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.

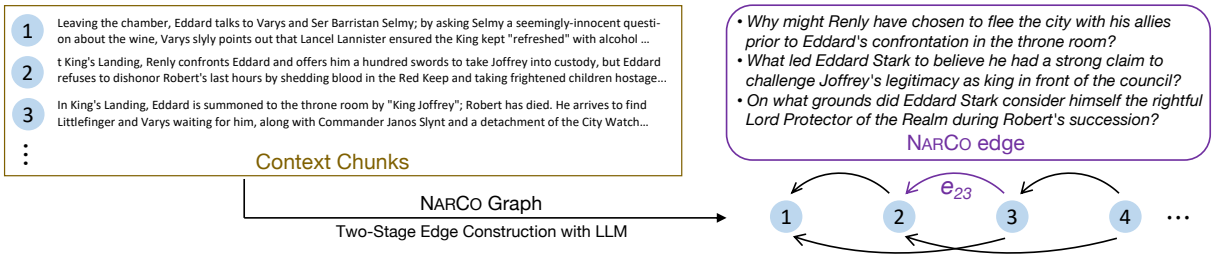

Figure 1: Our proposed NarCo graph described in Section 3, with retrospective questions connecting two nodes.

• **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.

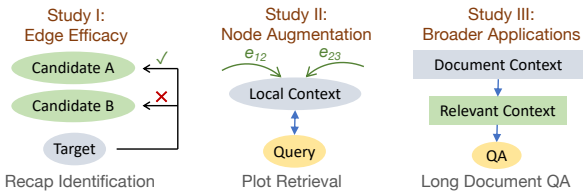

Figure 2: Three presented studies leveraging NARCO.

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 λ rank; relation and degree ranks are interpolated by α and β. 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 ith node, $\mathbf{h}_{ij}^e$ for its jth outgoing questions, $\mathbf{h}^q$ for the query. The interpolated similarity $\mathcal{S}_i$ between the query and ith 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 λ, 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 ith 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 ith 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; λ 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*