LFED: A Literary Fiction Evaluation Dataset for Large Language Models

Linhao Yu[1], **Qun Liu**[2], **Deyi Xiong**[1*]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Huawei Noah's Ark Lab
{linhaoyu, dyxiong}@tju.edu.cn, qun.liu@huawei.com

Abstract

The rapid evolution of large language models (LLMs) has ushered in the need for comprehensive assessments of their performance across various dimensions. In this paper, we propose LFED, a **L**iterary **F**iction **E**valuation **D**ataset, which aims to evaluate the capability of LLMs on the long fiction comprehension and reasoning. We collect 95 literary fictions that are either originally written in Chinese or translated into Chinese, covering a wide range of topics across several centuries. We define a question taxonomy with 8 question categories to guide the creation of 1,304 questions. Additionally, we conduct an in-depth analysis to ascertain how specific attributes of literary fictions (e.g., novel types, character numbers, the year of publication) impact LLM performance in evaluations. Through a series of experiments with various state-of-the-art LLMs, we demonstrate that these models face considerable challenges in effectively addressing questions related to literary fictions, with ChatGPT reaching only 57.08% under the zero-shot setting. The dataset will be publicly available at `https://github.com/tjunlp-lab/LFED.git`.

Keywords: Evaluation, Large Language Models, Literary Fiction Question Answering

1. Introduction

Numerous datasets have been developed to facilitate machine reading comprehension tasks, e.g., MCTest (Richardson et al., 2013), MCScript (Ostermann et al., 2018), RACE (Lai et al., 2017), CoQA (Reddy et al., 2019), WYWEB (Zhou et al., 2023), to name a few. However, as large language models (LLMs) have made remarkable progress recently, these passage-based datasets are no longer capable of evaluating such large models. More challenging datasets with long documents that go beyond the context windows of LLMs (even for the 100K-token context window of Anthropic Claude[1]), complicated reasoning (e.g., character relationship reasoning, counterfactual reasoning), skills of connoisseurship, etc., are much desirable for evaluating highly capable LLMs.

To bridge this gap, we curate LFED, a Literary Fiction Evaluation Dataset for large language models. LFED is a comprehensive dataset derived from a diverse collection of literary fictions that are either originally written in Chinese or translated into Chinese. It encompasses 8 distinct question types, which focus on the core aspects of the fictions, such as content, character relationships, storyline, writing techniques, and thematic values. In order to automate and standardize the evaluation of LLMs on LFED, we construct multiple-choice questions under each question type, providing carefully-prepared multiple answer choices for each question. The construction of the dataset is via crowdsourcing, with rigorous quality control. Ultimately, we have curated a total of 1,304 questions derived from 95 fictions. This dataset can serve as a comprehensive and challenging evaluation benchmark for assessing the fact understanding, logical reasoning, contextual comprehension, common-sense reasoning, and value judgment capabilities of large language models.

Our main contributions are summarized as follows.

1. We propose LFED, which, to the best of our knowledge, is the first Chinese dataset curated for evaluating LLMs on long literary fictions.
2. We define a question taxonomy according to the nature of literary fictions, which exhibits a wide coverage on the skills necessary for reading and understanding these fictions.
3. We have evaluated a number of LLMs on the curated dataset under the zero- and few-shot setting. Evaluation results demonstrate that long literary fiction comprehension is very challenging for LLMs, with ChatGPT achieving an accuracy of 57.08% under the zero-shot setting.

2. Related Work

We review existing machine reading comprehension (MRC) and question answering (QA) datasets within the scope and page constraint of this paper, highlighting representative Chinese datasets

*Corresponding author

[1] https://www.anthropic.com/index/100k-context-windows

LREC-COLING 2024, pages 10466–10475
20-25 May, 2024.

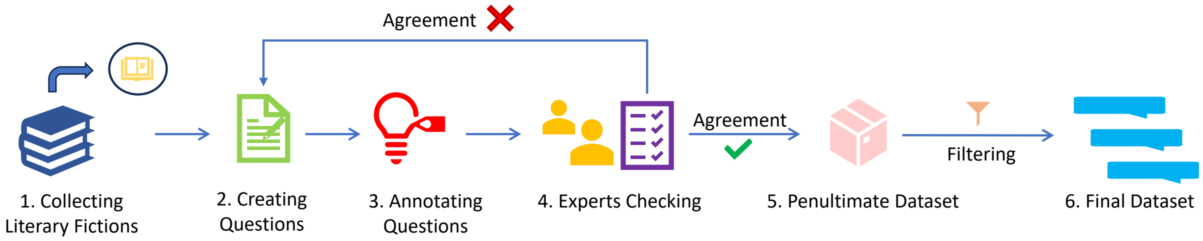

Figure 1: The overall pipeline for collecting questions in LFED.

in different categories (Guo et al., 2023). **Multiple-Choice QA Datasets** Multiple-choice questions are a specific question format that provides answer choices for each question. Numerous existing multiple-choice QA datasets are sourced from school examinations. For instance, RACE (Lai et al., 2017) encompasses a vast collection of over 28,000 essays and nearly 100,000 questions, extracting from both general and specific subjects covered in Chinese middle and high school English exams. NCR (Xu et al., 2021), on the other hand, comprises remarkably long modern and classical Chinese essays on various topics derived from high school Chinese language courses. It is tailored to evaluate the language proficiencies of native speakers. MCTest (Richardson et al., 2013) presents single-choice reading comprehension questions based on fictional stories. Additionally, recent efforts have been dedicated to curating datasets in the multiple-choice QA form for evaluating LLMs from different perspectives (Guo et al., 2023), such as CBBQ (Huang and Xiong, 2023), covering stereotypes and societal biases in 14 social dimensions related to Chinese culture and values, RoleEval (Shen et al., 2023), a bilingual benchmark designed to assess the memorization, utilization, and reasoning capabilities of role knowledge, etc. M3KE (Liu et al., 2023) collects 20,477 questions from 71 tasks, covering all major levels of the Chinese education system, from primary school to university, and a wide range of subjects including humanities, history, politics, law, education, psychology, science, technology, art and religion. LHMKE(Liu et al., 2024) encompasses 10,465 questions across 75 tasks covering 30 subjects, ranging from primary school to professional certification exams. Notably, LHMKE includes both objective and subjective questions, offering a more holistic evaluation of the knowledge level of LLMs.

Extractive MRC Datasets There has been a significant surge in the development of various extractive MRC datasets. One prominent example is SQuAD (Rajpurkar et al., 2016), which comprises questions generated by crowdsourced workers based on a collection of Wikipedia passages. Each question is designed to elicit an answer that corresponds to a specific text or span within the associated reading passage. Another dataset is BiPaR (Jing et al., 2019), which is a manually annotated bilingual parallel novel machine reading comprehension dataset. It facilitates monolingual, multilingual, and interlingual reading comprehension tasks specifically focused on novels. CMRC2018 (Cui et al., 2019), on the other hand, is an extractive dataset designed for Chinese machine reading comprehension. It contains a substantial collection of 20,000 real-world questions derived from Wikipedia sources. Furthermore, CJRC (Duan et al., 2019) is a dataset specifically created for Chinese judicial reading comprehension. The documents in this dataset are sourced from judicial documents, and the questions are annotated by legal experts, providing a valuable resource for exploring legal domain comprehension tasks.

Generative MRC Datasets The most authentic approach for human question answering involves generating answers independently, without being constrained to selecting predetermined options or extracting fragments from given documents as answers. This format enables the exploration of various question types. Notably, MS MARCO (Nguyen et al., 2016) is designed as a generative dataset that emphasizes deep learning in the search domain. In the case of NarrativeQA (Kociský et al., 2018), questions and answers are crafted by crowdsourcing workers based on book summaries. This format necessitates models to comprehend the underlying narrative in order to provide accurate answers. Additionally, DRCD (Shao et al., 2018) serves as a standard Chinese machine reading comprehension dataset. It consists of 10,014 paragraphs sourced from 2,108 Wikipedia articles, accompanied by over 30,000 questions generated by annotators.

Our LFED is unique in its utilization of long literary fictions as the data source, deviating from passage-based QA datasets. Furthermore, LFED offers a comprehensive assessment of LLMs capabilities in fact understanding, logical reasoning, context comprehension, common sense reasoning, and value judgment across eight distinct question categories.

3. Dataset Cre

Figure 1 shows the overall da pipeline. We design a very rig process to ensure the quality of th step, from the source of the data tion and review of the dataset. question taxonomy to guide ann

3.1. Data Source

A wide variety of novels are se their complex narratives, chara profound themes, and rich lingu These aspects make novels su ing Large Language Models (LL pacities, including fact understa reasoning. Unlike academic a structured and precise, or news reports, which are concise and direct, novels provide a deeper, more nuanced content that challenges LLMs to understand underlying themes and cultural nuances.

We crawl a list of the top 200 literary novels' name according to the recommendations on Douban[2], a Chinse community site with reviews of books and movies. We only select literary novels according to the reviews of readers published in Douban. We do NOT use any electronic versions of these fictions. All our hired crowdsourced workers read these fictions either with copyright or with bought hardcopies. Crowdsourced workers creating these questions give informed consent for the use of their contributions in LLM evaluation. Subsequently, each literary fiction in the list is manually checked to see if it satisfies with specific requirements. Only those that pass this manual selection are kept. The specific requirements are as follows:

- Choosing classic novels: Classic novels usually have literary, historical and cultural values, which are widely recognized and read.
- Considering the genre of fictions: Such a consideration aims to diversify the selected literary fictions in terms of genres.
- Scrutinizing Fiction content:The selected novel should be in line with human values. we eliminate novels that do not conform to contemporary values in the screening process, even though no such novels are ultimately selected. But in doing so, our dataset can prevent possible bias and ethical problems that current LLMs attempt to avoid too.
- Taking the popularity and influence of a fiction into account: The popularity of a fiction would make it easy for us to find crowdsourced workers to create questions and answers for it.

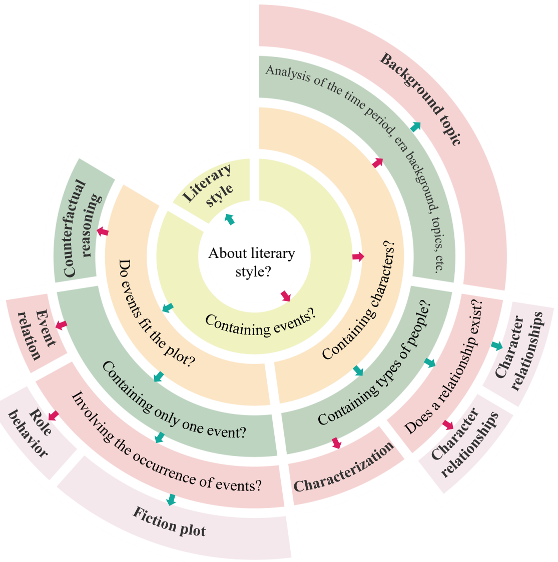

Figure 2: Decision-tree-style Illustration of the question taxonomy. Green arrows denote yes, while red arrows indicate no.

3.2. Question Taxonomy

We develop a question taxonomy to guide the creation of questions according to the nature and characteristics of literary fictions. The taxonomy covers 8 question categories, illustrated in Figure 2, which are character relationships, characterization, literary style, role behavior, event relations, fiction plot, background topic, and counterfactual reasoning. We provide the descriptions and examples of the 8 question categories in Table 1.

The design of the question taxonomy follows a systematic investigation of characters and events featured in literary fictions, which aims at the diversity and coverage of curated questions in the dataset.

Though we have clarified the meaning of the content of each judgment node with the workers and reviewers involved in annotation and review process, mistakes still occur in the three types of questions: role behavior, event relationship and plot analysis. For example, when we are determining whether the question is involving the occurrence of events to classify the question into role behaviour category or fiction plot category, we are referring to the conditions that the event needs to occur, such as the person and the reason. When the question is about where and when the event occurs, these are not necessary conditions but attributes of the event. Besides, when we are determining whether the question containing only event, we should first pretend answer is filled in the question and then determine if the question contains two complete events, that is, necessary conditions that the event needs to occur.

For instance, "Regarding the fiction *'Water Mar-*

[2] https://book.douban.com/

Q. Category	**Description**	**Example**
Character relationships	Relationships between two characters, such as master and apprentice, lovers, and so on.	Regarding the fiction *“The Return of the Condor Heroes”* , who is Yang Guo’s favorite master? A. Little Dragon girl B. Huang Rong C. Guo Jing D. Master Jin Lun
Characterization	The emotional transformation and personality change of a character in the story.	Regarding the novel *“Pride and Prejudice”* , what are the character traits of Mr. Darcy? A. He is arrogant B. He is ruthless C. He is cold D. He is kind
Literary style	The literary style , e.g., expository, narrative.	Regarding the fiction *“White Night Walk”* , what is the genre of the fiction? A. Fantasy novel B. Fairy novel C. Mystery novel D. Historical novel
Role behavior	The connections between the role and his/her behavior, including the reasons for the role to do the behavior and so on.	Regarding the novel *“The Kite Runner”* , why did Amir win the championship in a kite competition in 1975? A. In order to get the championship prize B. To stand out in front of friends C. To win the favor of my father D. To win a bet
Event relation	The relations and effects of events described in a fiction, e.g., causation, correlation.	Regarding the fiction *“Xu Sanguan Selling Blood”* , what relationship does Xu Sanguan sell blood the second time and Yi Le injure others? A. No relationship B. Mutually exclusive relationship C. Causal relationship D. Time relationship
Fiction plot	The reason and background of the plot and events in a fiction.	Regarding the novel *“Water Margin”* , what was the final result of Liang Shanbo’s heroes fighting against the imperial court? A. Defeated and disintegrated B. Give up the fight and submit to the court. C. Win and establish a new regime D. Continue to wander around
Background topic	The background, era background and theme of a fiction, e.g., the values and themes conveyed by the novel.	Regarding the fiction *“The White Night Walk”* , what message is the novel trying to convey? A. Positive energy B. Darkness of human nature C. Eternal goodness D. Money and depravity
Counterfactual reasoning	A situation or description that does not align with a fiction, such as a false character relationship or an event that does not exist in the fiction.	Regarding the fiction *“Fortress Besieged”* , what is the name of Fang Hung-chien and Tang Hsiu-fu’s child? A. Fang Hongtu B. Fang Fengyi C. Fang Feicong D. Characters do not exist

Table 1: Descriptions and examples of the 8 question categories.

gin’, who pulled the weeping willows” , “Regarding the fiction *’Water Margin’*, when did Zhishen pull the weeping willows?” and “Regarding the fiction *’Water Margin’*, what did Zhishen do before he pulled the weeping willows?” The first belongs to character behavior, the second is event relation, while the third belongs to fiction plot. This means that questions with similar meaning may fall into different categories when asked in different forms, and this is something to distinguish between them.

Q. Category	Average length	Count	Ratio(%)
Character relationships	27.99	120	19.20
Characterization	30.05	184	14.11
Literary style	29.80	102	7.82
Role behavior	31.80	277	21.24
Event relation	38.79	154	11.81
Fiction plot	33.89	140	10.74
Background topic	27.97	207	15.87
Counterfactual reasoning	37.07	120	9.20
Overall	31.97	1304	100

Table 2: Distribution and average length of questions in the dataset.

Category	Subcategory	Count
Novel Type	Love & Friendship	18
	Growth & Life	16
	Society & Human nature	15
	Military & History	15
	Crime & Mystery	15
	Science fiction & Fantasy	8
	Fables & Philosophies	5
	Literature & Art	3
Character numbers	30k-100k	31
	100k-1m	58
	over 1m	6
Publish year	before 1900	7
	1900-1950	18
	1951-2000	43
	2001-now	27

Table 3: Dataset statistics.

3.3. Collecting Questions

We design a fine-grained and strict procedure for data annotation, which ensures the quality of LFED. The procedure is well illustrated in Figure 1. We hire eight crowdsourced workers, all senior students from the Faculty of Arts, and their reading volume can cover the 95 novels we have selected. We also hire two experts to review the answers. Both annotators and experts meet the requirements of having extensive reading experience and a strong understanding of literary works. The crowdsourced workers follow the annotation convention and our carefully defined question taxonomy (shown in Section 3.2) to create questions and answers for their assigned fictions. Each novel are assigned to at least 2 workers. Each created data instance, including a question, four options, an answer, and associated annotations (e.g., question type according to our question taxonomy), is reviewed by two experts, and if one expert disagrees with the results, the question and feedback are sent back to the worker for recomposition until the questions can pass both experts' reviews.

During the expert checking process, both of them should answer a series of questions as follows:

1. Are there grammatical errors or typos in the questions and answers given?

2. If the category of the question is not counterfactual reasoning, is the category annotated by the crowdsourced worker correct according to the defined question taxonomy?

3. If the category of the question is counterfactual reasoning, is the question contrary to the content of the novel?

4. Is only one of the given multiple answers for the question is correct and can be selected as the final answer to the question?

Ultimately, we observe affirmative response rates of 0.84%, 87.92% (1,041 out of 1,184), 98.33% (118 out of 120) and 99.85% to these four questions answered by reviewers, respectively. The accuracy of annotation is low relatively. This mainly stems from the fact that a novel is annotated by at least two annotators, and they have the difficulty in achieving consensus on three main question types mentioned in section 3.2, which are role behavior, event relation and fiction plot. In addition, we also find that if a wrong question is annotated by one crowdcourced worker, the wrong question is prone to be one of the above three question types, because he/she misunderstands judgment nodes in decision-tree-style question taxonomy.

We identify questions in a format of "Regarding the fiction *'Fiction Name'*, which choice is wrong / right?" because these questions can not be categorized according the question taxonomy and can always be transferred to a question format adhering to the question taxonomy without changing the question meaning. Since each novel is annotated by more than one annotator, there are some redundancies in questions created by different annotators but related to the same novel. However, the proportion of such cases is very low. We filter out questions that have similar meaning and obtain 1,304 unique questions covering all 8 question categories pre-designed in our question taxonomy.

Model	Size	Base Model
ChatGPT	175B	instruct GPT (Ouyang et al., 2022)
ChatGLM-6B	6.2B	GLM (Du et al., 2022)
BELLE-7B-0.2M	7.1B(0.25M intructions)	bloomz-7b1-mt
BELLE-7B-2M	7.1B(2.05M intructions)	bloomz-7b1-mt
BLOOM-560M	560M	bloomz-560m
BLOOM-1B7	1.7B	bloomz-1b7

Table 4: Evaluated large language models.

3.4. Dataset Statistics

Table 2 provides the distribution and average length of questions in the created dataset. It reveals that 277 out of 1,304 questions are on *Role behavior*, accounting for 21.24% . This is followed by questions of *Chatacter relationships* (19.20%) and *Background topic* (15.87%).

Table 3 shows the additional statistics on LFED. First, among the 1,304 selected fictions, *Love & Friendship* emerges as the most prevalent themes, while fictions *Literature & Art* themes are relatively less common. Specifically, 66 of the fictions are originally written in Chinese, accounting for 69.47%. Second, the majority of selected fictions (i.e., 64 fictions) are very long, containing more than 100K Chinese characters (longer than the context window of most LLMs). 6 fictions are even longer than 1M Chinese characters, far beyond the largest context window of current LLMs. Besides, Table 3 showcases the time periods of selected fictions, which cover several centuries, with the earliest dating back to the 14th century.

The average length of questions is 25.516 characters. Each question is accompanied with four answer choices.

4. Experiments

We evaluated a number of LLMs on the curated dataset to investigate the capability of current LLMs on fiction comprhension.

4.1. Settings

LLMs We evaluated 6 large language models, which are displayed in Table 4.

Prompts We conducted zero-shot and few-shot tests of LLMs on the LFED. For zero-shot tests, we utilized two types of prompts: long prompt and short prompt. The short prompt is in a uniform format of "Select the desired answer based on the given question" while the long prompt is "Give you a multiple-choice question about a fiction, and you need to provide an answer. The provided string can be divided into three parts: the first part represents the title of the fiction, the second part is a question about the fiction, the third part includes four answer choices. Please only output the answer indicator (e.g., A, B, C or D)." In the few-shot tests, we augmented the long and short prompts used in the zero-shot tests by providing n examples based on the number of shots, where n is an integer from 0 to 5. And we make sure that the examples provided to LLMs in the prompt are from different novels.

The input for all large language models consisted of the prompt, question, answer choices, and the suffix "the correct answer is:" .

Evaluation Process Given the prompt, LLMs may not only output the answer choice indicator. For example, we find that ChatGPT under the zero-shot setting usually output not only the answer indicator but also rationals for the answer. We hence provide a script to deal with this issue. We run this process at most three times, or we will treat as LLMs answer wrong, because some model outputs will stay the same after multiple iterations:

1. Checking whether the output contains only one answer indicator. If so, the corresponding answer indicator is treated as the answer.

2. If the output contains multiple answer indicators, we choose the indicator occurring most frequently in the output as the answer.

3. If the above conditions are not met, we change the suffix to "According to the above question, please output the answer directly and do not output any rationals. " .

All text inputs to LLMs are in Chinese, as we are evaluating Chinese LLMs.

Evaluation Integrity The potential exposure of novels in LFED to LLMs during their pre-training stage is acknowledged, but measures have been taken to ensure that the evaluation reflects the capabilities of LLMs in "truly understanding" these novels. These measures include:

1. Manual Annotation and Review: This guarantees that the created questions are unique and not present in pre-training and alignment training data of LLMs. Additionally, we require hu-

Model	Short Prompt		Long Prompt	
	zero-shot	few-shot	zero-shot	few-shot
ChatGPT	**57.08**%	**51.58**%	**51.53**%	**47.26**%
ChatGLM-6B	39.15%	32.65%	34.14%	30.45%
BLOOM-560M	31.57%	31.71%	28.46%	33.06%
BLOOM-1B7	28.05%	31.23%	29.99%	32.18%
BELLE-7B-0.2M	42.18%	21.35%	38.48%	23.71%
BELLE-7B-2M	39.95%	20.13%	39.78%	35.46%

Table 5: Zero-shot results and average results under the few-shot setting (over different shots) with different prompts.

Model	PT	CR	CH	LS	RB	ER	FP	BT	CRE
ChatGPT	sp	**45.00%**	**67.39%**	**65.69%**	**50.54%**	**51.30%**	**60.71%**	**61.84%**	54.17%
	lp	**45.83%**	**60.87%**	**65.69%**	**49.46%**	**50.65%**	**60.00%**	**64.73%**	15.00%
ChatGLM-6B	sp	40.00%	46.20%	55.88%	42.96%	35.71%	36.43%	46.86%	9.17%
	lp	33.33%	40.22%	38.24%	36.10%	30.52%	40.00%	46.38%	8.33%
BLOOM-560M	sp	33.33%	33.15%	29.41%	28.52%	34.42%	35.71%	37.20%	20.83%
	lp	31.67%	28.26%	32.35%	31.05%	21.43%	30.00%	36.23%	16.67%
BLOOM-1B7	sp	28.33%	30.98%	25.49%	29.96%	33.12%	27.14%	37.68%	11.67%
	lp	26.67%	34.24%	25.49%	32.13%	33.77%	30.00%	43.48%	14.17%
BELLE-7B-0.2M	sp	44.17%	45.65%	46.08%	33.21%	21.43%	44.29%	43.48%	**59.17%**
	lp	38.33%	38.59%	43.14%	31.77%	37.66%	37.14%	45.41%	35.83%
BELLE-7B-2M	sp	40.00%	39.67%	47.06%	32.85%	20.78%	41.43%	44.44%	53.33%
	lp	37.50%	42.39%	49.02%	33.57%	22.08%	35.00%	47.83%	**50.83%**

Table 6: Zero-shot results over different question categories. **sp**: short prompt; **lp**: long prompt; **PT**: prompt type; **CR**: Characterization; **CH**: Characterization; **LS**: Literary style; **RB**: Role behavior; **ER**: Event Relation; **FP**: Fiction plot; **BT**: Background topic; **CRE** Counterfactual reasoning.

man annotators to focus on various cognitive abilities when producing questions.

2. Diverse Question Design: We encourage human annotators to create diverse questions, including those on assessing abilities in fact understanding, logical reasoning, etc.

4.2. Results

Overall The dataset is mainly developed to evaluate two capabilities of LLMs: the ability of applying knowledge of long novels obtained during training and the ability of reasoning over long novels. The zero- and few-shot results do have some implications for this evaluation goal. Besides, the design of long and short prompts can also give some clues in evaluating LLMs. Results are presented in Table 5. We observe that ChatGPT achieves the highest performance under the zero- and few-shot setting with both short and long prompts, reaching accuracy of 57.08% in zero-shot with short prompt. While the accuracy gaps between other evaluated LLMs and CharGPT are around 20% or even greater.

Zero-shot results over different question categories are reported in Table 6. We observe that ChatGPT has the best performance in other categories except counterfactual reasoning. On the counterfactual reasoning, Belle-7b-0.2M performed best with short prompt, reaching accuracy of 59.17%, while Belle-7b-2M performed best with long prompt, reaching accuracy of 50.83%.

Table 7 displays average few-shot results of the evaluated LLMs over different categories. Surprisingly, ChatGPT achieves the highest accuracy across all question categories with short prompts and long prompts.

Further Analysis As shown in Table 5, among evaluated LLMs, BLOOM series models may have deeper understand of long prompts as these models has higher performance in few-shot setting or with long prompt, while other evaluated models perform poorly in these settings.

Let's compare Table 6 and Table 7 together. We can see that ChatGPT has the best overall performance among the evaluated LLMs across all experiments, and leads almost all other evaluated LLMS by more than 10% in most experiments. The lowest results on counterfactual reasoning and event relation indicate that LLMs struggle on reasoning on extremely long documents while relatively good re-

Model	PT	CR	CH	LS	RB	ER	FP	BT	CRE
ChatGPT	sp	**44.27%**	**61.85%**	**68.63%**	**45.63%**	**48.44%**	**55.14%**	**59.71%**	**29.00%**
	lp	**40.35%**	**54.89%**	**61.76%**	**47.00%**	**47.01%**	**53.00%**	**53.72%**	**20.33%**
ChatGLM-6B	sp	35.04%	40.54%	37.84%	31.99%	34.42%	35.29%	37.10%	9.00%
	lp	33.33%	36.74%	31.57%	31.12%	31.69%	32.43%	37.87%	8.83%
BLOOM-560M	sp	36.08%	34.46%	33.14%	34.30%	34.68%	36.43%	39.61%	5.00%
	lp	34.88%	35.00%	33.33%	35.96%	34.81%	40.29%	43.19%	7.00%
BLOOM-1B7	sp	33.17%	36.96%	35.88%	33.57%	37.14%	32.14%	37.10%	3.83%
	lp	33.35%	37.07%	32.94%	34.08%	37.14%	36.00%	40.68%	6.17%
BELLE-7B-0.2M	sp	23.07%	24.46%	24.51%	19.78%	19.61%	19.29%	17.29%	22.83%
	lp	21.90%	25.65%	29.41%	25.34%	27.01%	24.43%	20.77%	15.17%
BELLE-7B-2M	sp	21.54%	20.11%	23.14%	18.84%	17.40%	17.00%	14.98%	28.00%
	lp	41.05%	39.89%	44.31%	34.08%	37.79%	33.86%	44.35%	8.33%

Table 7: Average few-shot results over different question categories. **sp**: short prompt; **lp**: long prompt; **PT**: prompt type; **CR**: Characterization; **CH**: Characterization; **LS**: Literary style; **RB**: Role behavior; **ER**:

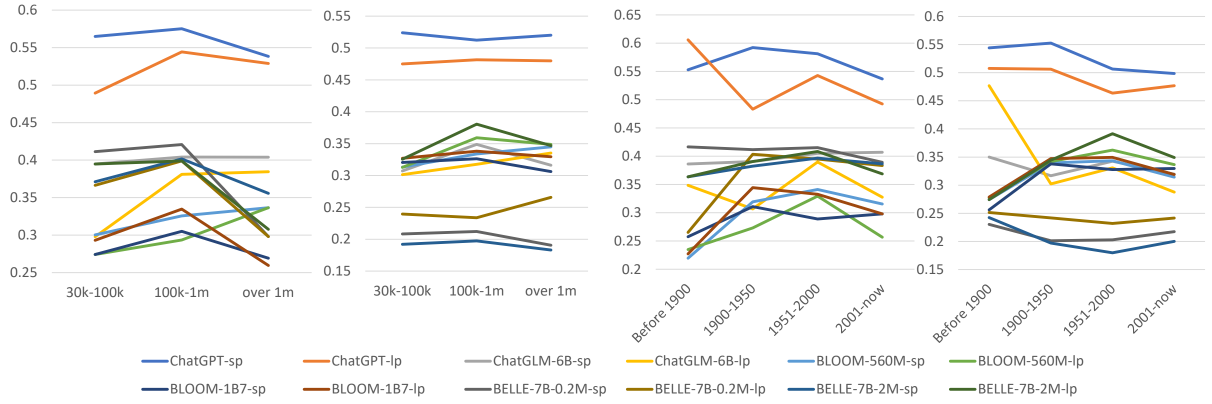

Figure 3: Results on different novel attributions under the zero- and few-shot setting. The suffixes **-sp** and **-lp** in the model name represent short prompt and long prompt respectively. The left two subfigures demonstrate results on different range of chatacter numbers under the zero- and few-shot setting respectively. The right two subfigures demonstrate results on different range of publish years under the zero- and few-shot setting respectively.

sults on characterization, background topic and literature style suggest that LLMs performs better on knowledge acquisition on novels than reasoning on novels. Surprisingly, BELLE-7Bt-0.2M and BELLE-7B-2M achieve high accuracies on counterfactual reasoning. This may be because a large proportion of questions in the training data which are "Unanswerable" , rather than having strong reasoning capabilities because the two models perform generally in event relation, a category of questions that require reasoning. This makes these two models more inclined to choose "unanswerable" as the answer, so the accuarcy on counterfactual reasoning are relatively high.

Besides, results shown in Table 6 and Table 7 indicate significant variability in performance across different question types and prompt lengths. For instance, it is noteworthy that the short prompts tend to yield higher accuracy in certain categories like Characterization and Fiction Plot, whereas long prompts seem to facilitate better performance in Background Topic and Counterfactual Reasoning.

We studied whether the number of characters of selected novel and the year of publication the novel was published would have an impact on the results. The results are shown in Figure 3. Most models perform better on novels with character numbers between 100,000 and 1 million in zero- and few-shot setting. It's possible that it's more difficult to acquire knowledge and make inferences in shorter or longer novels. Because shorter novels may have omissions in the storyline, this will make it difficult for LLMs to acquire knowledge. Extreme long novels (over 1 million characters in our dataset) have complex character relationships and storylines, which can make it difficult for LLMs to extract the right knowledge and various relations. Through the information obtain from the right two subfigures in Figure 3, we can find that there is no obvious trend in the results of the model on novels

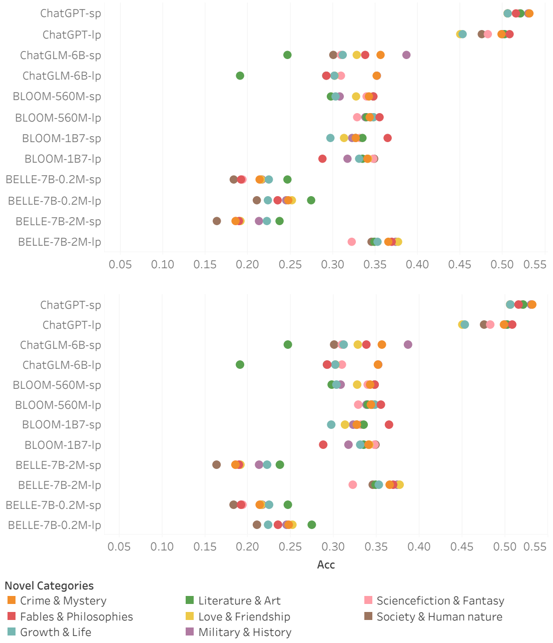

Figure 4: Results on different novel types under the zero- and few-shot setting. The suffixes **-sp** and **-lp** in the model name represent short prompt and long prompt respectively. The top figure shows zero-shot results while the bottom one demonstrates few-shot results.

of different publication years, which shows that the differences in language habits reflected in different times of novels do not affect the evaluation of the model.

We also evaluated performance by the type of novels. Results (see Figure 4) show that the performance of ChatGLM-6B varies greatly across different novel types with both long and short prompts, and it performs poorly on literary style questions under both zero- and few-shot setting. Besides, we can also observe that the performance of BELLE-7B-0.2M on different novel categories decreases substantially with long prompt when several demonstrations are provided, indicating that BELLE-7B-0.2M does not understand long prompt and learn from demonstrations well.

5. Conclusion

We have presented LFED, a literary fiction evaluation dataset which consisting of 1,304 questions from 95 fictions, designed to evaluate the reasoning capability of large language models. Our dataset curation features carefully selected fictions, a question taxonomy aiming at diversity and coverage, engaged workers and reviewers. Experiments demonstrate that the dataset is challenging for state-of-the-art LLMs under both zero- and few-shot setting.

Acknowledgements

The present research was partially supported by Huawei and Zhejiang Lab (No. 2022KH0AB01). We would like to thank the anonymous reviewers for their insightful comments.

6. Bibliographical References

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5882–5888. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 320–335. Association for Computational Linguistics.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, Heng Wang, and Zhiyuan Liu. 2019. CJRC: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 of *Lecture Notes in Computer Science*, pages 439–451. Springer.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Yufei Huang and Deyi Xiong. 2023. CBBQ: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. *CoRR*, abs/2306.16244.

Yimin Jing, Deyi Xiong, and Yan Zhen. 2019. BiPaR: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. In *Proceedings of the 2019 Conference* *on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2452–2462. Association for Computational Linguistics.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794. Association for Computational Linguistics.

Chuang Liu, Renren Jin, Yuqi Ren, and Deyi Xiong. 2024. LHMKE: A large-scale holistic multi-subject knowledge evaluation benchmark for chinese large language models.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, Xiaowen Su, Qun Liu, and Deyi Xiong. 2023. M3KE: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *CoRR*, abs/2305.10263.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 193–203. ACL.

Chih-Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. DRCD: a chinese machine reading comprehension dataset. *CoRR*, abs/1806.00920.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *CoRR*, abs/2312.16132.

Shusheng Xu, Yichen Liu, Xiaoyu Yi, Siyuan Zhou, Huizi Li, and Yi Wu. 2021. Native chinese reader: A dataset towards native-level chinese machine reading comprehension. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Bo Zhou, Qianglong Chen, Xiaomi Zhong, and Yin Zhang. 2023. WYWEB: A NLP evaluation benchmark for classical chinese. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3294–3319. Association for Computational Linguistics.