Let LLMs Take on the Latest Challenges ! A Chinese Dynamic Question Answering Benchmark

Anonymous ACL submission

Abstract

How to better evaluate the capabilities of Large Language Models (LLMs) is the focal point and hot topic in current LLMs research. Previous work has noted that due to the extremely high cost of iterative updates of LLMs, they are often unable to answer the latest dynamic questions well. To promote the improvement of Chinese LLMs' ability to answer dynamic questions, in this paper, we introduce **CDQA**, a **C**hinese **D**ynamic **QA** benchmark containing question-answer pairs related to the latest news on the Chinese Internet. We obtain high-quality data through a pipeline that combines humans and models, and carefully classify the samples according to the frequency of answer changes to facilitate a more fine-grained observation of LLMs' capabilities. We have also evaluated and analyzed mainstream and advanced Chinese LLMs on *CDQA*. Extensive experiments and valuable insights suggest that our proposed *CDQA* is challenging and worthy of more further study [1]. We believe that the benchmark we provide will become one of the key data resources for improving LLMs' Chinese question-answering ability in the future.

1 Introduction

Due to the excellent emergence capabilities and unified task paradigm, Large Language Models (LLMs) are undoubtedly the more popular stars in the field of Natural Language Processing (NLP) or Artificial Intelligence (Wei et al., 2022; Li et al., 2023; Shanahan, 2024). To promote the improvement of LLMs capabilities, more and more researchers have invested in building various LLMs evaluation benchmarks (Chang et al., 2023; Huang et al., 2023a). In the era of LLMs, high-quality evaluation benchmarks allow researchers to better understand the capabilities of LLMs, thereby stimulating further research on how to enhance LLMs.

Static Question	ACL 主会每年举办几次? How many times does the ACL annual meeting take place each year?	
GPT-4's Answer	一年一次。 Once a year.	✓
Dynamic Question	下一次ACL 将在哪里举办? Where will the next ACL be held?	
GPT-4's Answer	我无法提供相关信息。 I can't provide the information.	✗

Table 1: Examples of static and dynamic questions. The **GPT-4** is on Feb 11, 2024.

Question answering is an important and long-standing topic in NLP (Rajpurkar et al., 2016; Joshi et al., 2017; He et al., 2018). Especially for LLMs, QA tasks have almost become the indispensable basic task in LLMs research (Pan et al., 2024). Various forms of QA benchmarks can be used to measure the capabilities of LLMs in different dimensions (Adlakha et al., 2022; Bosselut et al., 2022; Rein et al., 2023; Huang et al., 2023b). Recently, the introduction of English FreshQA (Vu et al., 2023) has attracted widespread attention. It challenges LLMs through questions with dynamically changing answers, aiming to test LLMs' mastery of the latest factual knowledge. Obviously, being able to answer the latest questions determines to some extent whether LLMs can truly move towards large-scale daily applications. **Urgently, we note that there is still no such benchmark in the Chinese community, although LLMs in the Chinese scenario still face the same challenges and dilemmas**, as shown in Table 1.

To let LLMs in Chinese scenarios take on the latest challenges and empower them to answer dynamic questions, in this work, we present **CDQA**, a **C**hinese **D**ynamic **QA** benchmark. Specifically, we design a semi-automatic data production pipeline to construct our benchmark. In this pipeline, we first automatically generate a large number of raw

[1]Our dataset and code will be publicly available after the anonymous review period.

queries with the help of two LLMs with different roles, one is to extract key entities from the latest Chinese news, and the other is to automatically generate question queries based on the extracted entities that will be as the corresponding answers. Then we ask the well-trained annotators to filter, rewrite, and classify the automatically generated question samples to ensure the quality of *CDQA*. Through such a semi-automatic data construction method with human participation, we obtain 1,339 question-answer pairs for *CDQA*, classified by how frequently their answers change (i.e., fast-changing, slow-changing, and never-changing). The purpose of classifying *CDQA* samples by the frequency of answer changes is to provide finer-grained evaluation for LLMs, facilitating researchers to better perceive the true performance of LLMs.

Based on our constructed *CDQA*, we select a series of widely used and advanced LLMs in the Chinese community for evaluation. Results show that **Qwen1.5-72B-Chat** performs the best across all models with retrieval augmentation as it has better Chinese instruction following abilities and related knowledge while **Deepseek-67B-Chat** has the best knowledge of our questions without retrieval augmentation and **GPT-4** is weak at Chinese knowledge but has better retrieval augmented generation (RAG) ability than the Deepseek model. However, no LLM baselines achieves above 40 and 70 in F1-recall scores by standalone and RAG respectively, demonstrating the challenge of our dataset. Besides, **in-context learning** and **prompting methods** like Chain-of-Thought generally increase performances with searched evidence but also elicit more hallucinations in LLMs. For **search engines** in the RAG scenario, Google consistently takes advantage over Bing for all baseline models, showing its strength as a good retriever for LLMs.

In summary, the contributions of our work are summarized as follows:

1. We first introduce the idea of using dynamic questions to challenge Chinese LLMs, which provides a new direction for the development of LLMs in Chinese community.

2. We construct the high-quality *CDQA* benchmark composed of dynamic questions, which will become an important data resource for promoting the progress of Chinese LLMs.

3. Extensive experiments and detailed analyses based on *CDQA* provide valuable insights and discoveries, which are instructive for subsequent research about how to enhance LLMs to handle dynamic questions.

2 Chinese Dynamic Question Answering (CDQA)

2.1 Overview

Our **CDQA** mainly originates from latest news in Chinese Internet from different areas such as finance, daily life, politics, technology and so on. Besides, there are also queries collected from Chinese labelers. They represent the information-seeking cases of Chinese people. The generation pipeline could be illustrated in Figure 1. The dataset currently consists of 1,339 questions covering a range of topics with evolving answers which are mostly extracted entities from the raw corpus scraped from Chinese Internet and it is being regularly updated. We believe this initial data scale is suitable for benchmarking LLMs in the dynamic QA challenge(Joshi et al., 2017; Kasai et al., 2022; Rein et al., 2023; Vu et al., 2023; Mialon et al., 2023).

2.2 Data Collection

We collect *CDQA* dataset in two stage. **The first stage is automatic generations with Entity Extraction and Doc2Query**, for which we use SeqGPT (Yu et al., 2023), and GPT-4 (OpenAI, 2023), which could give great amount of raw question-answering pairs as SeqGPT extracts entities from latest Chinese news and GPT-4 is prompted into generating corresponding questions. For GPT-4 prompts, we use few-shot prompting in generating diverse questions from entities. **The second stage is manual labeling from crowd-sourced workers**. The Chinese labelers not only filter questions which are answered with biases, ambiguities and obsolete[2] knowledge but also annotate with **tags**, check the correctness and **rewrite** the question answer pairs to be more time-related and dynamic. At the very beginning, the labelers are shown with pre-annotation examples and annotation guides.

Tags The tags are annotated for questions and answers. For questions, we have the same taxonomy as *FreshQA* (Vu et al., 2023). The questions are categorized as **fast-changing**, **slow-changing**, and **never-changing**. For answers, we categorize these entities or short texts as **person**, **location**, **time**,

[2]The answer should be only supported with the knowledge after Jan 1, 2019 except for static knowledge, i.e., never-changing.

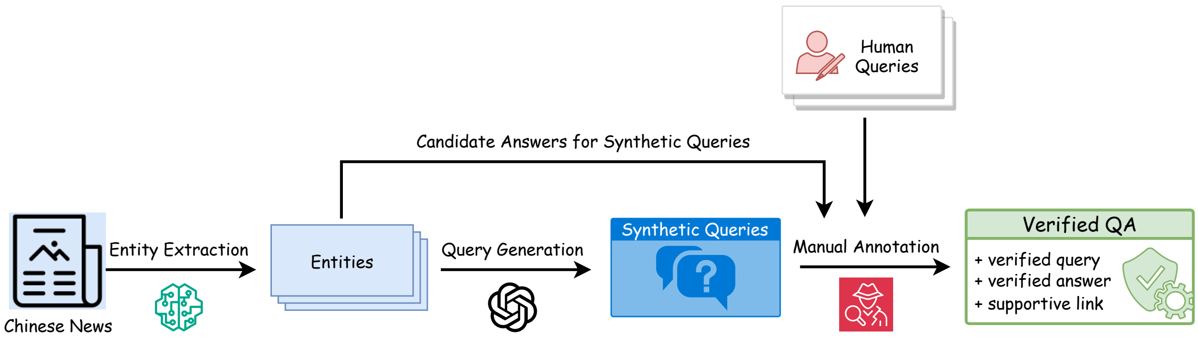


Figure 1: Data Generation Pipeline for **CDQA** dataset. We first collect Chinese News from Internet and then extract entities from these news passages. Based on GPT-4, we generate synthetic queries from passages and corresponding entities. Manual annotation is conducted to verify the synthetic data and extra human-crafted queries, providing the verified queries, answers and supportive evidence links.

event, **artificial work**, **group**, **nature**, **quantity** and **other**. Therefore, we could evaluate the models' latest world knowledge from various perspectives. The taxonomy and corresponding examples are illustrated in Appendix A.

Quality Control After getting the synthetic queries, the human annotators could rewrite and calibrate the questions and answers to make the QA pairs correct, consistent and dynamic. For example, annotators are required to provide the supporting evidence `URLs` along with correct answers using search engines. This calibration process could solidify our answers with supplementary valid information and help us better iterate the dataset as the generation process in the previous stage is not well-evaluated with supportive documents, let alone the correctness. Moreover, in order to facilitate the periodic updates, we filter out the questions with more than one valid answer.

For inter-annotator agreement, we randomly sample 100 examples from synthetic question-answer pairs and annotations from two annotators in the same annotation vendor are measured by **acceptance** (*whether the pair is accepted or discarded*), **question tags** and **answer types**. The ground-truth labels are provided by authors. For each category, we calculate their Cohen Kappa scores (McHugh, 2012). From Table 2, the averaged score across all types of annotations are above 63.1, representing "substantial agreement" for our dataset annotations.

	Acceptance	Question Tags	Answer Types
Ann1 v.s. Ann2	62.3	87.2	96.6
GT v.s. Ann1	79.6	59.1	100
GT v.s. Ann2	47.3	68.3	100
Avg	**63.1**	**71.5**	**98.9**

Table 2: Inter-annotator agreement for different annotation sections are calculated by **Cohen Kappa scores**. Ann1/2 represents Annotator1/2 respectively and GT represents Ground Truth. Our annotations could be considered as "substantial agreement" as the average scores are above 60.

2.3 Regular Updates

Our dataset is highly sensitive to time since the ground truth is evolving along the world development. Therefore, we commit to updating the dataset regularly and researchers are strongly encouraged to stay tuned with our latest version for evaluation. And the datasets are mainly calibrated with information from Chinese Internet. Currently, we are going to maintain it yearly.

2.4 Data Statistics

Due to limitations in automatic query generation by GPT-4 and SeqGPT from the first stage, our dataset has low **retention rate** in which only 44.6% synthetic data are accepted by human annotators. Among the accepted data, 53.1% of them still need further modifications because of improper questions or wrong answers. For **question tags**, we have relatively balanced distributions between *fast-changing* and *slow-changing* questions with fewer *never-changing* questions. For **answer types**, we have biased distributions as nearly 70% of entities extracted from passages lie in "person" and "group" categories. This is because most of entities in first stage by automatic generation are "person" and "group". However, question tags and answer types could be changed or calibrated over time by re-annotation of the dataset. These distribution graphs and more analysis about our dataset are in Appendix B.

2.5 Evaluation

As *CDQA* is constructed from Internet, our evaluation is mainly based on **retrieval-augmented generation (RAG)** (Chen et al., 2017; Gao et al., 2023) of LLMs with different search engines and the evaluation metrics are *answer rate* and *F1-recall*. Results from standalone LLMs are used as comparison. Overall, our evaluation provides a comprehensive understanding of current LLMs in factuality, especially for evolving knowledge. Besides, due to the safety implementation for different LLMs from helpful and harmless responses in training data (Bai et al., 2022), **F1-recall only counts on questions with effective responses by default** while **answer rate is used in representing the ratio of answered questions to the total questions**, which is a practical metric for the real world application of LLMs and could directly indicate the degree of hallucination in generated responses.

Evaluation Metrics For **F1-recall**, we calculate *the ratio of common tokens between model-generated responses and ground truth to the ground truth*. Specifically, we first segment the generated text and golden text into token lists using word segmentation tools [3], then calculate the ratio of tokens generated by models belonging to the golden token list to golden tokens. For **answer rate**, we directly calculate *the ratio of effectively answered questions to total questions*, i.e., responses of refusal, summarized from our empirical observations on predictions from these baseline LLMs, are filtered out in our evaluation.

3 Experiments

3.1 Experiment Setup

Baselines We experiment with a series of baseline models pretrained with Chinese data, including **Qwen1.5-72B-Chat** (Bai et al., 2023), OpenAI's **ChatGPT** (*gpt-3.5-turbo-1106*) (OpenAI, 2022) and **GPT-4** (*gpt-4-1106-preview*) (OpenAI, 2023), open-sourced Chinese-oriented models such as **Internlm2-20B-Chat** (Cai et al., 2024), **Aquila2-34B-Chat** (BAAI, 2023), **Yi-34B-Chat** (01-ai, 2023), **Deepseek-67B-Chat** (DeepSeek-AI, 2024). In the close-book scenario, we only use the standalone LLM to directly answer questions. For the open-book scenario, we use retrieval augmented generation with LLMs in which search engines are used for retrieving question-related results on the Internet and then fed into LLMs for reading.

[3] https://github.com/fxsjy/jieba

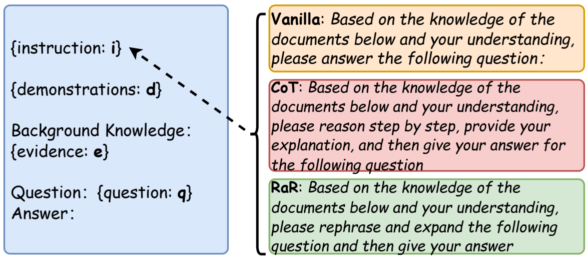


Figure 2: Our prompts are formulated under this framework. Different prompting methods are used with different instructions **i**. The Chinese version of our prompts is in Appendix C.

Search Engines Except for language models for information synthesis, we select two representative search engines to recall relevant passages from the Chinese Internet namely **Google** and **Bing**. These search engines are mainly used by Chinese people for information seeking. **Baidu** is omitted due to the difficulty in scraping its contents. The Top-10 searched results are provided to models in the **RAG** setting.

Prompt Design Our prompt framework, which is in Chinese, could be framed as concatenation of $(\mathbf{i}, \mathbf{d}, \mathbf{e}, \mathbf{q})$, in Figure 2 where **i** represents the instruction, **d** for question-answer pairs from crowdsourced labelers, **e** for search results and **q** for current question. Different instructions **i** are used with three widely adopted prompting styles, **Vanilla**, **Chain-of-Thought (CoT)** (Wei et al., 2023) and **Rephrase-and-Respond (RaR)** (Deng et al., 2023). **Vanilla** instruction is directly asking models to answer questions with the context. **CoT** instruction is asking models to first explain and analyze the question **q** step by step and then give their answers. **RaR** instruction, however, is asking models to first rephrase and expand the question **q** and then give their answers, which could be viewed as a complement of CoT as CoT is for diving deeper while RaR is for exploring broader. Besides, for demonstrations **d**, we have used zero-shot and different few-shot settings, i.e., 5-shot and 16-shot. More specifically, our few-shot demonstrations are made up of human written questions and answers similar to *CDQA* dataset without contexts or other explanations as it costs longer time without any improvement.

Models	fast-changing		slow-changing		never-changing		average F1-recall	
	w/o RAG	RAG	w/o RAG	RAG	w/o RAG	RAG	w/o RAG	RAG
Internlm2-20B-Chat	18.0 (99.6%)	58.4	17.8	68.2	34.8	77.0	23.5	67.9
Aquila2-34B-Chat	14.9	51.5	17.7	62.5	35.6	69.4	22.7	61.1
Yi-34B-Chat	22.9	56.5	30.8	68.8	46.9	76.9	33.5	67.4
Deepseek-67B-Chat	24.3	58.4	**37.2**	70.0	53.1	79.2	**38.2**	69.2
Qwen1.5-72B-Chat	28.9 (67.6%)	**65.2 (97.3%)**	29.1 (83.7%)	**72.5 (98.7%)**	**55.6 (88.4%)**	**85**	31	**73.3**
ChatGPT	18.1 (96.6%)	59.2 (98.3%)	14.1 (93.3%)	66.3 (98.3%)	34.7 (99%)	73.7 (99.7%)	21.7	65.6
GPT-4	**35.1 (13.5%)**	61.2 (96.4%)	33.8 (25.4%)	68.4 (96.5%)	54.4 (56.1%)	78.8 (98.6%)	14.6	67.6

Table 3: Best performance over different few-shot settings for **Vanilla** prompt with Top10 searched results from Google. We report in the form of *F1-recall (answer rate)* for different types of questions and omit the answer rate if it is 100%. For "average F1-Recall", they are *F1-recall* calculated **among all questions** in our dataset for better comparing baseline models. Data with the highest F1-recall scores are marked in bold.

Models	fast-changing		slow-changing		never-changing		average F1-recall	
	w/o RAG	RAG	w/o RAG	RAG	w/o RAG	RAG	w/o RAG	RAG
Internlm2-20B-Chat	16.4	55.2	17.4	64.8	34.3	72.4	22.7	64.1
Aquila2-34B-Chat	14.5	51.9	17.1	61.4	35.6	69.8	22.4	61.0
Yi-34B-Chat	23.2	57.4	30.4	68.5	47.0	77.3	33.5	67.7
Deepseek-67B-Chat	22.9	59.2	**37.0**	70.6	53.0	80.2	**37.6**	70
Qwen1.5-72B-Chat	**26.0 (86.7%)**	**71.0 (86.9%)**	26.7 (91.4%)	**77.5 (89.4%)**	**58.2 (77.2%)**	**85.6 (98.3%)**	30.6	**71.7**
ChatGPT	17.9 (97.3%)	61.4 (96.6%)	13.9 (98.3%)	65.7 (98.7%)	36.0 (99.7%)	74.9 (98.6%)	22.3	66.0
GPT-4	22.1 (82.9%)	68.0 (89.0%)	19.8 (86.7%)	74.7 (90.4%)	48.2 (56.1%)	83.5 (98.3%)	20.8	70.0

Table 4: Best performance over different few-shot settings for **CoT** prompt with Top10 searched results from Google. We report in the form of *F1-recall (answer rate)* for different types of questions and omit the answer rate if it is 100%. For "average F1-Recall", they are *F1-recall* calculated **among all questions** in our dataset for better comparing baseline models. Data with the highest F1-recall scores are marked in bold.

3.2 Results and Analyses

Table 3, 4, 5 summarize best performances over few-shot prompting across different baselines for Vanilla, CoT and RaR prompts respectively. Our default search engine for analysis is **Google** as it is most widely used around the world.

Baseline Comparison From the **average F1-recall** in above tables, we see that **Deepseek-67B-Chat** has the best performance without retrieval augmentation, showing its superior memorization of Chinese knowledge related to CDQA questions. On the contrary, **Qwen1.5-72B-Chat** ranks the best in RAG scenario, surpassing 70 in average F1-recall scores with all different prompts styles for all questions. Moreover, for detailed results among different types of questions, we notice that Qwen1.5-72B-Chat, ChatGPT and GPT-4 have higher answer rates with retrieval augmentation while other baseline models actively answer all questions (i.e. 100% answer rate) in both scenarios which indicates that these three models are aligned with hallucination reduction measures such as refusal of questions.

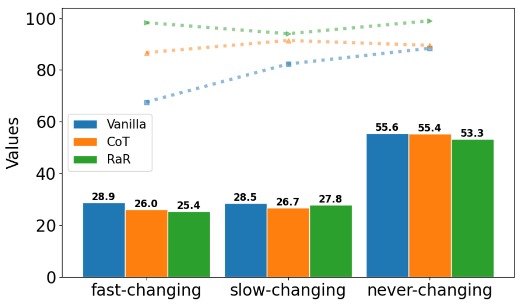
How do different styles of prompts work in LLMs? To rule out the other influences such as few-shot demonstrations, we use zero-shot setting with Qwen and GPT-4 models as open-sourced and closed representative models in the following analysis. In Figure 3, **without RAG**, we see that Qwen1.5-72B-Chat has higher answer rates over different prompts on different questions than GPT-4 and there are more apparently different answering behaviors in GPT-4. Specifically, GPT-4 answers with great care in vanilla prompts with lowest answer rates but high F1-recall scores while GPT-4 suffers from hallucination in CoT and RaR prompts with at most +522% and +176% in answer rates but -43% and -17% in F1-recall scores compared to Vanilla prompt. For both models, Vanilla prompt outperform the other two kinds of prompts with higher F1-recall scores. **This indicates that verbose explanation or expansion could increase hallucination especially when without evidence**.

In Figure 4, **with RAG**, we see that Qwen1.5-72B-Chat and GPT-4 both have fewer gaps in answer rates across different prompts and question types compared to close-book counterparts, representing adding contextual information elicits LLMs in answering questions more efficiently. Besides, with search results, CoT and RaR both outperform Vanilla prompt and CoT performs the best in GPT-4 and Qwen1.5-72B-Chat with less hallucination, i.e., lower answer rate and higher F1-recall score. **This indicates that CoT and RaR could improve LLMs on complex tasks but CoT elicits more reasoning abilities to improve the answering**.

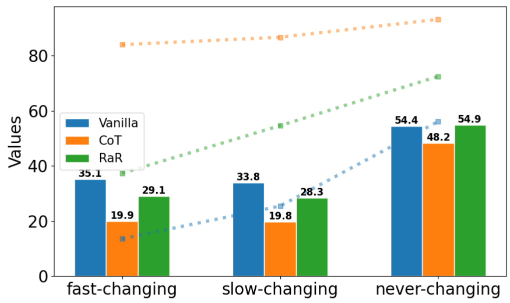
Nevertheless, model sizes and training data are both fundamental for these prompts to work. In

Models	fast-changing		slow-changing		never-changing		average F1-recall	
	no-RAG	RAG	no-RAG	RAG	no-RAG	RAG	no-RAG	RAG
Internlm2-20B-Chat	17.2	57.7	17.8	67.8	33.4	76.4	22.8	67.3
Aquila2-34B-Chat	15.5	51.4	17.5	61.9	36.1	69.5	23.0	60.9
Yi-34B-Chat	22.8	57.0	30.6	68.5	47.7	76.8	33.7	67.4
Deepseek-67B-Chat	23.3	58.9	**37.7**	70.7	54.2	79.8	**38.4**	69.8
Qwen1.5-72B-Chat	25.4 (98.3%)	**66.5 (97.1%)**	27.8 (94.0%)	**72.9 (99.6%)**	53.3 (99.0%)	**84.1**	34.6	**73.8**
ChatGPT	19.2	61.7	15.9	67.6 (99.6%)	35.6	76.5 (99.7%)	23.6	68.4
GPT-4	**29.1 (37.3%)**	64.9 (95.2%)	28.3 (54.6%)	71.6 (96.2%)	**54.9 (72.5%)**	82.6 (99.3%)	22.0	70.9

Table 5: Best performance over different few-shot settings for **RaR** prompt with Top10 searched results from Google. We report in the form of *F1-recall (answer rate)* for different types of questions and omit the answer rate if it is 100%. For "average F1-Recall", they are *F1-recall* calculated **among all questions** in our dataset for better comparing baseline models. Data with the highest F1-recall scores are marked in bold.

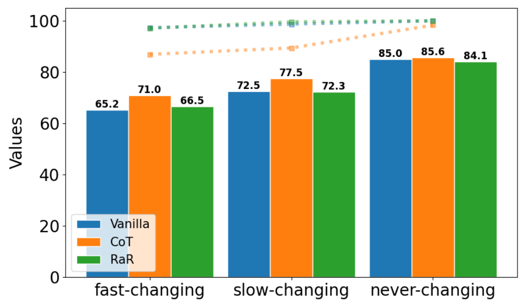


(a) Qwen1.5-72B-Chat

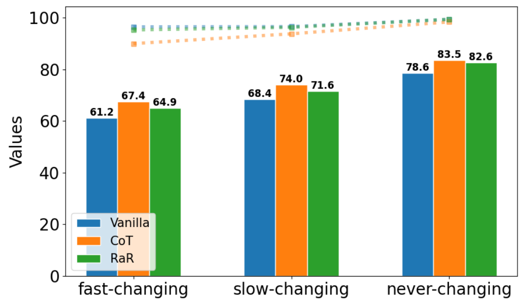


(b) GPT-4

Figure 3: F1-recall scores and Answer Rates of **different prompts** for LLMs **without RAG** under zero-shot setting. We represent F1-recall scores with bar plots and answer rates with dotted lines.



(a) Qwen1.5-72B-Chat



(b) GPT-4

Figure 4: F1-recall scores and Answer Rates of **different prompts** for LLMs **with RAG** under zero-shot setting. We represent F1-recall scores with bar plots and answer rates with dotted lines.

Figure 5, **not every model improves with CoT or RaR compared to Vanilla prompt**. For example, Deepseek-34B-Chat and Internlm2-20B-Chat's performances decrease in CoT and RaR; ChatGPT prefers RaR while Qwen1.5-72B-Chat, GPT-4 and Yi-34B-Chat prefer CoT for larger gains; Aquila2-34B-Chat is robust to all prompt types.

Does few-shot prompting always work for all LLMs? For better analyzing the influence of few-shot prompting, we collect experiments results with and without RAG in **vanilla** prompt. In Figure 6, based on nearly 100% answer rate, four (i.e. Internlm2-20B-Chat, Aquila2-34B-Chat, Yi-34B-Chat, Deepseek-67B-Chat) without RAG and three (i.e. Internlm2-20B-Chat, Yi-34B-Chat, Deepseek-67B-Chat) with RAG out of all five open-sourced Chinese-oriented models have better performance with more few-shot demonstrations, which are sampled in the same data distribution during the generation of *CDQA* dataset.

However, we also notice that Qwen1.5-72B-Chat, ChatGPT and GPT-4 have shown different trends compared to other open-sourced models, i.e., more few-shot examples lead to decreases in

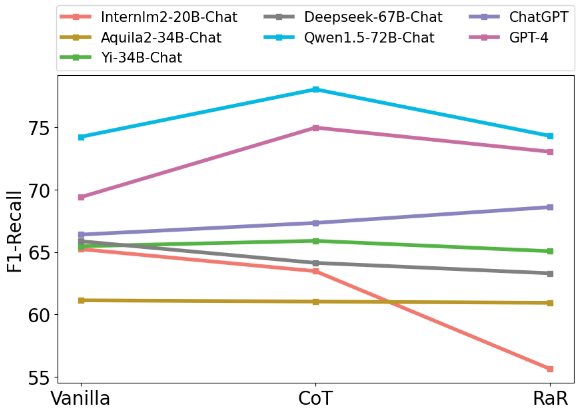


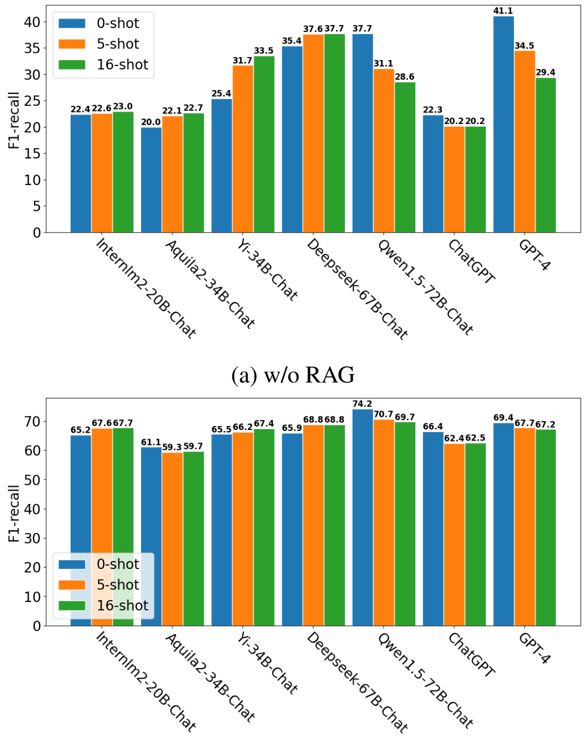
Figure 5: F1-recall scores averaged over all three different questions for all models with **different prompts** in open-book scenario under zero-shot setting. We present F1-recall score only since all answer rates $\geq$ 90%.

Models	w/o RAG			RAG		
	0-shot	5-shot	16-shot	0-shot	5-shot	16-shot
Qwen1.5-72B-Chat	79.4	86.5	93.2	98.7	98.3	**99.2**
ChatGPT	96.3	95.0	96.7	98.8	99.7	**99.9**
GPT-4	31.7	52.2	64.7	97.4	97.7	**98.0**

Table 6: Answer rates (%) for ChatGPT and GPT-4 averaged on all types of questions with **different few-shot settings**.

F1-recall scores. Therefore, we check their averaged answer rates over all types of questions in Table 6 where ChatGPT stays in fairly high answer rates ($\geq$ 95%) and Qwen1.5-72B-Chat and GPT-4 increase their answer rates with more few-shot examples. Combined with their monotonic decrease in F1-recall scores, we reveal that they hallucinate more with more few-shot examples in prompts. **This indicates that few-shot demonstrations are not always useful for LLMs. For models in weaker abilities, it might help on teaching LLMs on how to answer instructions by analogy while induce more hallucinations and distraction on LLMs**.

How do different search engines help? For fair comparison between search engines across all baselines, we use vanilla prompt under zero-shot setting as CoT and RaR have different effects on models behaviors from previous analysis. In Figure 7, searched results from Google consistently outperform Bing among all baseline models, which indicates that the **Google currently provides more useful retrieved evidence for question answering about Chinese news**.



(a) w/o RAG

(b) RAG

Figure 6: F1-recall scores averaged over all types of questions for different models with **different few-shot settings**.

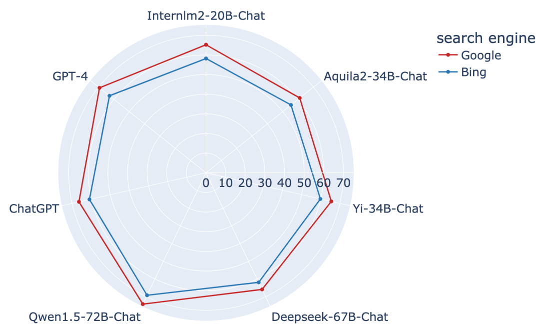


Figure 7: F1-recall scores averaged over all questions for different models with **different search engines**.

How do LLMs perform across different answer types? As answers in *CDQA* are mainly entities from news, we conduct analysis across different answer types for three representative LLMs, i.e., Deepseek-67B-Chat, Qwen1.5-72B-Chat and GPT-4. In Figure 8, we observe that **GPT-4's internal knowledge is poorer than Chinese-oriented models such as Deepseek-67B-**

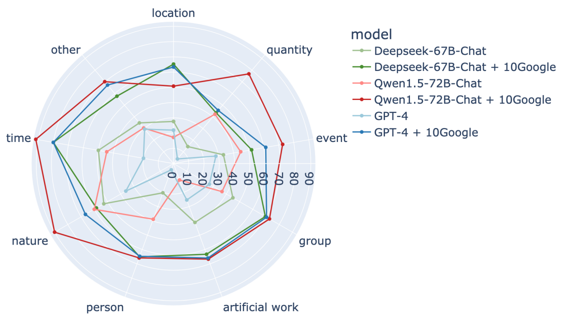


Figure 8: F1-recall scores on **different answer types** for Qwen-72B-Chat and GPT-4 in close-book and open-book scenarios with Vanilla prompt. We use Top 10 searched results from Google. Under close-book scenario, Qwen1.5-72B-Chat holds larger answer rates than GPT-4 whose drastically increases to 100% with searched results from Google.

Chat and Qwen1.5-72B-Chat for Chinese users. However, with enough retrieved evidence, **GPT-4 has stronger abilities in learning from contexts than Deepseek-67B-Chat and Qwen1.5-72B-Chat** where this "learning efficiency", i.e., *the ratio of gaps between open-book scores and close-book scores to the close-book* could reach at most 1370% compared to 219% in Deepseek-67B-Chat and 450% in Qwen1.5-72B-Chat. Moreover, from Figure 8, we also could notice that "quantity" and "location" groups are hardest for GPT-4 and Qwen1.5-72B-Chat respectively to figure out the correct answers, which is due to the granularity of answers and the need of reasoning abilities.

4 Related Work

Question Answering (QA) is a long-standing task in NLP area (Wang et al., 2024; Li et al., 2024), ranging from classic single-turn benchmarks such as *SQuAD* (Rajpurkar et al., 2016, 2018), *TriviaQA* (Joshi et al., 2017) and *Natural Questions* (Kwiatkowski et al., 2019) to conversational QA like *TopiOCQA* (Adlakha et al., 2022).

Temporal and Dynamic QA Benchmark *StreamingQA* (Liska et al., 2022) is a QA dataset where questions are generated on given dates, showing how open-book and close-book QA models adapt to new knowledge over time and importance of retrieval augmentation in up-to-date search space. *TimeQA* (Chen et al., 2021) is formed from extracted evolving facts in *WikiData* by manual extraction and verification while we extract entities to directly formulate them as answer candidates based on the documents. *RealTimeQA* (Kasai et al., 2022), a dynamic QA benchmark with automatic weekly updates from the weekly News Quiz section in social media such as CNN, is most related to our semi-automatic question generation with the latest Chinese news corpus.

Chinese QA benchmark In contrast to prosperous English QA benchmarks, Chinese counterparts are still under-explored. *DuReader* (He et al., 2018) is a classic free-form QA benchmark collected by Baidu from its own products and *CLUE* (Xu et al., 2020) is the first large scale NLU benchmark in Chinese. After the recent debut of powerful large language models, a series of Chinese QA benchmarks are proposed for better evaluating them. *C-Eval* (Huang et al., 2023b) is a multiple-choice questions answering dataset from Chinese Standard Exams. *WebCPM* (Qin et al., 2023) collects questions from web forums through web searching and browsing and *SuperCLUE* (Xu et al., 2023) is a comprehensive Chinese benchmark for question answering in aligning users needs. But they all suffer from either data leakage or the risk of saturated performance which hinders the accurate evaluation on questions requiring fresh knowledge to answer as static questions are readily overfitted.

5 Conclusion

The creation of *CDQA* addresses the urgent need for the evaluation of Chinese LLMs, thereby improving LLM-driven applications for Chinese users. Given the cultural influences in LLMs' training data, it is our aspiration that *CDQA* will foster development in various capabilities of LLMs, particularly within Chinese contexts. While *CDQA* progresses further with a semi-automatic generation pipeline with more data than *FreshQA*, we acknowledge that it is far from a perfect LLM evaluation. Other critical dimensions, including tool learning, LLMs safety, and robustness, remain to be explored. However, we believe that our constructed *CDQA* and the series of insights obtained based on it will provide valuable resources and guidance for subsequent research on Chinese LLMs. In the future, we will conduct more in-depth analyses of the capabilities of LLMs based on *CDQA* and investigate how to enhance the LLMs' ability to handle dynamic questions. This will empower LLMs to better cope with the complex and ever-changing real-world application environments.

Limitations

One of the limitations of our work is that the language we study is Chinese only. As the two most widely used languages in the world, English and Chinese have always been equally valued and widely concerned in the NLP community. In fact, our work is inspired by previous *FreshQA* in the English scenario and aims to provide similar data resources to Chinese LLMs researchers. We also encourage and welcome more researchers from other languages to engage in similar research.

In addition, another limitation that cannot be ignored is how to keep our *CDQA* updated. Because *CDQA* focuses on questions whose answers change dynamically, it is critical to ensure that the answers to questions in *CDQA* are always correct and up-to-date. Therefore, we also commit to updating our *CDQA* regularly and providing researchers with the latest version of *CDQA* for LLMs evaluation.

Ethics Statement

The task we focus on is the evaluation of LLMs, and the LLMs we evaluate are all public and widely used LLMs, so they do not bring potential ethical risks. The data samples of *CDQA* that we collect have been manually cleaned and pre-processed to ensure that they do not contain any data that will cause moral risks, such as politically sensitive, violent, and private data. In addition, we also have signed legal labor contracts with the human annotators we employ, and pay them higher than market prices based on their workload.

References

01-ai. 2023. Yi series large langugae models. https://github.com/01-ai/Yi.

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching.

BAAI. 2023. Aquila2. https://github.com/FlagAI-Open/Aquila2.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Antoine Bosselut, Xiang Li, Bill Yuchen Lin, Vered Shwartz, Bodhisattwa Prasad Majumder, Yash Kumar Lal, Rachel Rudinger, Xiang Ren, Niket Tandon, and Vilém Zouhar, editors. 2022. *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*. Association for Computational Linguistics, Dublin, Ireland.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions.

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Hai-Tao Zheng. 2023a. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. *arXiv preprint arXiv:2308.10855*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now?

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023. On the (in)effectiveness of large language models for chinese text correction. *CoRR*, abs/2307.09007.

Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S Yu. 2024. When llms meet cunning questions: A fallacy understanding benchmark for large language models. *arXiv preprint arXiv:2402.11100*.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants.

OpenAI. 2022. Chatgpt. `https://chat.openai.com/chat`.

OpenAI. 2023. Gpt-4 technical report.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. Webcpm: Interactive web search for chinese long-form question answering.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark.

Murray Shanahan. 2024. Talking about large language models. *Communications of the ACM*, 67(2):68–79.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2024. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. Clue: A chinese language understanding evaluation benchmark.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark.

Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, Hai-Tao Zheng, Ningyu Zhang, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. Seqgpt: An out-of-the-box large language model for open domain sequence understanding.

A Tag Taxonomy of CDQA

The tag taxonomy of *CDQA* and examples are presented in Table 7.

B Dataset Distributions

Knowledge types for queries and answer types are visualized in the following Figure 9, 10. More specifically, we have further visualized the answer type distributions in each question tag. From Figure 11, Figure 12 and Figure 13, we see that nearly 80% of slow changing questions are about person and group. Although it seems to be biased, CDQA is based on News articles in which who, what, when, where, why and how (5Ws and H) are key components and protagonists or characters are the most significant. So it is reasonable that our data comprises many 'persons' and 'groups' answers as they are indeed under frequent changing phase and reflect the dynamic aspect. Except for person and group, artificial work should be the third largest category for answers, which includes jobs, titles, knowledge and so on. These observations are all consistent with our data sources as information for the protagonists, places and events are compulsory and most frequent in news reports. Besides, percentages of time reach the maximum in never-changing tag as currently most of questions answered with time are about the frequencies.

As our data generation pipeline is semi-automatic, it is important to demystify how our dataset would represent the real-world dynamic QA challenge. In such, we compare the data distributions before and after the manual annotation process by t-SNE analysis where concatenated QA pairs are transformed into embedding representations. The resultant t-SNE graphical representations in Figure 14 indicate minimal alteration in the structural framework of the data. Furthermore, the spatial analysis, measured via the L2-norm distance between the centers of synthetic QA

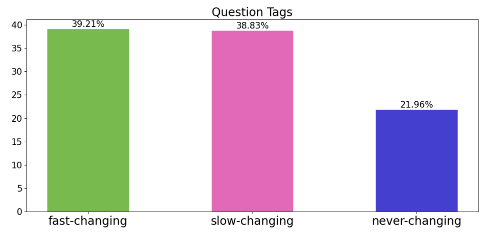


Figure 9: Distributions of question tags for full data.

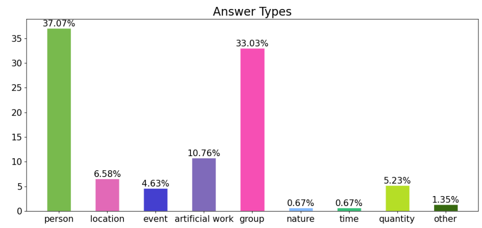


Figure 10: Distribution of answer types for full data.

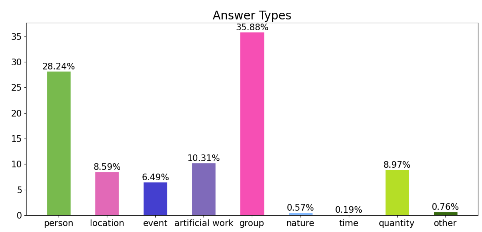


Figure 11: Distributions of answer types for fast-changing questions.

Category	Description	Example
`fast-changing`	The answer to the question is prone to changing within **one year**	(*How many sessions has the Maritime Silk Road Cultural Heritage Forum been held?, Four*)
`slow-changing`	The answer to the question is prone to changing in **several years**	(*Which ancient city site in China has recently been recognized as a UNESCO World Cultural Heritage?; the Liangzhu Ancient City Site*)
`never-changing`	The answer to the question is from **static knowledge** such as scientific theories, historical facts and so on	(*In rural areas during winter heating, it is necessary to guard against the risk of poisoning from which gas?; Carbon monoxide*)
`person`	Specific individual, usually referring to a human being.	(*Who among the current representatives of the Fuxin County People's Congress was one of the first batch of anti-epidemic heroes to rush to support Wuhan?; Xin Li*)
`location`	Geographical position.	(*Which province has recently strengthened the regulation of the intellectual property agency industry?; Hainan*)
`time`	Points or intervals of a continuous sequence of events or conditions.	(*In which year was the recent "Haikou Cup" sailing competition held?; 2023*)
`event`	Something that happens, which can be planned or spontaneous.	(*What themed event was recently launched in Suzhou High-speed Railway New Town to promote the development of private enterprises?; "Suzhou Sentiments, Private Enterprises Connected at Heart"*)
`artificial work`	Items or intellectual achievements created by humans, which have artistic, academic, or practical value.	(*What is the latest TV series aired starring Xin Jiang?; As Long As We Are Together*)
`group`	Entities formed by multiple individuals for a specific purpose.	(*Which undergraduate university is rencently established in the Ningxia Hui Autonomous Region recently?; Ningxia Minjiang Institute of Applied Technology*)
`nature`	Phenomena or entities in the natural world.	(*Please explain to me what Nucleases is?; Small RNA molecules with catalytic function, belonging to the category of biological catalysts?; capable of degrading specific mRNA sequences.*)
`quantity`	Numeric value for times or stuff.	(*How many base pairs in human Y chromosome have been observed from the latest sequencing results?; More than 30 million*)
`other`	Other answer not classified to the above categories.	(*Is there any fee for withdrawing WeChat balance to bank card?; Yes*)

Table 7: **Descriptions** and **examples** of *question tags* (first three rows) and *answer types* (last nine rows). We represent (<*question*>; <*answer*>) as examples. Original language for these examples is Chinese. We translate them here for better preview.

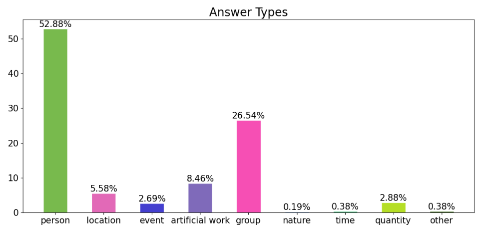


Figure 12: Distributions of answer types for slow-changing questions.

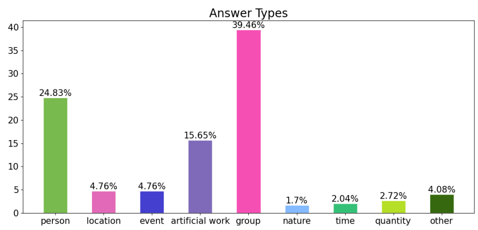
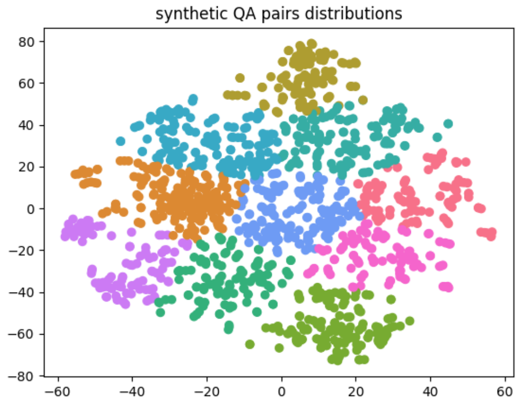
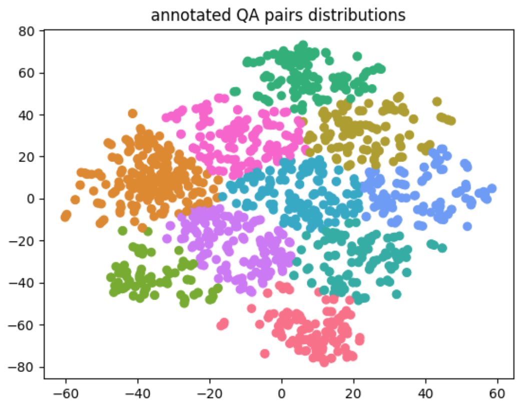


Figure 13: Distributions of answer types for never-changing questions.



(a) Synthetic QA generated by GPT-4



(b) Annotated QA produced by manual annotations

Figure 14: t-SNE analysis for CDQA QA pairs

and annotated QA embeddings, yields a negligible value of approximately 0.54. Such a small divergence suggests that even with human intervention, the essence of the synthetic QA data is largely preserved showing the satisfaction of human labelers in using them as information-seeking questions. This finding strengthens our confidence in the generalizability and real-world applicability of our models derived from the dataset in question.

C Translated Chinese Prompts

The translated prompt framework is illustrated in Figure 15.

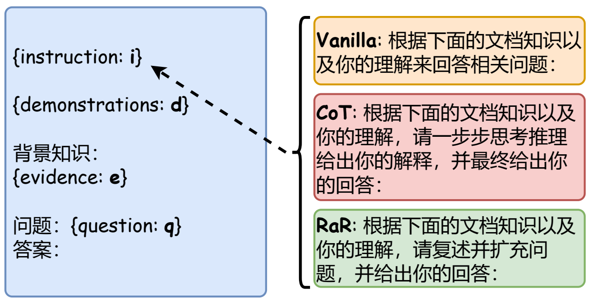


Figure 15: The Chinese prompt framework for Figure 2.