NovelQA: A Benchmark for Long-Range Novel Question Answering

Anonymous ACL submission

Abstract

The rapid advancement of Large Language Models (LLMs) has introduced a new frontier in natural language processing, particularly in understanding and processing long-context information. However, the evaluation of these models' long-context abilities remains a challenge due to the limitations of current benchmarks. To address this gap, we introduce NovelQA, a benchmark specifically designed to test the capabilities of LLMs with extended texts. Constructed from English novels, NovelQA offers a unique blend of complexity, length, and narrative coherence, making it an ideal tool for assessing deep textual understanding in LLMs. This paper presents the design and construction of NovelQA, highlighting its manual annotation, and diverse question types. Our evaluation of Long-context LLMs on NovelQA reveals significant insights into the models' performance, particularly emphasizing the challenges they face with multi-hop reasoning, detail-oriented questions, and extremely long input with more than 100,000 tokens. The results underscore the necessity for further advancements in LLMs to improve their long-context comprehension and computational literary studies.

1 Introduction

Recent years have seen a remarkable surge in the development of Large Language Models (LLMs) (OpenAI, 2023a; Touvron et al., 2023). Among these developments, long-context LLMs stand out for their ability to process and interpret extended pieces of text (Tworkowski et al., 2023; Team, 2023; Anthropic, 2023). This capability is essential for complex tasks that require a deep and nuanced understanding of lengthy documents, such as legal cases (Xiao et al., 2021) or academic papers (Groeneveld et al., 2024), where the key is to understand extended narratives (Xu et al., 2023). Meanwhile, the ability to analyze multiple long documents simultaneously is increasingly crucial,

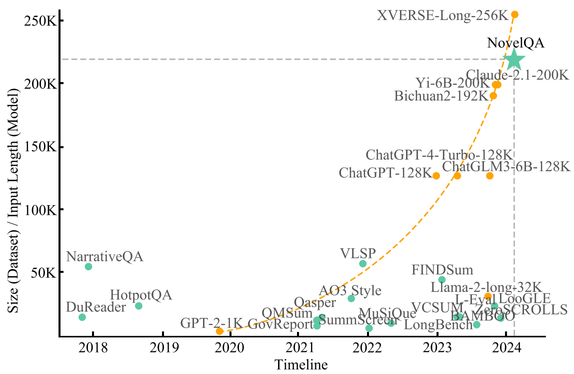


Figure 1: Trend of context window size of LLMs (Orange) and average token length of long-range benchmarks (Green). NovelQA is highlighted with a star.

supporting more informed decision-making in various fields (Deng et al., 2023; Lin et al., 2023)

However, evaluating the long-context capabilities of models presents challenges, as existing benchmarks (Yang et al., 2018; Tay et al., 2021) no longer align with the advanced processing abilities of current LLMs (Anthropic, 2023; Team, 2023; OpenAI, 2023b). The disparity is further highlighted by the increasing context window size of LLMs, which now outpaces the average token lengths found in long-range datasets. This gap is evident, as the most advanced long-context LLMs are capable of processing over 250,000 tokens, a stark contrast to the longest average token length in current benchmarks, which is around 60,000 tokens. This mismatch underscores the need for updated evaluation methods that can accurately reflect the capabilities of current and future LLMs, as illustrated in Figure 1.

To fill this gap, we introduce NovelQA, a benchmark crafted to specifically evaluate LLMs' performance on texts with context windows exceeding 100,000 tokens. Unlike existing benchmarks (Shaham et al., 2023; An et al., 2023; Adams et al., 2024), NovelQA addresses the need for assessing extremely long-context understanding, offering a

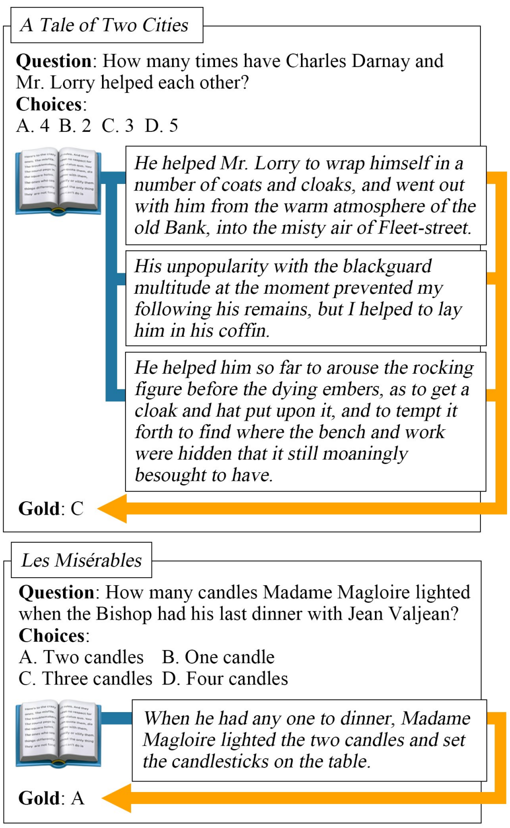


Figure 2: Illustrative Examples from NovelQA: This figure showcases two sample questions. For each question, models are evaluated under two distinct settings – Multichoice, where the task is to select the correct answer from four options, and Generative, where the model generates an answer.

refined and comprehensive tool for advancing natural language processing capabilities.

We construct NovelQA based on novels in English. NIn addition to their value in computaitonal literature ad creativity studies, novels are also ideal for testing long-context modeling because they are long and complex, with plots that are closely linked from start to end. We select novels from various eras, genres, and formats to enhance diversity. The annotation process is performed by a group of skilled annotators, all of whom are holding or pursuing a degree in English Literature and have a strong interest in and familiarity with the novels they annotate. Each question is paired with a 'golden answer' and corresponding textual evidences from the novels. And we categorize them by complexity and aspect for detailed analysis. Figure 2 presents two examples, while Table 8 details the distribution across types, with multi-hop, single-hop, and detailed questions constituting 35.0%, 42.8%, and 22.2% of the dataset, respectively.

In our NovelQA experiments, we assess various long-context LLMs, including commercial models such as GPT-4-128K (OpenAI, 2023b) and Claude 2.1-200K (Anthropic, 2023), alongside open-source models like InternLM2-Chat (Team, 2023). The primary findings offer a significant insight: even the most advanced long-context LLMs face challenges in consistently extracting and processing accurate information from extended texts. For example, GPT-4, the top performer, achieves a 46.88% accuracy rate, whereas the open-source InternLM2-20b achieves 32.37% in a generative setting. This difficulty is particularly apparent in answering multi-hop questions and queries that probe meanings, relationships, spans, and timelines, highlighting a significant gap in the models' long-range comprehension. Moreover, our data shows a decline in performance for evidence situated beyond the 100,000-token mark, including information at the novel's end, diverging from the anticipated *lost-in-middle* phenomenon (Liu et al., 2023). This shift suggests a distinct challenge faced by LLMs when processing texts exceeding 100k tokens in length. These results highlight challenges not only in memory optimization but also in the nuanced comprehension and integration of lengthy texts, indicating a substantial obstacle on the path to truly effective long-context LLMs.

To the best of our knowledge, NovelQA is the first long-context QA benchmark featuring manually crafted questions, golden answers, and evidences, with contexts extending beyond 200,000 tokens. It also is the first long-range QA dataset within computational literary studies. [1]

2 Related Work

Long-Range Benchmarks Evaluating the ability of Long-Context Pretrained Language Models has been a hot topic (Tay et al., 2021; Shaham et al., 2023; Pang et al., 2022). When entering the era of Large Language Models, the context window length has been much longer than ever, and the decoder-only LLMs have been the mainstream of language models and they have faced specific problems, such as the Lost-in-middle issue (Liu et al., 2023). Thus, increasing benchmarks are created for evaluating long-context LLMs. Among those benchmarks, several are representatives. Zero-

[1] We will only make some demonstrations publicly available. For the testset, we plan to launch a leaderboard website and provide an API for evaluation upon acceptance at `RemoveforSubmission`.

SCROLLS (Shaham et al., 2023) and LooGLE (Li et al., 2023a) both emphasize the importance of understanding and aggregating information from long texts, presenting challenges that highlight areas for improvement in current models. L-Eval (An et al., 2023) introduces a suite of tasks with human-labeled query-response pairs to assess LLMs' performance in processing long inputs effectively, using advanced metrics for a more accurate evaluation. Furthermore, benchmarks like LongBench (Bai et al., 2023), BAMBOO (Dong et al., 2023a), and LongBench-Chat (Bai et al., 2024) offer a diverse set of tasks across languages and domains, from reasoning and coding to summarization and multilingual translation. These benchmarks are designed to rigorously test the ability of LLMs to manage extensive contexts, with LongBench-Chat specifically focusing on instruction-following capabilities in long-context interactions. Additionally, LongHealth (Adams et al., 2024) addresses the need for LLMs to interpret long clinical documents accurately, providing a specialized benchmark for evaluating models on medical texts. This focus on domain-specific challenges underscores the broader necessity for LLMs to not only handle long texts but to do so in a manner that is accurate and contextually relevant across various fields.

NovelQA distinguishes from existing benchmarks in three points. Firstly, NovelQA features an average token length exceeding 200,000, far surpassing the tens of thousands typically found in other benchmarks. Secondly, while other datasets often rely on LLMs or existing datasets for content creation, the questions, golden answers, and evidences of NovelQA are entirely crafted through human effort. Lastly, it is the first QA dataset specifically designed for the literature domain.

Long-Context Language Modeling can be divided into several parts including efficient attention, preserving long-term memory like using KV cache, extrapolative positional embedding module, context pre/post-processing (Huang et al., 2023; Dong et al., 2023b). Using efficient attention can reduce computational complexity and memory usage (Beltagy et al., 2020; Wang et al., 2020; Tworkowski et al., 2023; Li et al., 2023b). Preserving KV cache or context-level cache allows models to recall and leverage past information without reprocessing, enhancing coherence over long texts (Chevalier et al., 2023; Wu et al., 2022a,b; Lin et al., 2024; Zhong et al., 2023; Chen et al., 2023a; Hooper et al., 2024). Deploying an extrapolative positional embedding to extend beyond the sequence lengths seen during training (Chen et al., 2023b; Su et al., 2024). Pre/post-processing the context can make models focus on key information (Li et al., 2023c; Jiang et al., 2023; Zhang et al., 2024; Han et al., 2023). Moreover, some training methods are put forward to support further long-context language modeling development (Wang et al., 2024; Press et al., 2021). Since these models are more about exploring long-context language modeling without scaling in data and model size, the focus is on language modeling indicators, such as BLEU and Perplexity, NovelQA can serve as a benchmark for evaluating those long-context language modeling methods.

	Multi-hop	**Single-hop**	**Detailed**	**Sum**
Times	463	0	0	463
Meaning	34	126	206	366
Span	34	0	0	34
Setting	24	177	63	264
Relation	119	14	32	165
Character	69	255	98	422
Plot	64	414	113	591
Sum	807	986	512	2305

Table 1: Distribution of Question Types in NovelQA: This table provides a breakdown of questions across different complexity categories (Multi-hop, Single-hop, Detailed) and aspect categories (Times, Meaning, Span, Setting, Relation, Character, Plot).

Computational Literary Studies in Novels as an inseparable part of the literature field, are a notable theme in NLP. Past research on novels has covered topics such as word segmentation and POS-tagging (Zhang et al., 2014; Liu and Zhang, 2012; Qiu and Zhang, 2015) and event extraction (Makazhanov et al., 2014; Sims and Bamman, 2020). Previous models' limited input window size hindered a comprehensive examination of long novels. With the advancement of models with long-context capabilities, research has been launched on benchmarks of novel summarization (Kryściński et al., 2021; Chang et al., 2023b) and evaluation on LLM's inherent novel knowledge (Chang et al., 2023a). NovelQA pioneers in question-answering on lengthy novels, marking a contribution to novel studies in the NLP field.

3 Data

3.1 Dataset Description

Data Formulation Every novel (N) in the dataset corresponds to multiple pieces of annotated data (d_i). Each piece of data consists of the following domains, *question* (Q_i), *answer* (A_i), *multi-choices* ($a_{i,0}$, $a_{i,1}$, $a_{i,2}$, and $a_{i,3}$), *gold label* ($a_{i,gold}$), *evidences* ($s_{i,0}$, $s_{i,1}$, ...), and *type* ($Complx_i$ and $Aspect_i$), among which the *answer* domain is a short answer to the question, while the *multi-choices* and *gold label* domain forms multi-choices, indicating that our dataset can serve to the evaluation of both the generative task and the multichoice task. In the generative setting, a novel text N and a question Q_i are combined to send into the model each time, and the generated answer is compared with the answer A_i. In the multichoice setting, the novel N, a question Q_i, and the four choices $a_{i,0}$ to $a_{i,3}$ are sent into the model, and the output is evaluated according to the gold label $a_{i,gold}$. Meanwhile, the *evidences* domain consists of either the original excerpts from the novel, or the reasoning steps written by the annotator.

Book Selection We aim to enhance diversity by selecting novels from various eras, genres, and formats. Recognizing that newer, popular novels are still under copyright protection, NovelQA inevitably includes some restricted-access works. Our efforts to respect copyright laws are detailed in the Ethical Statements section. We source open-copyright novels from Project Gutenberg[2], and purchase e-books for titles still under copyright protection online. All selected books exceed 50k words (approximately 67k tokens) and are in English.A visualization of the distribution of book token lengths is provided in Appendix A.2.

Question Distribution We classify our data by the complexity of solving the question and the aspect that the question focuses on. By complexity, the data are classified into three complexity-levels, with a subjective ordering of the complexity of multi-hop > detailed > single-hop. By the aspect that each question focuses on, the data entails seven types. A detailed specification of each type is listed in Appendix A.3. According to the classification above, the distribution of the questions in our dataset is displayed in Table 1.

3.2 Data Collection and Annotation

Procedure Overview The annotation process is performed by a group of skilled annotators, all of whom are holding or pursuing a degree in English Literature and have read the target novels carefully, and they are aware of the usage of the data. The annotation procedure consists of two phases: (1) Template-based phase: The annotators are required to fill entities into 19 templates (see Appendix A.1) that we design to be related to multi-hop or detailed information. This phase entails half of the data, mainly contributing to the multi-hop ones. (2) Free-formed phrase: To ensure the diversity of question expression and align the questions to the natural distribution, our second half of the data is annotated without a template, namely, the annotators contribute to any difficult questions that they come up with freely.

Template Design The first annotation phase relies on a question template, which requires the annotator to fill in the entities from the novel to form valid questions. To design templates, we carried out sufficient pre-tests on GPT-4-128K and Claude2-100K to analyze their possible weaknesses in long-input QA and novel knowledge memorization. Our pretest shows that they usually fail to tackle information spanning over multiple chapters, as well as lack attention to details that have no contribution to the main theme. Meanwhile, we also refer to around fifteen books on novel and narration theories (e.g. Forster, 1927; Tobias, 2012; Schmidt, 2012; McKee, 2005) to ensure our template covers more aspects that a novel can discuss (e.g., character, setting, theme). Aggregating all the information above, the template consisting of the 19 kinds of questions is formed.

Annotation Guideline While creating the QA tuples, our annotators are instructed to follow several principles below: (1) The annotators are either senior-grades English language & literature college students, or students with high English test scores. This ensures the annotators' proofreading ability. (2) The annotators are required to read through the GPT-4 responses on example questions, ensuring them to be familiar with the task goal. (3) The annotators should choose novels above 50K tokens and read through the books they choose before annotation. (4) Evidence of each answer should be provided for validation purposes. (5) The annotation reward, of \$1.11 per tuple, is explicitly informed in the guidelines. Data of high quality are rewarded by \$1.39 per tuple. As an average annotator can write 5 to 6 pieces of data at full speed according to our observation, the \$5.56 to \$8.34 hourly wage is above the local legal minimum wage of \$2.78/hour.

Quality Control The created data is manually double-checked by authors. Data with factual errors or questions that are too simple are either modified or eliminated. Only 79.4% of the data collected

[2] https://www.gutenberg.org/

are preserved to form the final dataset of 2305 QA tuples. Meanwhile, the inter-annotator agreement test, carrying on the multichoice versioned QA annotation on randomly selected 100 cases, shows a score of 94.66% in Cohen's Kappa, indicating a high agreement among annotators.

Distractions for Multi-choice Setting We use GPT-4 to generate three distracting options for each question and its golden answer and randomly permute the four answers.

3.3 Advantages

Our NovelQA dataset serves as a new benchmark for evaluating long-context understanding, distinguished by several key advantages. Firstly, it surpasses existing benchmarks in length, offering a rigorous test of a model's ability to navigate and comprehend significantly longer texts. Secondly, the inclusion of clear evidences alongside questions ensures that evaluations are grounded in concrete textual support, enhancing the reliability of assessments. Furthermore, the dataset emphasizes questions that require attention to detailed information, challenging models to move beyond superficial impressions to extract specific, nuanced answers. Questions, golden answers, and evidences of the dataset are entirely manually annotated and carefully checked, ensuring high-quality, nuanced questions and answers that reflect complex human thought processes. To prevent against data leakage, we will not release golden answers for the test set, minimizing the risk of overfitting. These features, combined with the dataset's comprehensive coverage of diverse narratives and meticulous construction, make NovelQA a valuable resource for advancing long-context understanding.

4 Experiments

We focus on long-context models meeting three criteria: a context window of at least 128,000 tokens, accessibility via a full API or public release, and chat functionality. For commercial models, our selection includes GPT-4-128k (OpenAI, 2023b) and Claude 2.1-200k (Anthropic, 2023). Among open-source options, we evaluated models like InternLM2-chat (Team, 2023).

4.1 Implementations

Settings To thoroughly test the abilities of these LLMs, we employ two evaluation settings: a generative setting where models directly generate short answers, and a multichoice setting with four provided options.

	Max Length	Multi-choice	Generative
GPT-4	128K	71.80	46.88
Claude 2.1	200K	66.84	46.04
InternLM2 -Chat-7b	200K	43.51	30.90
InternLM2 -Chat-20b	200K	49.18	32.37

Table 2: Evaluation of Long-Context LLMs on NovelQA. This table presents the performance of four long-context LLMs, including both commercial models (GPT-4 and Claude 2.1) and open-source, locally deployed models (InternLM2-Chat-7b/20b). Accuracy percentages are reported under two testing scenarios: multi-choice and generative. The Max Length column denotes the maximum token length of each model.

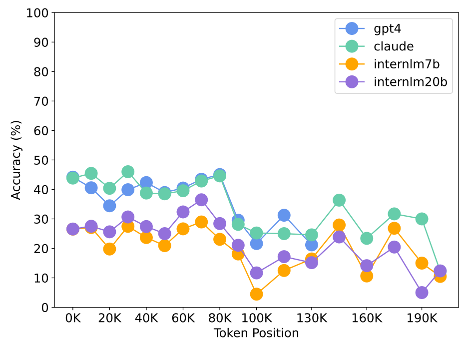


Figure 3: Analysis of Accuracy in Generative Setting by Token Indexes: This figure illustrates the accuracy, plotted against the token indexes of relevant evidences in the novels. The x-axis, reflecting token indexes, is folded on the right due to the long-tail distribution.

Prompts We use uniform prompts for all LLMs, with a start part, novel content, questions, choices in the multichoice setting, and end part. The prompt structure is shown in Appendix Table 9.

Truncation Due to input length limitations, we truncated the novel content from the end to the front, to meet with the max input length, while keeping questions and other prompts complete.

Evaluating Generative Results: Following the findings in Wang et al. (2023) which highlight GPT-4's proficiency in assessing the accuracy of short machine-generated answers, we employ GPT-4 (gpt-4-0125-preview) for the evaluation of generative responses in our study, which is also applied in other long-range benchmark studies (An et al., 2023; Li et al., 2023a). We further conducted a human evaluation on 800 pieces of generative outputs and carried an inter-evaluator agreement test

	chara	mean	plot	relat	settg	span	times	avg
mh	57.81	61.76	52.46	45.30	56.52	18.18	21.23	32.83
sh	66.12	56.56	69.33	21.43	57.23	-	-	63.93
dtl	52.63	12.87	61.68	37.50	58.06	-	-	37.58
avg	62.04	32.40	65.93	41.72	57.38	18.18	21.23	46.88

(a) GPT-4

	chara	mean	plot	relat	settg	span	times	avg
mh	48.94	64.29	58.18	40.00	82.61	18.52	17.10	30.34
sh	72.41	55.24	67.96	23.08	59.60	-	-	65.47
dtl	55.22	12.43	66.30	27.59	55.77	-	-	37.65
avg	65.90	31.61	66.60	35.61	61.06	18.52	17.10	46.04

(b) Claude 2.1

	chara	mean	plot	relat	settg	span	times	avg
mh	23.81	38.24	42.62	24.35	39.13	15.15	21.32	24.62
sh	35.80	28.10	42.18	14.29	32.10	-	-	36.42
dtl	26.67	9.18	55.14	29.03	34.43	-	-	27.02
avg	32.02	18.52	44.77	24.38	33.33	15.15	21.32	30.90

(c) InternLM2-Chat-7b

	chara	mean	plot	relat	settg	span	times	avg
mh	32.84	44.12	29.03	37.61	25.00	15.15	26.81	29.29
sh	42.97	34.17	43.22	21.42	36.90	-	-	40.57
dtl	30.93	7.00	48.21	25.00	29.03	-	-	24.65
avg	38.50	19.77	42.66	33.74	33.86	15.15	26.81	33.07

(d) InternLM2-Chat-20b

Table 3: Model Performance by Question Type in Generative Setting: This table details the accuracy scores of four models across different question types. Question types include details (dtl), multi-hop (mh), single-hop (sh), and character (chara), meaning (mean), plot, relation (relat), setting (settg), others, and an average score (avg) for each category, with '-' indicating the absence of data for a category.

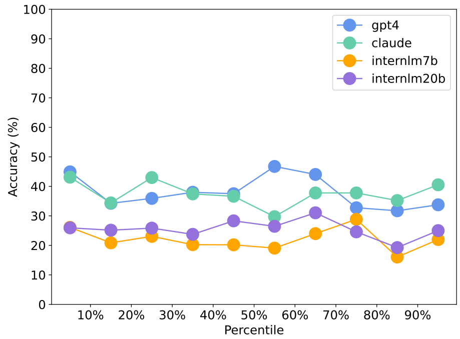


Figure 4: Analysis of Accuracy in Generative Setting by Relative Token Positions: This figure illustrates the accuracy, plotted against the percentage position of each question's relevant evidence within the novel.

between two human evaluators and the GPT-4 evaluator. The result of 89.25% in Cohen's Kappa indicates a high agreement towards the GPT-4 evaluating results (details in Appendix B.1).

Commercial LLMs: The APIs of commercial LLMs utilized are gpt-4-0125-preview[3] and anthropic.claude-v2:1[4].

Open-source LLMs: Running long-context LLMs on extremely long inputs, such as 200k tokens, is a challenge due to the immense GPU memory required, for example, it takes roughly 2.5T memory to calculate one attention matrix for a 7B model with a 200K-token input, while our local device is a 4 × 80G A100. To address this, we utilize the LMDeploy[5] (based on Dynamic NTK (emozilla, 2023)) for memory reduction, which is only compatible with several LLMs. Therefore, we choose InternLM2-Chat-7b-200K and InternLM2-Chat-20b-200K for our experiments. Attempts to run models like Yarn-Mistral-7b-128K (Peng et al., 2023), chatglm3-6b-128k (Zeng et al., 2022), XVERSE-13B-256K (xverse, 2023) and Yi-6B-200K (01-ai, 2023) on inputs exceeding 100k tokens are unsuccessful. The challenges range from GPU memory overflow due to incompatibility with LMDeploy, to models imitating the input content, or continuing to generate following the inputs without answering the questions. However, reducing input lengths to under 50k tokens can normalize most outputs, following the given instructions. This situation underscores the current limitations in technical support for processing and generating content based on extremely long contexts.

Other Details Given the cost of running long-context APIs, we request that each model respond to all questions for a given book in a single session. To ensure fair comparisons, local-deployed LLMs alsoto answer all questions for a book at once. We set 'temperature = 0' to eliminate randomness and keep other hyper-parameters default.

4.2 Main Results

We present the main results in Table 2. Even the highest scores (71.80% and 46.88% for GPT-4 in generative and multichoice settings, respectively) suggest there is considerable room for improvement in long-context understanding. This is especially true in the generative setting where understanding and recall over long contexts are more challenging. Additionally, commercial models (GPT-4 and Claude 2.1) outperform open-source models (InternLM2-Chat-7b and 20b) in this benchmark. All models show a drop in performance in the generative setting compared to the multichoice setting. This indicates the increased challenge in generating a correct answer from scratch, as opposed to selecting from provided options. We have

[3] `https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo`
[4] `https://aws.amazon.com/cn/bedrock/claude/`
[5] `https://github.com/InternLM/lmdeploy`

also observed three typical errors, hallucination, overlooking, and miscounting, a detailed analysis is conducted in Appendix B.3.

4.3 Results by the Question Type

An in-depth analysis of model performance across question types reveals nuanced insights into their comprehension abilities in both generative and multichoice settings, detailed in Tables 3 and Appendix 10, respectively. This analysis not only highlights the models' weaknesses across different formats but also illuminates the challenges in narrative comprehension, contributing to both NLP and computational literary research.

The examination of accuracy scores across question categories such as character, meaning, plot, relation, and setting highlights distinct patterns in performance, pointing to the models' differential capabilities and limitations. Notably, models exhibit particular difficulty with questions centered around meaning, relation, span, and times. This difficulty suggests several underlying challenges:

- Meaning Questions: The struggle with meaning questions indicates a challenge in grasping abstract concepts and locating entities or sentences through interpretations within the text, which requires an advanced level of semantic understanding and inference.

- Relation Questions: Difficulty with relation questions points to a gap in the models' ability to identify and interpret the dynamic and often nuanced relationships between characters, events, or concepts, which are crucial for a holistic understanding of narratives.

- Span and Times Questions: The lower performance on span and times questions suggests a limitation in tracking temporal sequences and spatial extents within the narrative, reflecting challenges in maintaining and applying contextual information over long stretches of text.

The above findings underscore a critical aspect in both computational literary studies and long-context comprehension of LLMs—while models are adept at handling certain types of narrative questions, they encounter significant hurdles when required to synthesize abstract concepts, interpret complex relationships, or maintain a coherent understanding of temporal and spatial narratives over long context. These can be the domains requiring further improvement to enhance narrative comprehension and reasoning capabilities.

4.4 Results by the Position

Our analysis delves into how the positioning of evidence within novels affects the accuracy of long-context LLMs. Specifically, we explore the impact of both absolute and relative positions of evidence, where the absolute position refers to the specific token index within the text, and the relative position is normalized against the total length of the novel, scaled to a 0%-100% range.

Absolute Position Analysis: In the generative setting, as depicted in Figure 3, all evaluated models show improved performance on questions where the necessary evidence is located before the 100k token mark. This trend highlights a challenge for LLMs in accessing and processing information beyond this threshold, suggesting a diminished capacity to handle very long inputs. The multichoice setting, detailed in Appendix B.2, follows a similar pattern, reinforcing the importance of evidence position in model performance.

Relative Position Analysis: By normalizing the evidence positions within the entire novel, we aim to understand if the proportional location of evidence influences model accuracy. This analysis, shown in Figure 4 for the generative setting and in the Appendix B.2 for the multichoice setting, indicates that models maintain relatively consistent performance across various relative positions, suggesting that long-context LLMs' effectiveness is not significantly affected by the evidence's relative position within the standardized length of novels.

To delve deeper into how long-context LLMs navigate extremely long inputs, we segment novels into two categories based on length: 65k-100k and over 100k tokens. We then examine model accuracy in relation to relative evidence positions within these ranges, with results presented in Appendix Figure 8. For novels within the 65k-100k token range, we observe a *lost-in-middle* phenomenon in GPT-4, InternLM2-7b, and InternLM2-20b, akin to findings by Liu et al. (2023). This pattern indicates stronger performance at the beginning and end of texts but weaker in the middle. Conversely, in novels exceeding 100k tokens, model performance generally declines towards the end, potentially due to the scarcity of training data for contexts of this length. This behavior underscores a unique challenge faced by LLMs when processing exceptionally long texts over 100k tokens.

	Close-book QA	
	Multichoice	Generative
GPT-4	60.94 (-9.06)	34.30 (-13.58)
Claude 2.1	51.77 (-15.01)	22.36 (-20.68)
InternLM2-Chat-7b	33.58 (-10.71)	14.12 (-16.81)
InternLM2-Chat-20b	33.05 (-16.13)	15.51 (-16.86)

Table 4: Close-book Performance across four LLMs on NovelQA. Unlike the standard scenario, models rely solely on internal knowledge without access to the novels. The parentheses indicate the performance drop from the standard to the Close-book scenario, highlighting the models' dependency on external text for answering.

	Correctness	Relevance	Sufficiency	Avg.
GPT-4	29.47	38.05	27.67	31.73
Cluade 2.1	23.51	29.08	22.27	24.95
InternLM2-Chat-7b	2.05	3.88	1.90	2.61
IntnerLM2-Chat-20b	6.36	12.51	7.50	8.79

Table 5: Evidence recalling results of four LLMs.

4.5 Analysis

Close-book Question Answering We employ a Close-book QA scenario to assess the extent to which models rely on the content of novels versus using their internal knowledge to answer questions. In this approach, the models are not given access to the text of the novels and must rely solely on their pre-existing knowledge to provide answers. Given that our selected novels are well-known and representative of their genres, it's inevitable that LLMs have encountered their texts during training and retained some of their content. The results of this evaluation are presented in Table 4.

Models like GPT-4 and Claude 2.1 achieve notable scores in the Close-book setting (60.94% and 51.77% in multichoice, 34.30% and 22.36% in generative, respectively) indicating that they have internalized significant portions of the novels' content during training. This internal knowledge allows them to perform reasonably well even without direct access to the text. The difference in performance between the Close-book and standard settings underscores the challenges in long-context understanding. While models can retain and recall information from well-known texts, their ability to comprehend and use such information to answer questions accurately diminishes in the absence of the text. This suggests that long-context understanding, as measured by the main results, might still be more challenging than it appears, as models benefit from having the text directly available.

Evidences Recall We evaluate the ability to recall evidence, prompting the four models above to answer the questions in NovelQA again, with printing the supporting evidence simultaneously. We then prompt GPT-4 with the generated evidence alongside the annotated evidence to obtain its evaluation on the quality of retrieved evidence pieces. The evaluating matrix consists of the following three dimensions: *correctness* refers to whether the retrieved evidence is the same as the annotated evidence or with a similar correct meaning; *relevance* indicates whether the evidence is consistent with the answer; *sufficiency*, whether the retrieved pieces of evidence are enough to support the answer. Each dimension is scored between 0 and 100 and an average score is further obtained through calculating the algorithmic mean on these three dimensions. Prompts involved in this evaluation procedure are presented in Appendix A.4.

The results, detailed in Table 5, show higher performances of GPT-4 and Claude 2.1. Moreover, though the scoring range is from 0 to 100, the four models all perform with low scores in evidence recall, possibly because the models do not always follow the instructions after inputs and outputs of such an above-average length, resulting in a high percentage of 0 scores. This phenomenon is particularly severe for InternLM-Chat-7b and InternLM-Chat-20b models, whose outputs consist of a large proportion of invalid placeholders. Still, current long-context LLMs generally demonstrate inadequate abilities recalling the correct and supportive evidences from the context.

5 Conclusion

We introduced NovelQA, a long-range benchmark designed to assess the long-context comprehension abilities of LLMs. Utilizing representative English novels, NovelQA presents LLMs with the challenge of navigating complex, real-world texts. Our evaluations reveal that both commercial and open-source models face difficulties with detailed understanding, multi-hop reasoning, and accurately retrieving specific information from lengthy contexts, especially for lengths are 100,000. Moreover, operating LLMs on inputs exceeding 200,000 tokens faces technical challenges, notably in terms of memory requirements and associated costs. NovelQA's contributions extend to advancing research in both NLP and computational literary studies.

Limitations

Access to Close-source LLMs: One significant limitation is our inability to obtain APIs for certain close-source long-context LLMs, such as Baichuan-192K, GLM4-200K, and Moonshot-192K.
Technical Constraints for Open-source LLMs: The majority of open-source LLMs are not optimized for processing inputs exceeding 128K tokens with high GPU memory cost. This technical challenge renders direct inference on NovelQA infeasible for these models. Consequently, our experiments were limited to LLMs that have been adapted using LMDeploy.
Language Limitation:NovelQA, and all associated data are exclusively in English.

Ethics Statement

We are committed to ensuring that NovelQA serves purely for academic and scientific purposes. To facilitate widespread research use while adhering to copyright laws, we plan to launch a leaderboard website and provide an API. This approach allows users to engage with NovelQA for evaluation without the need for public data release, particularly concerning copyrighted novels. By taking these measures, we aim to respect the "fair use" principle under copyright protection[6], ensuring our project navigates legal and ethical boundaries responsibly.

References

01-ai. 2023. Yi-6b-200k.

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL. Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. 2024. LongHealth: A question answering benchmark with long clinical documents.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-Eval: Instituting standardized evaluation for long context language models.

Anthropic. 2023. Cluade-2.1.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024. LongAlign: A recipe for long context alignment of large language models.

[6]https://support.google.com/legal/answer/3463239?hl=en&ref_topic=4558877&sjid=14110422187432235906-EU

Yushi Bai, Xin Lv, Jiajie Zhang, Hong Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. LongBench: A bilingual, multitask benchmark for long context understanding. *ArXiv*, abs/2308.14508.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023a. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023b. BooookScore: A systematic exploration of book-length summarization in the era of llms.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023a. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. *arXiv preprint arXiv:2305.14788*.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Le Zhou, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. 2023. Learning a foundation language model for geoscience knowledge understanding and utilization. *arXiv preprint arXiv:2306.05064*.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models.

Zican Dong, Tianyi Tang, Lunyi Li, and Wayne Xin Zhao. 2023b. A survey on long text modeling with transformers. *arXiv preprint arXiv:2302.14502*.

emozilla. 2023. Dynamically scaled rope further increases performance of long context llama with zero fine-tuning.

Guhao Feng, Yuntian Gu, Bohang Zhang, Hao-Tong Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: a theoretical perspective. *ArXiv*, abs/2305.15408.

Edward Morgan Forster. 1927. *Aspects of the Novel*. Harcourt, Brace.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2023. Lm-infinite: Simple on-the-fly length generalization for large language models. *arXiv preprint arXiv:2308.16137*.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. 2024. KVQuant: Towards 10 million context length llm inference with kv cache quantization. *arXiv preprint arXiv:2401.18079*.

Yunpeng Huang, Jingwei Xu, Zixu Jiang, Junyu Lai, Zenan Li, Yuan Yao, Taolue Chen, Lijuan Yang, Zhou Xin, and Xiaoxing Ma. 2023. Advancing transformer architecture in long-context large language models: A comprehensive survey. *ArXiv*, abs/2311.12351.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. BookSum: A collection of datasets for long-form narrative summarization.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023a. LooGLE: Can long-context language models understand long contexts?

Xianming Li, Zongxi Li, Xiaotian Luo, Haoran Xie, Xing Lee, Yingbin Zhao, Fu Lee Wang, and Qing Li. 2023b. Recurrent attention networks for long-text modeling. *arXiv preprint arXiv:2306.06843*.

Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023c. Compressing context to enhance inference efficiency of large language models. *arXiv preprint arXiv:2310.06201*.

Bin Lin, Tao Peng, Chen Zhang, Minmin Sun, Lanbo Li, Hanyu Zhao, Wencong Xiao, Qi Xu, Xiafei Qiu, Shen Li, et al. 2024. Infinite-LLM: Efficient llm service for long context with distattention and distributed kvcache. *arXiv preprint arXiv:2401.02669*.

Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, et al. 2023. GeoGalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*.

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

Yang Liu and Yue Zhang. 2012. Unsupervised domain adaptation for joint segmentation and pos-tagging. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pages 745–754. Indian Institute of Technology Bombay.

Aibek Makazhanov, Denilson Barbosa, and Grzegorz Kondrak. 2014. Extracting family relationship networks from novels. *arXiv preprint arXiv:1405.0603*.

Robert McKee. 2005. *Story*. Dixit.

OpenAI. 2023a. GPT-4 technical report.

OpenAI. 2023b. GPT-4 technical report. *arXiv*.

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. 2023. YaRN: Efficient context window extension of large language models.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Likun Qiu and Yue Zhang. 2015. Word segmentation for chinese novels. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2440–2446. AAAI Press.

Victoria Lynn Schmidt. 2012. *45 Master Characters, Revised Edition: Mythic Models for Creating Original Characters*. Penguin.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding.

Matthew Sims and David Bamman. 2020. Measuring information propagation in literary social networks. *arXiv preprint arXiv:2004.13980*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*.

InternLM Team. 2023. InternLM: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Ronald B Tobias. 2012. *20 master plots: And how to build them*. Penguin.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2023. Focused transformer: Contrastive training for context scaling.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-QA evaluation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Y Wang, D Ma, and D Cai. 2024. With greater text comes greater necessity: Inference-time training helps long text generation. *arXiv preprint arXiv:2401.11504*.

Qingyang Wu, Zhenzhong Lan, Kun Qian, Jing Gu, Alborz Geramifard, and Zhou Yu. 2022a. Memformer: A memory-augmented transformer for sequence modeling. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. 2022b. Memorizing transformers. *arXiv preprint arXiv:2203.08913*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

xverse. 2023. XVERSE-13b-256k.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. GLM-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Type-supervised domain adaptation for joint segmentation and pos-tagging. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 588–597. The Association for Computer Linguistics.

Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. 2024. Soaring from 4k to 400k: Extending llm's context with activation beacon. *arXiv preprint arXiv:2401.03462*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, and Yanlin Wang. 2023. MemoryBank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*.

Model	Human Evaluator A	Human Evaluator B	Avg.
GPT-4	91.97	95.84	93.91
Claude 2.1	88.04	90.00	89.02
InternLM-Chat-7b	91.88	86.53	84.88
InternLM-Chat-20b	85.68	84.08	87.25
Avg.	89.39	89.11	89.25

Table 6: The Cohen's Kappa score in Inter-evaluator Agreement test on different model outputs between human-evaluator and GPT4-as-evaluator. Higher Cohen's Kappa scores indicate higher agreement between human-evaluators and the GPT-4-evaluator.

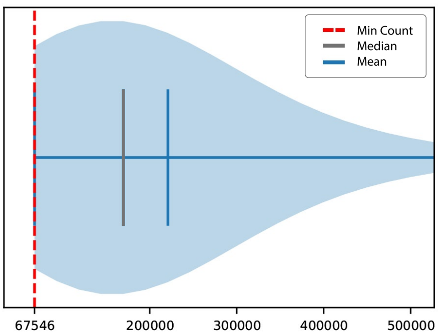


Figure 5: Token Count Distribution in NovelQA. The token count of both the novel and the questions are counted.

Appendix

A Data

A.1 Question Templates

The 17 pieces of question template adopted in the annotation procedure are presented in Table 7.

A.2 Distribution of Input Token Count

The input token count of each novel in NovelQA is calculated by adding the book length to the lengths of its related questions. The distribution of the token count is illustrated in Figure 5.

A.3 Data Classification

The data in NovelQA are classified into 3 complexity-levels and focus on 7 aspects. A detailed description of criteria and examples for each class is presented in Table 8.

A.4 Prompts

We present all the prompts involved in either tests or evaluations in Table 9.

B Experiments

B.1 Inter-evaluator Agreement for GPT-4-as-Evaluator

We also perform an inter-evaluator agreement test between human-evaluators and the GPT-4-evaluator. The results of Cohen's Kappa score are presented in Table 6.

B.2 Relation between Multi-choice Accuracy and Evidence Positions

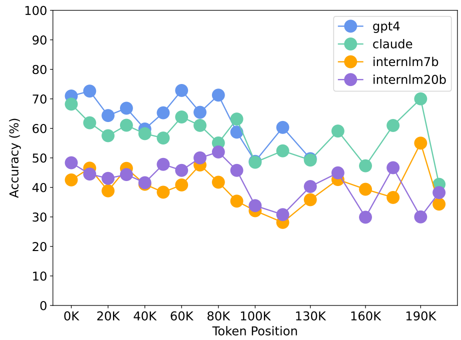


Figure 6: Analysis of Accuracy in Multichoice Setting by Token Indexes: This figure illustrates the accuracy in the generative setting of NovelQA, plotted against the token indexes of relevant evidences in the novels. Each subplot represents a different model. The x-axis, reflecting token indexes, is folded on the right due to the long-tail distribution in the lengths of the selected novels.

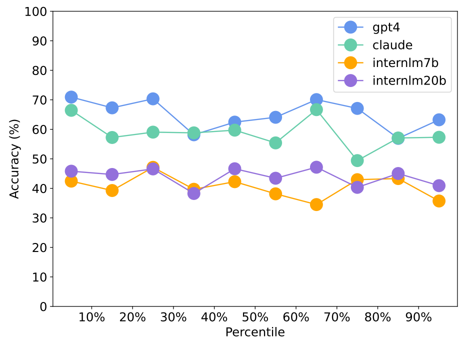


Figure 7: Analysis of Accuracy in Multichoice Setting by Relative Token Positions: This figure illustrates the accuracy in the multichoice setting of NovelQA, plotted against the percentage position of each question's relevant evidence within the novel. Each subplot represents a different model.

Besides the evaluation in generative settings, we also prompted the models to collect their responses for the multichoice versioned questions. Table 10 presents the accuracies of four models in multichoice settings in each question type. Figure 6 and Figure 7, presenting the relationships between the accuracy and the absolute or relative positions accordingly, show similar trends to which are observed in the generative setting.

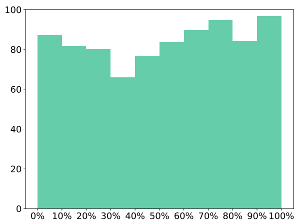
B.3 Representative Errors

Through reviewing the generation above, we concluded four common error types: hallucination, miscounting, overlooking, and reasoning failure. Examples for each type are presented in Table 11.

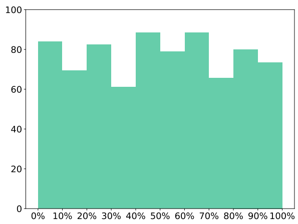
Hallucination refers to the information generated by the model with factual errors. In our generative QA setting, typical hallucination mistakes encompass two types: (1) factual errors about the fictional settings (e.g., mixing entities within the setting or between the settings of different books) and (2) factual errors about the narrations (e.g., whether a fact is narrated). The first category usually appears in questions asking minor characters, plots, and settings, where the model might output a non-existing one. Meanwhile, the second category is often associated with sentence-locating questions, which ask the model to locate a sentence. In this case, the model may fake a sentence that does not exist in the original text.

Overlooking refers to the model's neglect of details. As mentioned above, the questions in *detailed* category involve minor characters, plots, or settings. Diving further, the reasons why these details are difficult to be recalled lie in two aspects: (1) They do not contribute to the character development, other plots, or the main themes, and thus reading the rest of the novel does not help to remind this detail; (2) Since most novels have derivative works (e.g., films, fan works, and book reviews), where the detailed information is eliminated to form a condensed narration. As the derivative works spread further and appear more frequently in the model's training data, they have a higher probability of becoming the models' inner knowledge, which is similar to (Chang et al., 2023a)'s observation, and vice versa for those omitted details. These two factors contribute to the difficulty in the model's recalling details and thus result in overlooking errors.

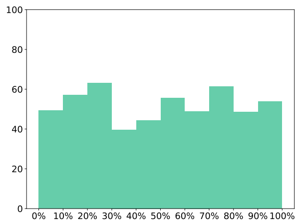
Miscounting Researches (Li et al., 2023a; Feng et al., 2023) has revealed shortcomings in the counting ability of LLMs, especially autoregressive-decoder-based models, and methods unfolding the outputs such as chain-of-thought prompting can enhance their counting ability. Our test does show that models make mistakes with numbers. Though the errors in the case of generative responses may be due to not following instructions and simply outputting "multiple times" instead of the desired specific times, the accuracy in the multichoice setting has still only reached 38.53% to 49.56% for the chosen four models, as shown in Table 4. Even in the simplest question which asks for the appearing frequency of certain phrases, the models still make mistakes.



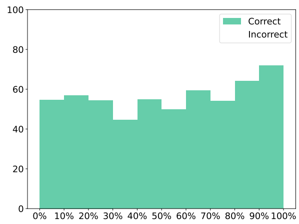
(a) GPT-4
65-100K



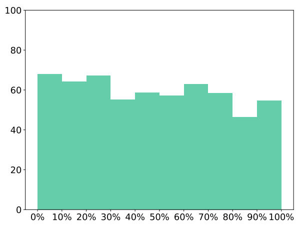
(b) Claude 2.1
65-100K



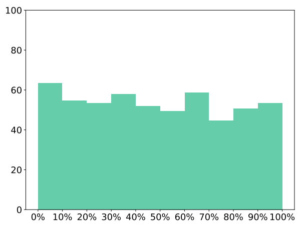
(c) InternLM-7b
65-100K



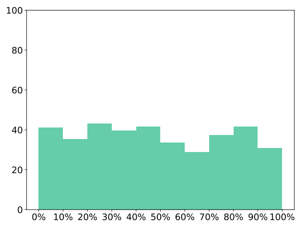
(d) InternLM-20b
65-100K



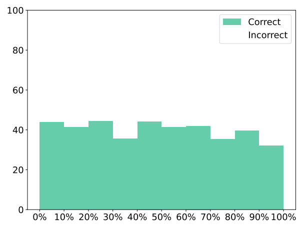
(e) GPT-4
100+K



(f) Claude 2.1
100+K



(g) InternLM-7b
100+K



(h) InternLM-20b
100+K

Figure 8: Performance Analysis by Relative Positions within Different Token Ranges: This figure illustrates the accuracy of various long-context LLMs in the multichoice setting, segmented by novels' length categories: 65k-100k tokens and over 100k tokens. It highlights the models' performance trends relative to the evidence position within these ranges, showcasing the 'lost-in-middle' phenomenon in shorter texts and a performance decline towards the end in longer texts, thus revealing distinct comprehension challenges faced by LLMs when processing texts of varying lengths.

Aspect	Template	Answer Format
times	Has the word or phrase <a word or phrase> appeared in the novel? If so, how many times does it appear in the text?	Yes/No + number
	Has the object <an object> appeared in the novel? If so, how many times does it appear in the text?	Yes/No + number
	Has the place <a place> appeared in the novel? If so, how many times does it appear in the text?	Yes/No + number
	Has the sentence <a sentence> appeared in the novel? If so, how many times does this sentence appear in the text?	Yes/No + number
	Has the character <a character> appeared in the novel? If so, how many times does he/she/they appear in the text?	Yes/No + number
	Has the plot <a plot> happened in the novel? If so, how many times does it happen in the text?	Yes/No + number
	How many times has <a character> done <a doing-verb phrase> in the novel?	Yes/No + number
	How many times have <a character> and <a character> <do something> together in the novel?	Yes/No + number
	How many times have <a character> and <a character> <met each other (or appear together)> in the novel?	Yes/No + number
	How many times have <a character> and <a character> <communicate or have verbal conflicts with each other> in the novel?	Yes/No + number
mean	Explain the meaning or implication of the symbol or metaphor <a symbol or a metaphor> in one sentence, which appears in the novel.	An explanation
	In which chapter does there exist a sentence in the novel <the novel> with the same or similar meaning as "<a sentence not from the original text>"? Please output the chapter name or index.	A chapter index or title
chara	<A character> is used to be <positive or negative> and finally becomes a <negative or positive> one in the novel. Tell in one sentence which episode marks this character's change.	A plot in one sentence
	Who are mentioned with names in the <an organization, a family, or a club> in the novel?	A list of names
	Please list 3 aliases or designations of <a character> in the novel.	3 aliases or designations
	Who is <a minor character, or a character without name, or a character that appears only once> in the novel?	A description of character
settg	In which <cities, or countries> does this story take place in the novel?	A list of cities or countries
	In which year does the earliest event happen, and in which year does the latest even happen in the novel?	A range in years
relat	What is the relationship between <a character> and <an alias or a nick-name of this character> in the novel?	A relationship

Table 7: Question template adopted in data annotation. Question types include character (chara), meaning (mean), plot, relation (relat), setting (settg), times and span. '<>' label indicates the entity for annotators to fill in.

Dimension	Type	Percentage	Description	Example
By Complexity	Multi-hop	34.98%	Questions requiring knowledge across multiple paragraphs, or even multiple chapters, to be solved.	*How many times has Bran jumped off and ran?* in *A Game of Thrones*
	Single-hop	42.74%	Questions requiring knowledge from one or several adjacent single sentences to be solved.	*According to the Colonel, what did he smoke in McQueen's compartment?* in *Murder on the Orient Express*
	Detailed	22.19%	Questions requiring knowledge from one or two adjacent sentences to be solved. Detailed questions are distinguished from the single-hop class by involving information that is too minor to impact other plots, making the details difficult to be recalled.	*How many candles Madame Magloire lighted when the Bishop had his last dinner with Jean Valjean?* in *Les Misérables*
By Aspect	Times	20.07%	About the number of times that a character, location, or plot appears in the novel.	*How many times has Kitty kissed Walter?* in *The Painted Veil*
	Meaning	15.86%	The understanding of certain sentences or metaphors, e.g., to interpret the relationship of a certain metaphor and the novel itself, or find a specific sentence according to a paraphrase provided by the annotator.	*In which chapter does there exist a sentence with the same or similar meaning as 'The Marquis responded, 'You do me too much honor. In any case, I lean toward that assumption."?* in *A Tale of Two Cities*
	Span	1.47%	About the range of the novel setting. To be specific, they either ask about the starting and ending year of the story or require listing all the cities or countries that are involved in the story.	*In which year does the earliest event happen, and in which year does the latest even happen?* in *Tess of the d'Urbervilles*
	Setting	11.44%	About the time or place settings, besides those in the span type, are classified in this type.	*Where did Diana's cousins leave for the Debating Club concert?* in *Anne of Green Gables*
	Relation	7.15%	About the relationship of multiple character entities. To be specific, they ask either about the relationship of a character and their alias or designation, or about the relationship between different characters.	*What is the relationship between Jean Valjean and Ultime Fauchelevent?* (designation) *Who are members of ABC friends?* (membership) in *Les Misérables*
	Character	18.29%	About the information of characters, besides those in the relation type, are classified into this type.	*Who is Miss Beirne?* in *Dubliners*
	Plot	25.62%	We define a plot as "some character does something for once". Questions that ask about the information of any plots are classified into this type.	*What does Clarissa repair in preparation for the night's party?* in *Mrs.Dallory*

Table 8: Data Distribution

QA Setting	Prompt
Generative	You are a literature professor. I will provide you with the full text of a novel along with a series of questions. Please thoroughly analyze the novel 's content to accurately respond to each of the following questions. Book title: <title>; Book Content: <content>; Book ends. Questions start here: N ×(Question: <question>); Questions end here. Try your best to answer the questions based on the given full text the novel. The answer should be in short with only one or several words. Your output format should be 'Answer0: <answer>Answer1: <answer>... Answern: <answer>', each answer in one line without outputing the questions and other info.
MultiChoice	You are a literature professor. I will provide you with the full text of a novel along with a series of questions and corresponding choices pertaining to it. Please thoroughly analyze the novel 's content to accurately respond to each of the following questions. Book title: <title>; Book Content: <content>; Book ends. Questions start here: N ×(Question: <question> Choices: 0: <choice0> 1: <choice1> 2: <choice2> 3: <choice3>); Questions end here. Try your best to select the correct choice to each question based on the given full text the novel. Your should output the choice to each question with the format 'Answer0: <choice> Answer1: <choice>... Answern: <choice>' (only the choice index is required), each answer in one line without outputing the questions and other info.
Closebook-Generative	You are a literature professor. I will provide you a series of questions. Please accurately respond to each of the following questions. Book title: <title>; Book Content: <content>; Book ends. Questions start here: N ×(Question: <question>); Questions end here. Try your best to answer the questions based on your own knowledge. The answer should be in short with only one or several words. Your output format should be 'Answer0: <answer>Answer1: <answer>... Answern: <answer>', each answer in one line without outputing the questions and other info.
Closebook-MultiChoice	You are a literature professor. I will provide you a series of questions along with four choices for each question. Please accurately select the correct choice to each of the following questions. Book title: <title>; Book Content: <content>; Book ends. Questions start here: N ×(Question: <question> Choices: 0: <choice0> 1: <choice1> 2: <choice2> 3: <choice3>); Questions end here. Try your best to answer the questions based on your own knowledge. Your should output the choice to each question with the format 'Answer0: <choice> Answer1: <choice>... Answern: <choice>' (only the choice index is required), each answer in one line without outputing the questions and other info.
Evaluating Generative	You are a literature professor reviewing a student's quiz paper. The question is about the novel <novel title>: <question>. The related evidences from the novel are: <evidences>. Correct ans is: <ca>. Student ans is: <sa>. Plz check whether the student's ans is correct wrt. the correct ans, and return "C" for correct and "N" for not correct. esp., if the student grabs the correct ans's meaning, return "C". However, if there are factuality errors in student ans, or the question requires a specific number but the student answers a rough number, you should return "N". Please only return the char C or N w/o any other output.
Evidence Recall	You are a literature professor. I will provide you with the full text of a novel along with a series of questions. Please thoroughly analyze the novel's content to accurately respond to each of the following questions. Book title: <title>; Book Content: <content>; Book ends. Questions start here: N ×(Question: <question>); Questions end here. Try your best to answer the questions based on the given full text of the novel. The answer should be in short with only one or several words. Your output format should be Ánswer0: <answer>$ <evidences> Answer1: <answer>$ <evidences>... Answern: <answer>$ <evidences>, each answer in one line with all the supporting evidences. Each evidence should be a sentence exactly from the original text without any paraphrase.
Evaluating Evidence Recall	You are a literature professor reviewing student's evidence for their answer about novel <novel title>. Question: <ques>. Correct answer: <ca>. Student answer: <sa>. Correct evidence: <ce>. Student evidence: <se>. You should evaluate the student evidence in 3 aspects: C) correctness: whether the student evidence is the same with the correct evidence or with a similar correct meaning. R) relevance: whether the evidence is relevant to the ans. S) sufficiency: whether sufficient evidences are retrieved to support the ans. And give a score of 1-100 to only the evidence (not the ans). You should **only** return 3 score numbers, e.g.in format C50R66S33, without any other outputs.

Table 9: Prompts used in each setting of question answering.

	chara	mean	plot	relat	settg	span	times	avg
dtl	69.39	30.10	85.84	53.12	80.95	-	-	57.62
mh	76.81	88.24	87.50	79.83	91.67	52.94	45.79	60.22
sh	86.27	88.10	92.03	57.14	87.01	-	-	88.64
avg	80.81	55.46	90.36	72.73	85.98	52.94	45.79	71.80

(b) Claude 2.1

	chara	mean	plot	relat	settg	span	times	avg
mh	71.88	76.47	85.00	70.69	78.26	51.52	47.71	58.17
sh	82.28	76.72	83.29	50.00	80.89	-	-	81.19
dtl	63.51	38.12	76.92	68.75	79.03	-	-	58.02
avg	76.80	54.55	82.21	68.52	80.17	51.52	47.71	66.78

(c) InternLM2-Chat-7b

	chara	mean	plot	relat	settg	span	times	avg
mh	42.19	38.24	45.90	46.15	65.22	39.39	42.92	43.87
sh	44.44	39.34	44.56	28.57	48.15	-	-	44.23
dtl	52.63	26.24	55.14	31.25	59.68	-	-	41.54
avg	45.69	31.84	46.79	41.72	52.63	39.39	42.92	43.51

(a) GPT-4

	chara	mean	plot	relat	settg	span	times	avg
mh	55.88	44.12	53.12	56.78	66.67	44.12	38.53	45.15
sh	53.94	53.97	58.94	35.71	62.15	-	-	57.26
dtl	47.96	25.98	55.75	18.75	55.56	-	-	30.00
avg	52.86	37.36	57.70	47.56	60.98	44.12	38.53	49.18

(d) InternLM2-Chat-20b

Table 10: Model Performance by Question Type in Multichoice Setting: This table details the accuracy scores of four models across different question types within the Multichoice setting of NovelQA. Question types include character (chara), meaning (mean), plot, relation (relat), setting (settg), and others, with '-' indicating the absence of data for a category. The table also provides an average score (avg) for each question category and model. Abbreviations used are dtl (details), mh (multi-hop), sh (single-hop).

Fault Type	Subtype	Setting	Example QA
Hallucination	On Fictional Setting	GPT-4, generative	**Book**: *Mansfield Park* **Q**: What is the relationship between Miss Maria Ward and Lady Bertram? **Correct A**: Miss Maria Ward and Lady Bertram are the same person. **Model A**: Sisters.
	On Narration	GPT-4, generative	**Book**: *The Night Land* **Q**: How many times has Aesworpth shouted? **Correct A**: 1 **Model A**: Not mentioned.
Overlooking		GPT-4, generative	**Book**: *Light in August* **Q**: Who is Percy Grimm? **Correct A**: Percy Grimm - the captain of the State National Guard who kills Joe Christmas and castrates him. **Model A**: Percy Grimm does not appear in the novel.
Miscounting		GPT-4, generative	**Book**: *Can You Forgive Her?* **Q**: Has the word or phrase 'take away another man' appeared? If so, how many times does it appear in the text? **Correct A**: Yes, 2. **Model A**: Yes, 1.

Table 11: Categories of representative errors observed in evaluating LLMs on NovelQA.