence dependencies between context, boosting up to 4.7 F1 over the GPT-4 baseline.

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

• Lastly, we utilize NARCO in a long document question answering task (Section 6), as a broader application of **Retrieval-Augmented Generation** (RAG) (Lewis et al., 2020). Experiments on QuALITY that requires global context 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 identify more relevant context through edge relations.

Overall, our key contributions in this work are:

- We propose a new paradigm of fine-grained context modeling to facilitate narrative comprehension, orthogonal to the end-to-end paradigm.
- Our introduced NARCO framework describes flexible relations of context dependencies by retrospective questions, which are realized by LLMs through our designed prompting scheme, without reliance on human annotations.
- We present three studies effectively utilizing NARCO on narrative tasks, empirically examining its edge properties and broader utilization.

## 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). QUD has been adapted by recent works for discourse analysis (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 also adopts question-form relations; though, the scope and motivation is different from discourse analysis. Consequently, NARCO differs from QUD works considerably on the following design choices.

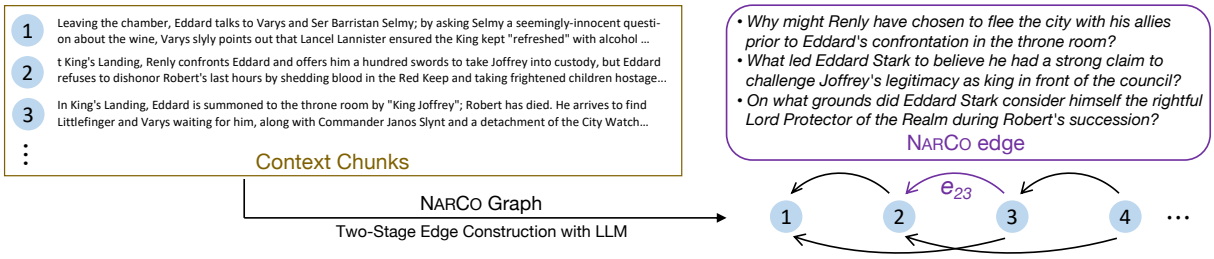


Figure 1: Our proposed NarCo graph described in Section 3, with retrospective questions connecting two nodes.

• **Coarse Granularity** While QUD tends to employ sentences as the basic discourse unit, NarCo opts for a coarser granularity, adopting passages (or chunks) as the 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 context.

• **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 latter context, such that all generated questions in NarCo are naturally grounded by the corresponding prior context.

• **Precision-Focused** Unlike previous QUD works that require dedicated human annotations, NarCo is formulated attainable by LLMs. Accordingly, we prioritize precision over recall for practical instantiation of graph edges, and do not necessitate strict linguistic criteria, as long as edges contribute positively for narrative understanding.

**Narrative Comprehension Assessments** A major task direction on narratives is question answering (QA), where past works have proposed several datasets with human annotations, such as NarrativeQA (Kočiský et al., 2018), TellMeWhy (Lal et al., 2021), FairytaleQA (Xu et al., 2022b), QuALITY (Pang et al., 2022). We adopt QuALITY as the broader application in this work, due to its challenging long context, requirement of global evidences, and simple evaluation by multi-choices.

Recently, several tasks have emerged focusing on modeling the reading process of long narratives, including TVShowGuess (Sang et al., 2022), PerSoNet (Yu et al., 2023), ToM-in-AMC (Yu et al., 2024), and retrieval tasks such as RELiC (Thai et al., 2022), PlotRetrieval (Xu et al., 2023b). These tasks require a holistic understanding of the long narratives to enhance contextual comprehension of specific segments. We reckon the significance of explicitly modeling context dependencies as a crucial aspect of narrative comprehension, motivating the inception of .

**LLM Understanding and Reasoning** LLMs have demonstrated remarkable capabilities on a wide spectrum of comprehension and reasoning tasks (Chen et al., 2024; Sun et al., 2024). The simple end-to-end solution is especially appealing with long context window enabled, using techniques such as scaling positional embeddings (Chen et al., 2023b; Xiong et al., 2023; Peng et al., 2024), efficient attention (Munkhdalai et al., 2024), cached attention (Wang et al., 2023; Ge et al., 2024a), recurrent attention (Dai et al., 2019), context compression (Chevalier et al., 2023; Ge et al., 2024b), context pruning (Anagnostidis et al., 2023), etc. Though being effective, certain narrative tasks demand beyond the end-to-end solution. Recently, new methods have been proposed for fine-grained task processing, e.g. reading agents such as MemWalker (Chen et al., 2023a). Nevertheless, our proposed approach depicts explicit context dependencies as an alternative paradigm, which is orthogonal to the existing directions and could be even further combined.

**Structured Representation** Various relational structures in text documents has attracted much attention by previous works, such as syntactic relations (Strubell et al., 2018; Xu et al., 2022a), discourse relations (Ji and Smith, 2017; Nair et al., 2023; Hu and Wan, 2023), entity or event relations (Ding et al., 2019; Li et al., 2020, 2021; Xu and Choi, 2020, 2022; Nguyen et al., 2022). As all these structures encompass pre-defined taxonomies on edge types, our propose graph representation is motivated to comprise open-world edge types that have been practiced in other tasks (Wu et al., 2019; Xu et al., 2023a; Su et al., 2024), while being practical and attainable by LLMs without requiring efforts of human annotations.

## 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 specific graph realization using LLMs, without 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 indicates the relations between the context. These relations are articulated as free-form 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 context $v_i$. Since the hypothesis space is huge without any regularization, we pose soft semantic constraints on questions, such that questions should primarily reflect on causal and temporal relations, which are significant to the narrative coherence.

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 coherent comprehension of the global context (Trabasso and Sperry, 1985; Graesser et al., 1994; Song et al., 2020). Unlike conventional QUD that features curiosity-driven questions in a forward direction, which could yield unanswerable questions, all edge in NARCO are fully grounded by the prior context, since all retrospective questions are addressable by prior nodes.

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

- It may have zero or many questions. An empty edge (zero questions) signifies $v_j$ is vaguely independent from $v_i$ in terms of coherence.
- Each question should be salient towards the comprehension of narrative development, rather than inquiring trivial details. Hence, the number of questions in an edge should reflect how cohesively related between two nodes.
- As we adopt coarse granularity for nodes, questions could probe higher-level relations based on the extrapolation over multiple sentences, which may be useful towards broader understanding.

### 3.2 Graph Realization

To obtain graph nodes, the full context is split by paragraph and sentence boundaries. 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 $O(N^2)$ full 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 is sufficient to obtain relation dependencies from neighboring nodes within a context window.

Despite the daunting formulation on edge relations, the emergence of LLMs presents an opportunity: through designed 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 to tackle the challenging edge construction.

**Question Generation** For an edge $e_{ij}$ to be instantiated, LLMs need 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 a 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 suggest that although LLMs can generate plausible questions by following the instructions, their quality is often unsatisfactory for NARCO, with common errors as follows (examples in Appx A.2):

- *Self-answerable*: LLMs often ask questions upon $v_j$ but also answerable by $v_j$ as well. Such self-answerable pattern aligns with the more conven-

tional QG setting (Du et al., 2017) that may exist plentifully during the supervised finetuning of LLMs, causing a bias towards this type of question. However, they are not desirable for NARCO, since they do not express dependencies between nodes to reflect their relations.

- *Hallucination*: LLMs could hallucinate the relations of two nodes 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 is empirically challenging even for strong LLMs (e.g. GPT-4), we perform QG with two heuristic turns that can be viewed as human-guided Chain-of-Thoughts (Wei et al., 2022):

> 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 above 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.

The designed two-turn QG scheme yields higher-quality questions than the rudimentary generation, mainly alleviating the self-answerable problem. However, noisy questions of the two identified error types still occur due to imperfect instruction following by LLMs. In light of these noises, we apply an optional second stage to filter out noisy questions through self verification.

**Self Verification** The second stage takes the generated questions from QG and in turn, performs question answering on the context:

> Given a context $\mathcal{C}_{ij}$ and a related question, determine whether it is answerable. If yes, reason the answer and provide original sentences of key supporting evidences.

In Particular, $\mathcal{C}_{ij}$ 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 the prior context $v_i$. If not, the question is attested noisy and gets discarded, as it does not bridge the two context effectively.

With the second stage, only questions that could be answered by prior nodes are eventually retained in NARCO, being a precision-focused approach. In this work, we adopt GPT-4 for the challenging QG stage, and ChatGPT for the easier verification stage. NARCO may also be derived with strong open-source LLMs as well. Our full prompts and more details are provided in Appx A.1.

As NARCO targets the practical utility to facilitate narrative comprehension, the obtained edges shall be directly consumed by downstream tasks. Sections 4-6 present three empirical studies, each from a distinctive perspective, to examine the edge properties and their utilization.

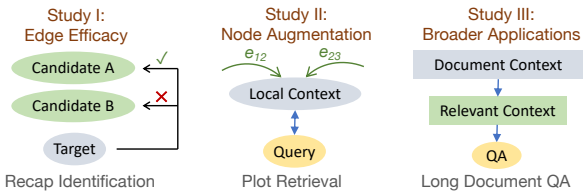


Figure 2: Three presented studies leveraging NARCO.

## 4 Study I: Edge Efficacy

Our first study examines the graph edges on whether they express useful relations, such that the generated retrospective questions should bridge the coherence between two context. For appropriate assessment, we conduct 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. The 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 designed to capture the coherence relations between context, 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 text content.

With NARCO, we regard each current snippet as a **t**arget graph node $v_t$, and the list of its $N$ pre-

ceding snippets $\{v_c|c = 1, .., N\}$ as the **c**andidates. For $v_t$ and each of its candidate $v_c$, the edge is realized as $e_{ct}$. As questions in $e_{ct}$ should reflect causal or temporal coherence, we directly utilize these questions in two following ways.

**Edge Relations** Normally, candidate snippets are processed by their text content as in the baseline. To evaluate the coherence depicted by edges, we instead propose to identify recap snippets solely based on the edge relations: for a candidate node $v_c$, we concatenate all its questions in edge $e_{ct}$, denoted by $q_c$, to identify recap, and completely neglect original text content, so to ensure an entirely isolated assessment of edge relations.

Specifically, given the context of a target snippet $v_t$, and $N$ candidates $\{q_c|c = 1, .., N\}$ represented by questions, we now ask a LLM to score which $q_c$ addresses important questions that are significant to comprehend the current context, with higher scores indicating better overall questions that provide recap information. Candidates with empty edges are directly assigned 0 score.

**Edge Degrees** Alternatively, as mentioned in Section 3.1, the number of questions between two nodes could suggest how cohesively related they are. We take this number as the edge degree, and propose to simply deem it as the score to rank candidates, without any inference on the node context or edge relations at all. Though being rather unconventional, ranking candidates by edge degrees further reflects the edge quality.

With either the relation score or degree score, it can 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 $r_i^{deg}$, along with binary selection $b_i$ from the original 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 decision $b_i$ 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 $\alpha/\beta/\lambda$ to 0 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.

| | 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.

### 4.3 Results

Table 1 shows the zero-shot evaluation results on the test set of RECIDENT. Notably, the interpolation with NARCO edges (`Full`) consistently brings 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 NARCO still advances 3.5 / 4.7 F1 upon `BL` on NDP / GOT as well.

Moreover, selection solely based on edge relations (`Rel`) obtains comparable or better performance than the baseline, with the only exception on GOT with ChatGPT. Overall, Table 1 effectively demonstrates the edge efficacy of NARCO that expresses coherence through retrospective questions.

For in-depth analysis, 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 **F**iltering by self verification.

| | 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$^{-F}$).

Table 2 shows the additional evaluation results, 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.

For Full$^{-F}$, the degradation is trivial from Full. It is also expected, as the LLM scoring on relations is based on the presence of "good" questions that reflect recap information, which should be retained by the verification stage. Thus, our approach with NarCo is shown robust against noisy questions.

### 4.4 Graph Insights

The majority of generated questions in NarCo are *what*/*why*/*how*-type of questions. Their ratios are provided in Table 3, along with the averaged number of questions per edge before / after the self verification stage (Section 3.2).

| | NDP | GOT |
|---|---|---|
| *What*-Questions Ratio | 61.5% | 58.4% |
| *Why*-Questions Ratio | 26.5% | 25.2% |
| *How*-Questions Ratio | 7.8% | 14.0% |
| # Questions per Edge | 3.4 | 3.5 |
| + Self Verification | 1.9 | 2.0 |

Table 3: Statistics of NarCo in Study I (Section 4).

## 5 Study II: Node Augmentation

Our second study underscores the NarCo utility of local context augmentation, examining whether the graph typology could enrich 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 tends to be more interconnected with others than isolated.

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., 2023b), 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, analogous 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 regard all candidate snippets as graph nodes to be retrieved from, and derive NarCo edges of neighboring nodes. Our proposed method focuses on fusing edge questions into node representation for enhanced retrieval.

Xu et al. (2023b) 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 as of this writing, 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 and supervised settings.

#### 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

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

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 node embedding is obtained attending its edge questions conditioned on the query. Let $\mathcal{A}_i$ be the attention scores of the $i$th candidate 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 embedding for the $i$th node $\mathbf{h}_i^a$, the model simply reranks top retrieved candidates from a baseline system. It is trained with the supervised contrastive loss (Khosla et al., 2020) to maximize the similarity between each query $q$ and its positive targets $P(q)$ among $N$ in-batch candidates (details in Appx A.3):

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

### 5.2 Experiments

**Data** For experiments situating our purpose, we adapt the data from (Xu et al., 2023b) 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 or 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 zero-shot method that ranks candidates based on the query-node similarity.
- ZS+NARCO: our proposed interpolation with query-edge similarity; $\lambda$ is tuned on the dev set.
- Supervised (SU): the baseline supervised model without leveraging NARCO.
- SU+NARCO: our proposed rerank model that utilizes the global-contextualized embeddings; the inference reranks top 50 candidates by SU.

| | nDCG | | |
|---|---|---|---|
| | @1 | @5 | @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 4: Evaluation results of zero-shot and supervised settings on our test set of the plot retrieval task. nDCG is evaluated on the top-1/5/10 retrieved candidates.

### 5.3 Results

Table 4 shows the evaluation results of the four settings. Notably, our proposed zero-shot interpolation with query-edge similarity improves upon its baseline on all nDCG metrics, leading 3.4% on nDCG@10 ($\lambda = 0.1$), confirming the positive information gain from edges that contribute useful contextual information. The same trend still holds up for the supervised model that learns enriched embeddings leveraging edge relations, especially by the 2.4% improvement on nDCG@1.

Overall, NARCO is shown helpful towards the acquisition of better local representation, through the explicit relational dependencies beyond local context. The empirical results advocate the direc-

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

tion of fine-grained context modeling, which could foster a more nuanced comprehension.

## 6 Study III: Broader Application

Our last study sheds light on the potentials of graph utility in broader applications. As a first step towards this new direction, in this work, we evaluate with Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) in the task of long document question answering. Experiments are conducted on QuALITY (Pang et al., 2022), a multi-choice QA dataset on narrative documents, mostly being fiction stories from Project Gutenberg. With an averaged length of 5k+ tokens per document, we adopt the retrieval-based approaches, where relevant snippets conditioned on the question are retrieved first, then fed to a LLM to generate answers, following a standard RAG paradigm.

Especially, QuALITY was constructed with global evidences in mind: questions may require multiple parts in the document to reason upon. Therefore, NARCO may assist to recognize more relevant snippets through the extracted relations across the narrative context, leading to improved QA performance benefited from enhanced retrieval.

**Methods** Retrieval-based approaches are commonly adopted for tackling long context, which have been evaluated on QuALITY by previous works (Pang et al., 2022; Xu et al., 2024; Sarthi et al., 2024). Following these setup, we split the full document by short snippets, and retrieve relevant snippets with regard to the question, which are then concatenated as the shortened context for subsequent zero-shot QA inference by LLMs. To leverage NARCO, we apply the same retrieval process described in Section 5.1.1 to identify relevant snippets, where the query-edge similarity is interpolated as in Eq (2) using BGE-Large encoder.

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

Table 5 & 6 present the evaluation results on the dev set and test set respectively. Results on the dev set suggest that ER can boost QA performance with all LLMs, especially with the smaller 7B model by 5% accuracy, fulfilling our initiative to effectively utilize NARCO in broader applications. The improvement from enhanced context retrieval is consistent, further confirmed by the 2% leading margin with ChatGPT on both the dev and test set.

Having demonstrated that NARCO can improve RAG in narratives through enhanced retrieval, its utility beyond the retrieval process may be further exploited, e.g. potential facilitation on LLM pretraining or inference directly. We leave future research to explore additional integration of fine-grained context modeling.

| | 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 5: Evaluation results on the dev set of QuALITY: accuracy with standard deviation (from three runs). Enhanced Retrieval (ER) improves QA consistently.

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

Table 6: 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). Performance of three retrieval-based experiments are directly comparable (same 1.5k context limit). We exclude another related work RAPTOR (Sarthi et al., 2024), as they use smaller QA models and different context limit, thus not directly comparable.

## 7 Conclusion

We address the distinctive characteristics of narratives, and propose a novel paradigm of fine-grained context modeling, which explicitly captures the inter-connective coherence within narrative context. A graph is thereby formulated, dubbed NARCO, with edges encompassing free-form retrospective questions to depict the relational dependencies. NARCO is practically realized by LLMs via our designed two-stage prompting scheme, leveraging the promising development of LLMs without reliance on human annotations. To examine the graph properties and its utility, three unique studies are conducted, where NARCO is shown to bring empirical improvement on various narrative applications.

## 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


Sotiris Anagnostidis, Dario Pavllo, Luca Biggio, Lorenzo Noci, Aurelien Lucchi, and Thomas Hofmann. 2023. Dynamic context pruning for efficient and interpretable autoregressive transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 65202–65223. Curran Associates, Inc.

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

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

Pei Chen, Boran Han, and Shuai Zhang. 2024. Comm: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico City, Mexico. Association for Computational Linguistics.

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

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.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. *arXiv preprint arXiv:1905.05460*.

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.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. 2024a. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *The Twelfth International Conference on Learning Representations*.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024b. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*.

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.

Xinyu Hu and Xiaojun Wan. 2023. Exploring discourse structure in document-level machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13889–13902, Singapore. Association for Computational Linguistics.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

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.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

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

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

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.

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of EMNLP 2021*, pages 6443–6456.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of EMNLP 2020*, pages 684–695.

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

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. 2024. Leave no context behind: Efficient infinite context transformers with infini-attention.

Inderjeet Nair, Shwetha Somasundaram, Apoorv Saxena, and Koustava Goswami. 2023. Drilling down into the discourse structure with LLMs for long document question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14593–14606, Singapore. Association for Computational Linguistics.

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.

Minh Van Nguyen, Bonan Min, Franck Dernoncourt, and Thien Nguyen. 2022. Joint extraction of entities, relations, and events via modeling inter-instance and inter-label dependencies. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4363–4374, Seattle, United States. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. TVShowGuess: Character comprehension in stories as speaker guessing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4267–4287, Seattle, United States. Association for Computational Linguistics.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. 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*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Zhenlin Su, Liyan Xu, Jin Xu, Jiangnan Li, and Mingdu Huangfu. 2024. Sig: Speaker identification in literature via prompt-based generation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wangtao Sun, Haotian Xu, Xuanqing Yu, Pei Chen, Shizhu He, Jun Zhao, and Kang Liu. 2024. Itd: Large language models can teach themselves induction through deduction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand. Association for Computational Linguistics.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. RELiC: Retrieving evidence for literary claims. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7500–7518, Dublin, Ireland. Association for Computational Linguistics.

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 Molybog, 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.

Ruidong Wu, Yuan Yao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2019. Open relation extraction: Relational knowledge transfer from supervised data to unsupervised data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 219–228, Hong Kong, China. Association for Computational Linguistics.

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.

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.

Liyan Xu and Jinho Choi. 2022. Modeling task interactions in document-level joint entity and relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5409–5416, Seattle, United States. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

Liyan Xu, Chenwei Zhang, Xian Li, Jingbo Shang, and Jinho D. Choi. 2023a. Towards open-world product attribute mining: A lightly-supervised approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12223–12239, Toronto, Canada. Association for Computational Linguistics.

Liyan Xu, Xuchao Zhang, Bo Zong, Yanchi Liu, Wei Cheng, Jingchao Ni, Haifeng Chen, Liang Zhao, and Jinho D. Choi. 2022a. Zero-shot cross-lingual machine reading comprehension via inter-sentence dependency graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11538–11546.

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. 2023b. Plot retrieval as an assessment of abstract semantic association.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022b. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023. Personality understanding of fictional characters during book reading. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14784–14802, Toronto, Canada. Association for Computational Linguistics.

Mo Yu, Qiujing Wang, Shunchi Zhang, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Liyan Xu, Jing Li, Yue Yu, and Jie Zhou. 2024. Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind. In *Forty-first International Conference on Machine Learning*.

# A Graph Realization

## A.1 Full Prompts and Details

Full prompts of our designed two-stage LLM prompting scheme (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 document 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 \$0.03$T$ as of this writing.

## A.2 Qualitative Examples

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

### A.2.1 Case1

**Current Context:**

*At Craster's Keep, Locke scouts the keep for the party of the Night's Watch sent to eliminate the traitors holed up there; in his reconnaissance, Locke finds the hut where Bran Stark, Jojen, Meera and Hodor are being held captive. Reporting back to Jon Snow and the others, Locke tells them that only eleven traitors are present and most of them are drunk and won't prove much of a threat. He also lies about the hut where Jon's brother Bran and his group are being imprisoned, claiming there are only hounds kept inside and that they should keep away from it to prevent the dogs alerting their enemy. Believing Locke, Jon agrees and tells the party they attack at nightfall.*

**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

**LLM** The usage of ChatGPT (*gpt-3.5-turbo*) and GPT-4 (*gpt-4-1106-preview*) is through OpenAI's paid API service. For the open-source Llama-2 (Touvron et al., 2023), we perform inference on Nvidia A100 GPUs.

**Training** For training a rerank model in Section 5, we initialize a BERT model with weights from BGE-Large (Xiao et al., 2023), and use the mean-pooled token embeddings as the sequence representation, following the standard S-BERT setup (Reimers and Gurevych, 2019). The training is conducted on one Nvidia A100 GPU, taking around 6 hours to finish, with 20 epochs, 20 queries within each batch, learning rate $2 \times 10^{-5}$, cosine learning rate schedule, 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). Slots in blue refer to the input texts.

```
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). Slots in blue refer to the input texts.

```
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. Slots in blue refer to the input texts.