Chinese SimpleQA: A Chinese Factuality Evaluation for Large Language Models

Yancheng He[1]*, Shilong Li[1]*, Jiaheng Liu[1,2]*†, Yingshui Tan[1], Weixun Wang[1], Hui Huang[1], Xingyuan Bu[1], Hangyu Guo[1], Chengwei Hu[1], Boren Zheng[1], Zhuoran Lin[1], Xuepeng Liu[1], Dekai Sun[1], Shirong Lin[1], Zhicheng Zheng[1], Xiaoyong Zhu[1], Wenbo Su[1], Bo Zheng[1]

[1]Alibaba Group, China [2]Nanjing University, China

heyancheng.hyc@taobao.com, liujiaheng@nju.edu.cn

Abstract

New LLM benchmarks are important to align with the rapid development of Large Language Models (LLMs). In this work, we present **Chinese SimpleQA**, the first comprehensive Chinese benchmark to evaluate the factuality ability of LLMs to answer short questions, and Chinese SimpleQA mainly has five properties (i.e., Chinese, Diverse, High-quality, Static, Easy-to-evaluate). Specifically, first, we focus on the **Chinese** language over 6 major topics with 99 **diverse** subtopics. Second, we conduct a comprehensive quality control process to achieve **high-quality** questions and answers, where reference answers are **static** and cannot be changed over time. Third, following SimpleQA, questions and answers are very short, and the grading process is **easy-to-evaluate**. Based on Chinese SimpleQA, we perform a comprehensive evaluation of the factuality abilities of existing LLMs. Finally, we hope that Chinese SimpleQA could guide developers to better understand the factuality abilities of their models and facilitate the growth of LLMs[1].

1 Introduction

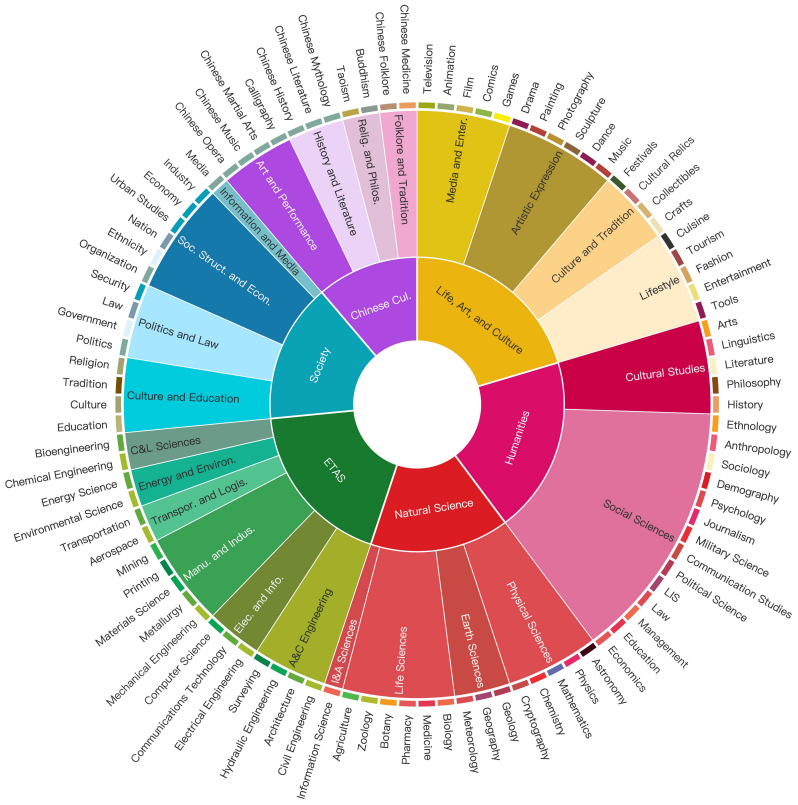
A significant challenge in AI development is to ensure language models generate factually accurate responses (Zhao et al., 2023; Liu et al., 2025; Bu et al., 2021; Li et al., 2024; Bai et al., 2024; Huang et al., 2025; Zhang et al., 2024; Liu et al., 2024). Current frontier models sometimes produce false outputs or answers that are not substantiated by evidence. This is the problem known as "hallucinations", which greatly hinders the extensive use of general AI technologies, such as large language models (LLMs). Besides, it is difficult to evaluate the factuality abilities of the existing LLMs. For example, LLMs usually generate lengthy responses containing numerous factual claims. Recently, to address the aforementioned evaluation problem, OpenAI has released the SimpleQA benchmark (Wei et al., 2024) with 4,326 concise and fact-seeking questions, which makes measuring factuality simple and reliable.

However, the SimpleQA benchmark primarily targets the English language, resulting in a limited understanding of LLMs' capabilities in other languages. Moreover, inspired by several recent Chinese LLM benchmarks (e.g., C-Eval (Huang et al., 2023), CMMLU (Li et al., 2023b)), to evaluate the factuality abilities of LLMs in Chinese, we present the **Chinese SimpleQA** benchmark, which consists of 3000 high-quality questions spanning 6 major topics, ranging from humanities to science and engineering, as shown in the left of Figure 1. Specifically, the distinct main features of our proposed Chinese SimpleQA dataset are as follows:

- **Chinese**: Our Chinese SimpleQA focuses on the Chinese language, which provides a comprehensive evaluation of the factuality abilities of existing LLMs in Chinese.
- **Diverse**: Chinese SimpleQA covers 6 major topics (i.e., "Chinese Culture", "Humanities", "Engineering, Technology, and Applied Sciences", "Life, Art, and Culture", "Society", and "Natural Science"), and these topic includes 99 fine-grained subtopics in total, which demonstrates the diversity of our Chinese SimpleQA.
- **High-quality**: We conduct a comprehensive and rigorous quality control process to ensure the quality and accuracy of Chinese SimpleQA.
- **Static**: Following SimpleQA, to preserve the evergreen property of Chinese SimpleQA, all reference answers would not change over time.
- **Easy-to-evaluate**: Following SimpleQA, as the questions and answers are very short, the grading procedure is fast to run via existing LLMs.

∗ First three authors contributed equally.
† Corresponding Author.
[1]Codes and datasets are available at https://github.com/OpenStellarTeam/ChineseSimpleQA.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19182–19208
July 27 - August 1, 2025



Question Category	Number
Total	3000
- Chinese Culture	326
- Humanities	609
- Engineering, Technology and Applied Sciences	481
- Life, Art and Culture	601
- Society	453
- Natural Science	530
Question Length	
- *maximum length*	106
- *minimum length*	9
- *avg length*	29.4
Ref. Answer Length	
- *maximum length*	33
- *minimum length*	1
- *avg length*	6.0

Figure 1: Left: Overview of Chinese SimpleQA. "Chinese Cul." and "ETAS" represent "Chinese Culture" and "Engineering, Technology, and Applied Sciences", respectively. Right: Dataset statistics of Chinese SimpleQA.

Moreover, we have performed a comprehensive evaluation and analysis on Chinese SimpleQA, and several insightful findings are as follows:

- **Chinese SimpleQA is challenging**. Only o1-preview and Doubao-pro-32k achieve the passing score (63.8% and 61.9% on the correct metric), and there is a long way to go for many existing LLMs.
- **Larger models lead to better results**. Based on the results of the Qwen2.5 series, InternLM series, Yi-1.5 series, etc, we observe that better performance is obtained when LLM is larger.
- **Larger models are more calibrated**. o1-preview is more calibrated than o1-mini, and GPT-4o is more calibrated than GPT-4o-mini.
- **RAG matters**. Performance gaps between different LLMs decrease a lot when using RAG (Retrieval-Augmented Generation). For GPT-4o and Qwen2.5-3B, the performance gap decreases from 42.4% to 9.3% using RAG.
- **Alignment tax exists**. Existing alignment or post-training strategies usually decrease the factuality of language models.
- **Rankings of SimpleQA and Chinese SimpleQA are different**. The performance of several LLMs focusing on Chinese (Doubao-pro-32k, and GLM-4-Plus) is close to the high-performance o1-preview. In particular, in the "Chinese Culture" topic, these Chinese community LLMs are significantly better than GPT or o1 series models.

2 Related Works

LLM Factuality. LLM factuality is the capability of large language models to produce contents that follow factual content, including commonsense, world knowledge, and domain facts, and the factual content can be substantiated by authoritative sources (e.g., Wikipedia, Textbooks). Recent works have explored the potential of LLMs to serve as factual knowledge bases (Yu et al., 2023; Pan et al., 2023; He et al., 2025). Specifically, existing studies have primarily focused on qualitative assessments of LLM factuality (Lin et al., 2022a; Chern et al., 2023), investigations into knowledge storage mechanisms (Meng et al., 2022; Chen et al., 2023), and analyses on knowledge-related issues (Gou et al., 2023).

Factuality Benchmarks. Many factuality benchmarks (Hendrycks et al., 2021; Zhong et al., 2023; Huang et al., 2023; Li et al., 2023b; Srivastava et al., 2023; Yang et al., 2018) have been proposed. For example, MMLU (Hendrycks et al., 2021) is to measure the multitask accuracies on a diverse set of 57 tasks. TruthfulQA (Lin et al., 2022a) focuses on assessing the truthfulness of a language model's generated answers. Additionally, HaluEval (Li et al., 2023c) is to examine the tendency of LLMs to produce hallucinations. Recently, SimpleQA (Wei et al., 2024) has been proposed to measure the short-form factuality in LLMs. However, SimpleQA only focuses on the English domain. In contrast, our Chinese SimpleQA aims to comprehensively evaluate factuality in Chinese.

3 Chinese SimpleQA

3.1 Overview

The left of Figure 1 shows the category distribution of Chinese SimpleQA, which encompasses a comprehensive set of 6 major topics with 99 distinct subtopics. The six main topics are: "Chinese Culture", "Humanities", "Engineering, Technology and Applied Sciences", "Life, Art and Culture", "Society", and "Natural Science", and the 99 subcategories are detailed in Appendix I. Moreover, we specially set up the category of "Chinese Culture" to evaluate the region-specific knowledge. We also present several examples from our dataset in Appendix G. The right of Figure 1 provides detailed statistics for the Chinese SimpleQA dataset. The dataset consists of 3,000 samples, and the distribution across the six major categories is relatively balanced, facilitating a comprehensive assessment of LLMs in diverse domains. In addition, the average answer length is approximately six tokens, ensuring efficient evaluation and minimizing potential errors.

3.2 Dataset Construction

The data construction process consists of the following steps: (1) extracting and filtering relevant knowledge content, (2) manually collecting high-quality question-answer examples for each topic, (3) generating question-answer pairs using predefined criteria, and (4) verifying and revising them if requirements are not met.

Specifically, we first collect a large volume of knowledge-rich text from various fields, such as Wikipedia[2] and Baidu Baike[3]. We then apply specific filtering rules and use a trained quality assessment model to remove low-quality data, such as texts with little information content or too specific. Next, we manually create several high-quality examples for each category to provide a few-shot learning foundation for the model's question generation. These examples are designed to demonstrate various ways of asking questions to ensure diversity in the generated outputs. Finally, we prompt the LLM to generate question-answer pairs using these high-quality knowledge contents based on predefined criteria. We have also established an automated multi-round feedback mechanism: when the model generates a question-answer pair, it triggers another model to evaluate whether the pair meets the criteria. If it does not, the secondary model provides suggestions for improvement, and the main model regenerates the pair based on this feedback.

Notably, the construction of question-answer pairs is based on the following criteria:

Questions must be based on factual knowledge. Questions should pertain to objective facts about the world and must not be influenced by subjective opinions. For example, questions that begin with "What do you think about" or "How would you evaluate" are inappropriate, as they invite personal judgment rather than factual inquiry.

Answers must be unique and unambiguous. Each question must have a single, definitive answer, eliminating any possibility for multiple correct responses. Questions that have contentious or ambiguous answers are not suitable. For instance, the question "Who is the author of Dream of the Red Chamber?" does not meet the requirements because there are different opinions.

Answers must not change over time. Answers should reflect timeless facts unaffected by the time the question is posed. For example, "What is the atomic number of carbon?" and the answer "6" remains unchanged. In contrast, questions regarding current affairs, such as "Who is the current president of a certain country?" are inappropriate, as their answers are subject to change.

Questions must be challenging. Questions should not be overly simple, and queries need to

[2] https://www.wikipedia.org/
[3] https://baike.baidu.com/

Benchmark	Venue	Data Size	Language	Data Source	Categories	Open-ended	Reasoning	Short-form	Metric
TriviaQA (Joshi et al., 2017)	ACL	650K	English	Human Collection	-	✓	✓	✓	Accuracy
NQ (Kwiatkowski et al., 2019)	TACL	3,610	English	Real World	-	✓	✓	✓	Accuracy
MMLU (Hendrycks et al., 2021)	ICLR	15,908	English	Exams	57	×	✓	✓	Accuracy
TruthfulQA (Lin et al., 2022b)	ACL	817	English	Human Writers	38	✓	×	×	ROUGE
C-Eval (Huang et al., 2023)	NeurIPS	13,948	Chinese	Exams	52	×	✓	✓	Accuracy
CMMLU (Li et al., 2023a)	ACL	11,528	Chinese	Exams	67	×	✓	✓	Accuracy
HaluEval (Li et al., 2023d)	EMNLP	5000	English	General	-	×	✓	×	Accuracy
ChineseFactEval (Chern et al., 2023)	Arxiv	125	Chinese	Human Collection	7	✓	×	×	LLM-as-a-Judge
AGI-Eval (Zhong et al., 2024)	NAACL	8062	Ch&En	Exams	20	×	✓	✓	Accuracy
SimpleQA (Wei et al., 2024)	Arxiv	4,326	English	Human Writers	10	✓	×	✓	LLM-as-a-Judge
Chinese SimpleQA (Ours)	-	3,000	Chinese	Self-constructed & Human Writers	99	✓	×	✓	LLM-as-a-Judge

Table 1: Comparisons between our Chinese SimpleQA and other benchmarks.

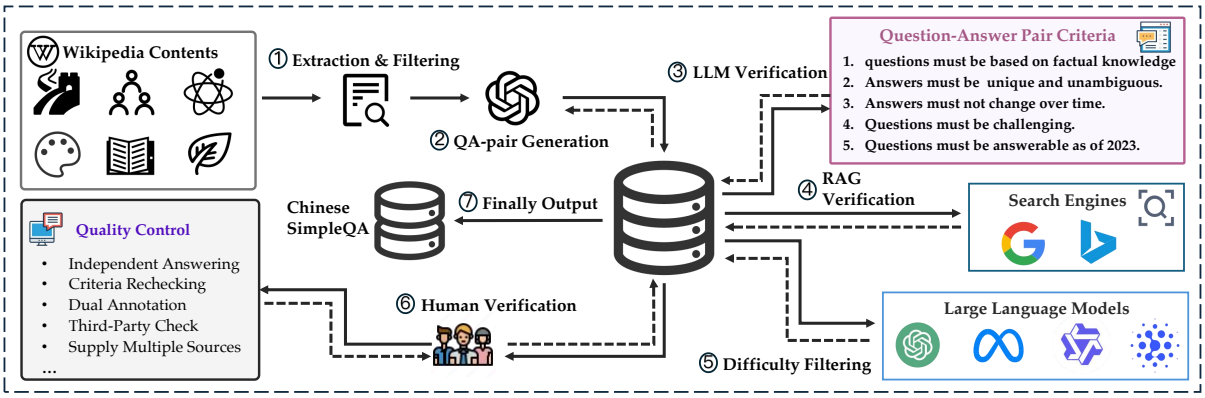


Figure 2: An overview of the entire production process of Chinese SimpleQA.

assess the knowledge depth thoroughly.

Questions must be answerable as of 2023. Each question must be answerable by December 31, 2023, ensuring fair evaluation for models trained on data available post this date.

3.3 Quality Control

Automatic Verification. We also build an automated verification mechanism, including LLM-based verification and RAG-based verification. Specifically, we first use LLM to determine whether the question meets the criteria and is relevant to the current category. If it does not meet the requirements or is irrelevant, the question and answer pair will be deleted directly. Then, we use the built RAG system to verify the answer. Here, we build it based on the LlamaIndex [4] framework and use Google and Bing search results as data sources. If there are contradictory answers in the retrieval results, they will be deleted directly. For more details, please refer to Appendix K.

Difficulty Filtering. In addition, we filter simple questions to discover the knowledge boundaries of LLMs and improve the difficulty of Chinese SimpleQA. Specifically, if a question could be correctly answered by all five powerful models[5], it is considered as a simple question and will be discarded.

Human Verification. Following the automated data collection, human verification is employed to further enhance the quality of the dataset. Specifically, each question is independently evaluated by two human annotators. Initially, the annotators assess whether the question adheres to the predefined criteria outlined earlier. If either annotator deems the question non-compliant, it is discarded. Subsequently, both annotators use search engines to retrieve relevant information and formulate answers. At this stage, the annotators are required to use content from authoritative sources (e.g., Wikipedia, Baidu Baike), and each must provide at least two supporting URLs. In cases where the annotators' answers are inconsistent, a third annotator reviews the sample. The final annotation is determined by the third annotator, who references the initial two assessments. Finally, the human annotation results are compared with responses generated by the large language model (LLM), and only those question-answer pairs that are fully consistent are

[4] https://github.com/run-llama/llama_index

[5] GPT-4o (OpenAI, 2023), Meta-Llama-3-70B-Instruct (Dubey et al., 2024), Qwen2.5-72B-Instruct (Team, 2024d), DeepSeek-7B-chat (DeepSeek-AI, 2024a) and Baichuan2-7B-chat (Baichuan, 2023).

retained. This rigorous human verification process ensures that the dataset is both accurate and meets established standards.

Desensitization. After completing the above verification, we also use our safety risk model to filter to ensure that the final questions and answers do not contain any security risks.

Preventing Data Contamination. We need to mention that our Chinese SimpleQA evaluates the factuality knowledge abilities, where the knowledge is saved in the training corpus from the website (e.g., Wikipedia). Thus, we rigorously check generated questions against our question database for highly similar or identical ones. When such similarities are detected, we rewrite the questions to ensure uniqueness, which prevents direct memorization and ensures that the evaluation reflects the model's ability to recall knowledge.

Analysis of the Retention Rate. In the collection process, many low-quality question-answer pairs are discarded. Specifically, 10,000 pairs are initially generated. After LLM-based verification and RAG-based verification, roughly 2,840 pairs are removed. After that, another 3,690 samples are removed after difficulty evaluation through testing with different models, which means that only about 35% of the original generated data remains. Finally, after a thorough and rigorous manual review, only about 3,000 samples are kept, which is approximately 30% of the original dataset.

3.4 Comparison to other benchmarks

In Table 1, we compare Chinese SimpleQA with several mainstream benchmarks, and our dataset is the first Chinese evaluation set to adopt a generative approach to evaluate the factuality abilities comprehensively. It is worth noting that the C-Eval and CMMLU mainly adopt multiple-choice evaluation methods that may introduce option bias and reduce the difficulty of questions, making it easier for models to guess the correct answer rather than truly understand the question (We do a detailed experimental analysis in Appendix C). In contrast, the generative evaluation method used in Chinese SimpleQA is closer to real-world scenarios. In addition, compared with other similar datasets, our benchmark has the advantages of high evaluation efficiency and more comprehensive coverage.

4 Experiments

4.1 Setup

We use the same prompt format in all experiments. The temperature and sampling parameters are the official configuration or default parameters of each LLM. The judge model we use is GPT-4o. We also evaluate using their smaller models as judge models, all of which achieved a high degree of consistency(see Appendix B for more details). For more details on the experimental implementation, please refer to Appendix H.

4.2 Baseline Models

We evaluate a total of 41 models, comprising 17 closed-source models and 24 open-source models. The closed-source models include: o1-preview [6], Doubao-pro-32k[7], GLM-4-Plus[8], GPT-4o[9], Qwen-Max (Team, 2024c), Gemini-1.5-pro (Team, 2024a), DeepSeek-V2.5 (DeepSeek-AI, 2024b), Claude-3.5-Sonnet [10], Yi-Large[11], moonshot-v1-8k[12], GPT-4-turbo (OpenAI, 2023), GPT-4 (OpenAI, 2023), Baichuan3-turbo[13], o1-mini[14], Doubao-lite-4k[15], GPT-4o-mini[16], GPT-3.5 (Brown et al., 2020). The open-source models include: Qwen2.5 series (Team, 2024d), InternLM2.5 series (Team, 2024b), Yi-1.5 series (AI et al., 2024), LLaMA3 (AI@Meta, 2024) series, DeepSeek Series (DeepSeek-AI, 2024a), Baichuan2 series (Baichuan, 2023), Mistral series (Jiang et al., 2023), ChatGLM3 and GLM-4 (GLM et al., 2024; Du et al., 2022).

4.3 Evaluation Metrics

Following SimpleQA, we also adopt the following five evaluation metrics. (1). **Correct (CO)**: The predicted answer fully includes the reference answer without introducing any contradictory elements. (2) **Not attempted (NA)**: The reference answer is not fully given in the predicted answer,

[6] https://openai.com/index/introducing-openai-o1-preview/
[7] https://www.volcengine.com/product/doubao
[8] https://bigmodel.cn/dev/api/normal-model/glm-4
[9] https://openai.com/index/hello-gpt-4o/
[10] https://www.anthropic.com/news/claude-3-5-sonnet
[11] https://platform.lingyiwanwu.com/
[12] https://platform.moonshot.cn/
[13] https://platform.baichuan-ai.com/
[14] https://openai.com/o1/
[15] https://www.volcengine.com/product/doubao
[16] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Models	Overall results on 5 metrics					F-score on 6 topics					
	CO	NA	IN	CGA	F-score	CC	HU	ETAS	LAC	SO	NS
Closed-Source Large Language Models											
o1-preview	63.8	12.2	24.0	72.7	**67.9**	45.7	**69.8**	**72.4**	**65.0**	**73.5**	**72.3**
Doubao-pro-32k	61.9	10.3	27.8	69.1	65.3	**61.8**	69.3	69.0	56.1	64.2	70.4
GLM-4-Plus	58.7	7.4	33.9	63.4	60.9	56.5	64.1	64.9	50.7	66.6	62.8
GPT-4o	59.3	1.4	39.3	60.1	59.7	39.4	64.0	65.1	53.3	68.6	62.0
Qwen-Max	54.1	11.3	34.6	61.0	57.4	47.8	59.9	63.5	49.9	61.2	59.3
Gemini-1.5-pro	54.4	8.0	37.6	59.1	56.7	41.4	59.1	60.8	52.2	56.3	64.3
DeepSeek-V2.5	54.1	5.9	40.0	57.5	55.7	50.4	57.6	58.8	50.1	59.4	56.9
Claude-3.5-Sonnet	46.2	27.4	26.4	63.6	53.5	28.7	61.3	60.4	42.2	59.8	57.7
Yi-Large	47.3	16.4	36.3	56.6	51.5	41.1	56.5	55.1	41.7	57.6	53.8
moonshot-v1-8k	48.7	5.4	45.9	51.5	50.1	49.8	54.1	56.8	41.4	53.0	46.6
GPT-4-turbo	45.6	14.2	40.2	53.1	49.1	24.2	55.2	58.9	43.9	52.5	50.8
GPT-4	45.4	8.4	46.2	49.6	47.4	25.2	54.0	52.8	41.8	52.8	50.6
Baichuan3-turbo	45.2	9.0	45.8	49.6	47.3	32.3	52.5	54.0	35.4	54.6	50.9
o1-mini	39.5	20.6	39.9	49.7	44.1	21.3	49.2	55.9	33.8	48.8	46.8
Doubao-lite-4k	36.7	31.2	32.1	53.3	43.4	40.2	44.8	51.0	31.1	41.4	50.4
GPT-4o mini	37.6	0.9	61.5	37.9	37.8	19.0	42.4	46.4	31.0	42.2	39.8
GPT-3.5	29.7	2.9	67.4	30.6	30.1	13.3	35.8	35.2	25.6	32.7	31.7
Open-Source Large Language Models											
QwQ-32B-Preview	39.8	11.7	48.5	45.1	42.3	35.9	46.0	44.7	27.3	43.5	41.7
Qwen2.5-72B	48.4	7.1	44.5	52.1	50.2	36.3	56.1	57.9	37.1	53.3	56.4
Qwen2.5-32B	38.8	11.1	50.1	43.6	41.1	33.7	45.8	48.7	27.3	44.7	44.9
Qwen2.5-14B	35.4	9.6	55.0	39.2	37.2	30.2	41.8	46.1	24.1	38.8	41.0
Qwen2.5-7B	26.6	9.9	63.5	29.5	27.9	20.1	32.7	33.8	18.0	28.6	32.0
Qwen2.5-3B	16.2	12.8	71.0	18.6	17.3	13.4	17.9	26.1	9.3	15.6	20.8
Qwen2.5-1.5B	11.1	14.6	74.3	13.1	12.0	11.0	11.3	18.7	6.7	12.2	12.9
GLM4-9B	25.9	12.5	61.6	29.6	27.6	28.8	32.1	32.0	17.6	28.9	27.8
ChatGLM3-6B	11.2	13.6	75.2	12.9	12.0	12.1	13.8	12.4	8.8	13.4	11.8
InternLM2.5-20B	31.5	7.7	60.8	34.1	32.8	32.0	37.1	37.7	21.2	35.7	34.3
InternLM2.5-7B	24.7	7.5	67.8	26.7	25.7	25.5	29.4	31.0	16.4	26.9	25.8
InternLM2.5-1.8B	5.3	31.1	63.6	7.6	6.2	6.1	8.7	7.2	3.3	4.5	7.4
Yi-1.5-34B	30.9	5.8	63.3	32.8	31.8	28.2	36.9	36.8	24.4	32.8	31.4
Yi-1.5-9B	18.2	2.9	78.9	18.7	18.4	17.2	20.2	24.3	10.2	20.1	19.8
Yi-1.5-6B	15.9	2.8	81.3	16.3	16.1	14.2	17.9	21.3	10.3	16.8	16.5
LLaMA3.1-70B	38.3	9.4	52.3	42.3	40.2	22.9	47.2	49.3	34.5	49.6	40.4
LLaMA3.1-8B	16.9	8.8	74.3	18.6	17.7	8.5	20.7	23.4	9.7	20.5	20.7
DeepSeek-67B	43.5	14.8	41.7	51.1	47.0	34.3	54.5	50.3	42.3	49.0	46.2
DeepSeek-V2-Lite-Chat	33.7	12.8	53.5	38.6	36.0	35.3	38.5	41.7	32.2	37.5	31.2
DeepSeek-7B	23.2	13.2	63.6	26.7	24.8	24.5	27.2	28.9	20.6	27.0	21.5
Baichuan2-13B	19.1	24.9	56.0	25.4	21.8	24.0	25.8	23.3	16.8	23.0	18.7
Baichuan2-7B	12.5	21.8	65.7	16.0	14.0	14.6	16.1	15.4	11.1	13.8	13.3
Mixtral-8x22B-Instruct-v0.1	27.3	2.2	70.5	27.9	27.6	10.6	32.3	36.0	21.0	34.1	26.9
Mixtral-8x7B-Instruct-v0.1	20.4	7.2	72.4	22.0	21.2	5.2	26.5	29.0	13.0	25.0	23.3
Mistral-7B-Instruct-v0.2	15.0	8.8	76.2	16.4	15.6	4.5	18.2	22.2	9.5	21.4	15.7

Table 2: Results of different models on Chinese SimpleQA. For metrics, **CO**, **NA**, **IN**, and **CGA** denote "Correct", "Not attempted", "Incorrect", and "Correct given attempted", respectively. For subtopics, **CC**, **HU**, **ETAS**, **LAC**, **SO** and **NS** represent "Chinese Culture", "Humanities", "Engineering, Technology, and Applied Sciences", "Life, Art, and Culture", "Society", and "Natural Science", respectively. **Bold** indicates the best results within the same group of models, while underline indicates the second best.

and there are no contradictory elements with the reference answer. (3) **Incorrect (IN)**: The predicted answer contradicts the reference answer, even if the contradiction is solved. (4). **Correct given attempted (CGA)**: The metric is the proportion of accurately answered questions among those attempted questions. (5). **F-score**: The metric represents the harmonic mean between correct and

correct given attempted.

4.4 Main Results

In Table 2, we provide the results of different LLMs on Chinese SimpleQA. Specifically, we provide the overall results on 5 evaluation metrics. Additionally, we report the F-score for 6 topics to analyze fine-grained factuality abilities. We have the following insightful observations:

- o1-preview achieves the best performance on Chinese SimpleQA, and results of several recent closed-source LLMs focusing on Chinese (Doubao-pro-32k and GLM-4-Plus) are very close to o1-preview.

- It is obvious that the "mini" series models (o1-mini, GPT-4o-mini) achieve lower results than the corresponding larger models (o1-preview, GPT-4o), which also indicates these "mini" series models do not pay attention to memorize factuality knowledge.

- A Larger LLM leads to better performance, where we can draw this conclusion based on many model series (e.g., GPT, Qwen2.5, InternLM2.5, Yi-1.5). The scatter plot in Figure 4 shows a clearer positive correlation between model scale and performance.

- Small LLMs usually lead to higher scores on "not attempted (NA)". The NA scores for o1-mini, InternLM2.5-1.8B are 20.5 and 31.2, respectively, which are larger than the scores of corresponding larger LLMs a lot (o1-preview with 12.2, InternLM2.5-20B with 7.7).

- There is a significant performance difference among different subtopics for different LLMs. Notably, the Chinese community LLMs (e.g., Doubao-pro-32k, GLM-4-Plus, Qwen-Max) are significantly better than GPT or o1 models in the Chinese Culture (CC) subtopic. In contrast, o1 has significant advantages in science-related subtopics (e.g., ETAS and NS).

In addition, we also provide the detailed results (CO and CGA metrics) on 6 topics in Figure 3.

4.5 Further Analysis

Analysis of Calibration. For the calibration of different LLMs, following SimpleQA, we instruct the model to provide a corresponding confidence level (from 0 to 100) when answering questions to measure the model's confidence in its answers (See the prompt in Appendix L). We know that a perfectly calibrated model's confidence (%) should match its answers' actual accuracy. The left plot in Figure 5 illustrates the alignment performance, which indicates that GPT-4o aligns better than GPT-4o-mini and o1-preview aligns better than o1-mini. For the Qwen2.5 series, the alignment order is Qwen2.5-72B > Qwen2.5-32B > Qwen2.5-7B > Qwen2.5-3B, which suggests that larger model sizes result in better calibration. Furthermore, for all evaluated models, their confidence in the range of confidence > 50 falls below the line of perfect alignment, which means that they all overestimate the accuracy of their responses and overconfidence exists.

Analysis of Test-Time Compute. We also evaluate the relationship between increased test-time compute and response accuracy for different LLMs. Specifically, we randomly sample 50 samples from Chinese SimpleQA, and for each sample, the model is asked to independently answer 100 times. Then, we obtain the model's response accuracy using the Best-of-N method as the inference counts increase. The results are shown in the right plot of Figure 5. We observe that as the times of inferences increase, the response accuracy of all models improves and eventually reaches a ceiling. This is reasonable for Chinese SimpleQA to probe the boundaries of a model's knowledge.

Analysis on the effect of RAG. We explore the effect of the Retrieval-Augmented Generation (RAG) in enhancing the factual accuracy of LLMs on Chinese SimpleQA. Specifically, we reproduce a RAG system based on LlamaIndex (Liu, 2022), incorporating Google search APIs(For an ablation study on different retrieval APIs, see Appendix F). In Figure 6, all models show a substantial improvement in accuracy with RAG. For example, the performance of Qwen2.5-3B improved more than threefold. Notably, nearly all models with RAG outperform native GPT-4o. Meanwhile, RAG also leads to a marked reduction in performance disparities among models. For example, the F-score difference between Qwen2.5-3B with RAG and Qwen2.5-72B with RAG is only 6.9%. This suggests that RAG reduces the performance gaps greatly, enabling even smaller ones to achieve high performance when augmented with RAG. Overall, this suggests that RAG serves as an effective shortcut for enhancing the factuality.

Analysis on the alignment tax. Recently, prior

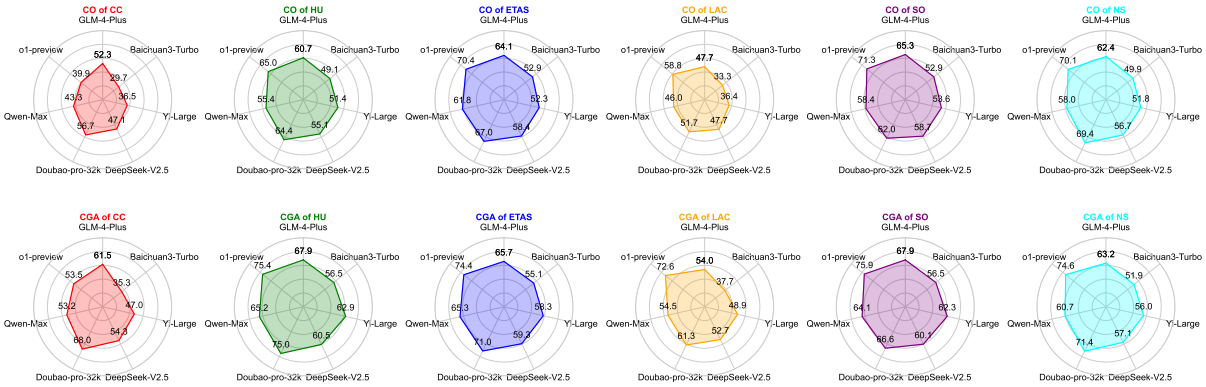


Figure 3: Results (**CO** and **CGA** metrics) of different models for six topics.

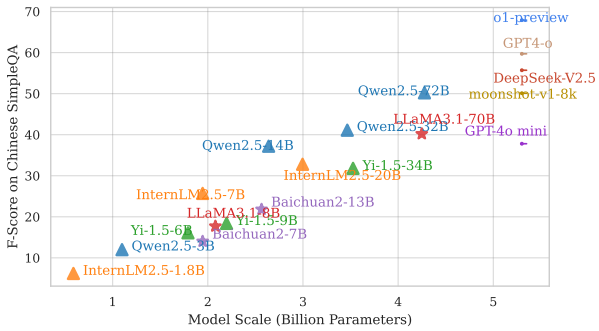


Figure 4: Relationship between model scale (in billion parameters) and F-score on Chinese SimpleQA.

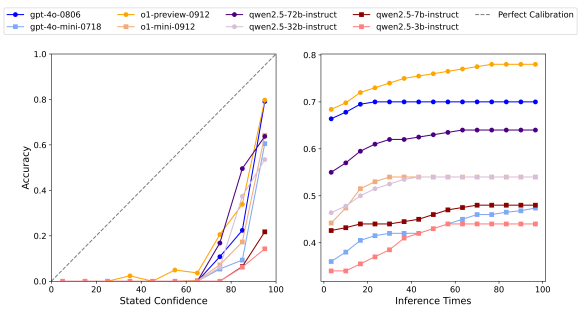


Figure 5: Left: Calibration of LLMs based on their stated confidence. Right: Improvement in accuracy with increased test-time compute using Best-of-N.

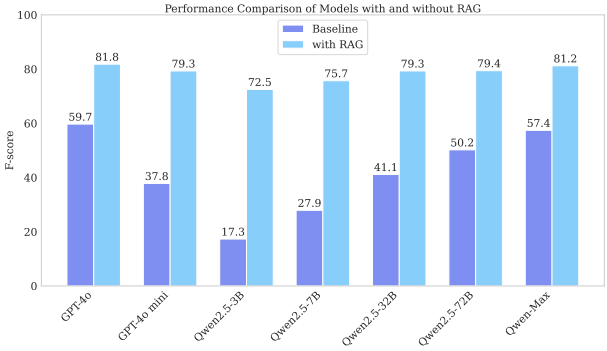


Figure 6: The effect of RAG strategy.

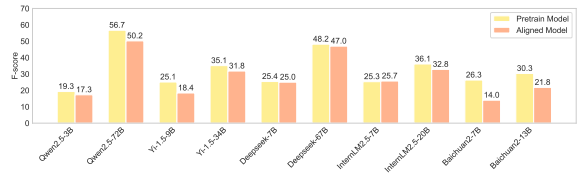


Figure 7: The effect of alignment in post-training.

studies (OpenAI, 2023; Song et al., 2023) have found that the alignment can lead to a decrease in the abilities of language models as known as the "alignment tax". To illustrate the effect of alignment on factuality, we conduct a comparative performance analysis between pre-trained models and aligned models that are trained with Supervised Fine-Tuning (SFT) or Reinforcement Learning from Human Feedback (RLHF). As illustrated in Figure 7, different models exhibit varying trends after post-training, but most models have a significant decline. Among these, the Baichuan2 series models show the most significant decreases, with Baichuan2-7B and Baichuan2-13B experiencing F-score reductions of 47% and 28%, respectively. This reflects that the alignment training of most current LLMs still has obvious drawbacks of knowledge hallucinations, which further reflects the necessity of our Chinese SimpleQA dataset.

Comparison between Chinese SimpleQA and SimpleQA. We also compare the ranking differences of various models on the SimpleQA and the Chinese SimpleQA. In Figure 8, there are notable discrepancies in model performance across these two benchmarks. For instance, Doubao-pro-32k ranks significantly higher on the Chinese SimpleQA, moving from 12th to 2nd place (+10). Conversely, GPT-4 shows a decline in performance on the Chinese SimpleQA, dropping from 3rd to 9th

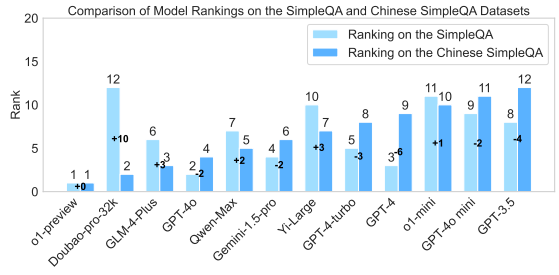


Figure 8: The rankings of different LLMs.

place (-6). These differences emphasize the importance of evaluating models on datasets in various languages and the need for research into optimizing model performance across different linguistic environments. In addition, most Chinese community-developed models (e.g., Qwen-Max, GLM-4-Plus, Yi-Large, Doubao-pro-32k) perform better on the Chinese SimpleQA than on the SimpleQA, showing their advantages on Chinese.

5 Conclusion

In this paper, we propose the first Chinese short-form factuality benchmark (i.e., Chinese SimpleQA), which mainly has five important features (i.e., Chinese, diverse, high-quality, static, and easy-to-evaluate). Besides, we comprehensively evaluate the performance of existing 40+ LLMs on factuality and provide detailed analysis to demonstrate the advantages of our Chinese SimpleQA.

6 Limitations

While Chinese SimpleQA provides valuable insights, it has some limitations. Its coverage of six main topics and 99 subtopics may not fully capture the diversity of all domains, particularly niche or emerging areas. As a static benchmark, it cannot reflect real-time advancements or evolving factual information. Additionally, its focus on short-form questions and reliance on simple evaluation metrics might overlook complex reasoning tasks or nuanced factuality. Addressing these limitations will be our focus in future work.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

AI@Meta. 2024. Llama 3 model card.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xingyuan Bu, Junran Peng, Junjie Yan, Tieniu Tan, and Zhaoxiang Zhang. 2021. Gaia: A transfer learning system of object detection that fits your needs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 274–283.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. *Preprint*, arXiv:2308.13198.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. *Preprint*, arXiv:2307.13528.

DeepSeek-AI. 2024a. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

DeepSeek-AI. 2024b. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

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)*, pages 320–335.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie

Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao,

Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *arXiv preprint arXiv: 2407.21783*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *Preprint*, arXiv:2305.11738.

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, et al. 2025. Can large language models detect errors in long chain-of-thought reasoning? *arXiv preprint arXiv:2502.19361*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hui Huang, Jiaheng Liu, Yancheng He, Shilong Li, Bing Xu, Conghui Zhu, Muyun Yang, and Tiejun Zhao. 2025. Musc: Improving complex instruction following with multi-granularity self-contrastive training. *arXiv preprint arXiv:2502.11541*.

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. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Tim Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *ArXiv*, abs/2306.09212.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023b. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023c. Halueval: A large-scale hallucination evaluation benchmark for large language models. *Preprint*, arXiv:2305.11747.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023d. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, et al. 2024. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. *arXiv preprint arXiv:2406.14550*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Jerry Liu. 2022. LlamaIndex.

Jiaheng Liu, Ken Deng, Congnan Liu, Jian Yang, Shukai Liu, He Zhu, Peng Zhao, Linzheng Chai, Yanan Wu, Ke Jin, et al. 2024. M2rc-eval: Massively multilingual repository-level code completion evaluation. *arXiv preprint arXiv:2410.21157*.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. 2025. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2023. Gpt-4 technical report. *PREPRINT*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying large language models and knowledge graphs: A roadmap. *Preprint*, arXiv:2306.08302.

Ziang Song, Tianle Cai, Jason D Lee, and Weijie J Su. 2023. Reward collapse in aligning large language models. *arXiv preprint arXiv:2305.17608*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

InternLM2 Team. 2024b. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Qwen Team. 2024c. Introducing qwen1.5.

Qwen Team. 2024d. Qwen2.5: A party of foundation models.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models.

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. *Preprint*, arXiv:1809.09600.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *International Conference for Learning Representation (ICLR)*.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, Emmanouil Benetos, Huan Yang, Junting Zhou, Kaijing Ma, Minghao Liu, Morry Niu, Noah Wang, Quehry Que, Ruibo Liu, Sine Liu, Shawn Guo, Soren Gao, Wangchunshu Zhou, Xinyue Zhang, Yizhi Zhou, Yubo Wang, Yuelin Bai, Yuhan Zhang, Yuxiang Zhang, Zenith Wang, Zhenzhu Yang, Zijian Zhao, Jiajun Zhang, Wanli Ouyang, Wenhao Huang, and Wenhu Chen. 2024. Map-neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv: 2405.19327*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. abs/2303.18223.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

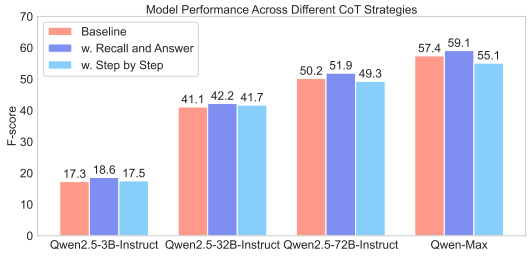


Figure 9: Model Performance Across Different CoT Strategies.

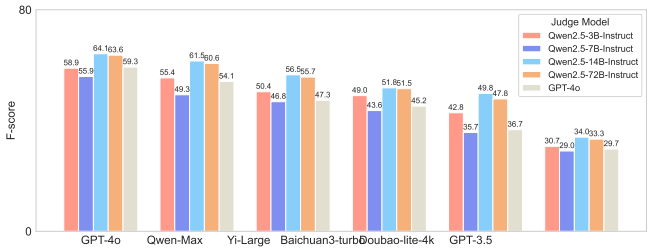


Figure 10: Robustness of Judge Models.

Model	Single-Choice Accuracy (%)	Shuffled-Options Accuracy (%)	Open-Ended QA Accuracy (%)	Accuracy Drop (%)
Qwen-Max	84.89	83.99	66.77	21.35
Yi-Large	71.60	68.88	56.80	20.68
Baichuan3-turbo	74.62	73.23	59.82	19.84
GPT-4o mini	60.12	60.42	42.30	29.65
GPT-3.5	48.20	46.83	30.21	37.32
Qwen2.5-14B	73.41	75.53	63.14	13.99
Qwen2.5-7B	75.23	75.83	54.38	27.71
Baichuan2-13B-Chat	48.64	47.73	25.08	48.45
ChatGLM3-6B	47.73	45.32	19.34	59.49

Table 3: Evaluation results of various models using three methods: original multiple-choice questions, shuffled-options multiple-choice questions, and open-ended QA. The table illustrates accuracy differences across evaluation formats and highlights the impact of option bias and the challenges posed by open-ended QA.

A Analysis on the effect of Chain-of-Thought

To evaluate the impact of Chain-of-Thought (CoT) prompting strategies on model factuality, we implemented and compared two approaches: "Recall and Answer" and "Step by Step". As illustrated in Figure 9, the performance differences induced by these strategies are not substantial across models of varying scales. Notably, larger models, such as Qwen2.5-72B-Instruct and Qwen-Max, display a marginal improvement with CoT strategies, but the gains remain relatively modest compared to baseline performances. These findings align with conclusions drawn in similar studies, such as those in C-EVAL and CMMLU, where CoT prompting did not universally enhance accuracy, and in some cases, resulted in performance degradation. This suggests that the effectiveness of CoT strategies in improving factuality may be limited for tasks that do not require intensive reasoning.

B Robustness of Judge Models

The figure 10 shows the evaluation results of six models selected from different performance levels of the current ranking. We use various evaluation models to evaluate these models. The evaluation models include Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-72B-Instruct, and GPT-4o used in the previous experiment. It can be seen that although the scores of different evaluation models are different, the relative scores and rankings of each model have not changed. This consistency demonstrates the robustness of our evaluation set and evaluation method, and means that using smaller-scale evaluation models (such as Qwen2.5-3B-Instruct or Qwen2.5-7B-Instruct) can also have a high consistency rate, and can be evaluated efficiently even with limited resources.

C Biases in Choice Evaluation

The experimental results presented in Table 3 demonstrate the impact of option bias in multiple-choice evaluations and highlight the importance of open-ended question-answering (QA) evaluations. To conduct the experiment, we used GPT-4 to transform the CEval dataset into a question-answering format, applying the Chinese SimpleQA pipeline to filter the data. This process resulted in a curated set of 300 high-quality QA items for experimentation.

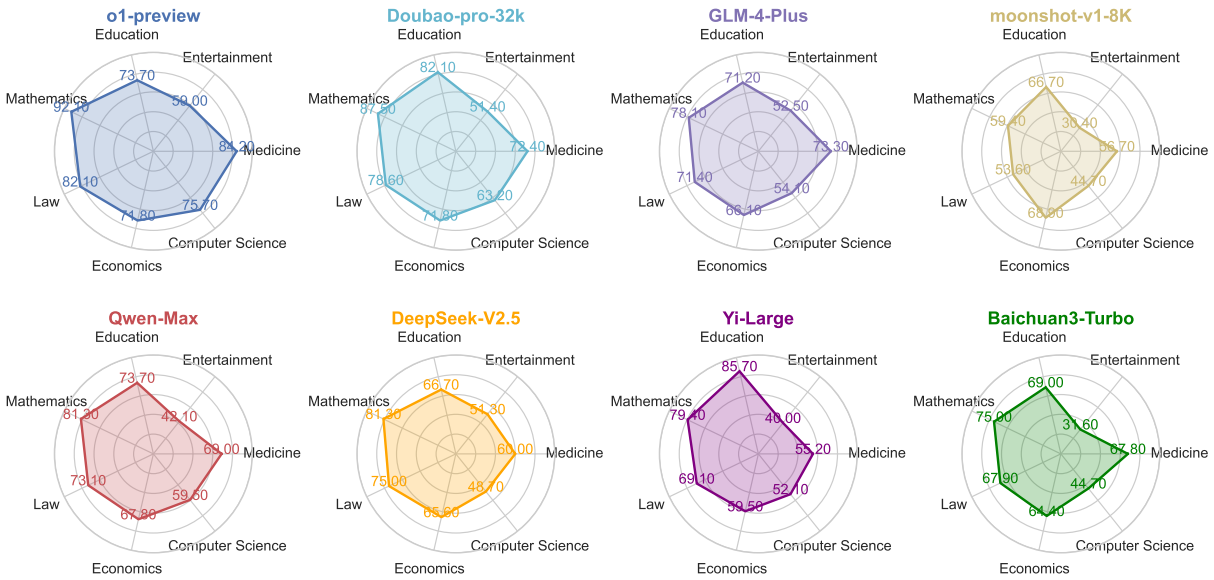


Figure 12: Detailed results on some selected subtopics.

Each model was evaluated using three methods: (1) the original multiple-choice question format, (2) the multiple-choice format with shuffled options, and (3) the open-ended QA format. Results show that most models exhibit a significant reduction in accuracy when evaluated with the open-ended QA format (note: the questions remained consistent across formats). For instance, the Qwen-Max model's accuracy dropped by 21%, while GPT-3.5's accuracy decreased by 30%. These findings suggest that the multiple-choice format simplifies the evaluation process, potentially underestimating the difficulty of the task. Additionally, some models exhibited changes in accuracy after the options were shuffled, indicating the presence of option bias. These observations raise concerns that the multiple-choice evaluation method may fail to fully capture the factual correctness of the models or adequately address hallucination tendencies. In contrast, the generation-based evaluation aligns more closely with realistic task scenarios. This format requires models to generate free-form answers, eliminating the possibility of "guessing" from predefined options and providing a clearer assessment of the model's reasoning and comprehension abilities.

We hope that Chinese SimpleQA can provide the research community with a more accurate perspective to advance model development, rather than just following the trend of multiple-choice based evaluation.

D Analysis Across Answer Types

In Figure 11, we show the accuracy of each model on different answer types. It can be observed that the accuracy of responses to questions involving dates and numerical information is significantly lower compared to other answer types such as organizations, places, and persons. This suggests that the models are more prone to generating hallucinations when processing numerical data, highlighting a persistent challenge in addressing arithmetic reasoning and date-related tasks within these systems.

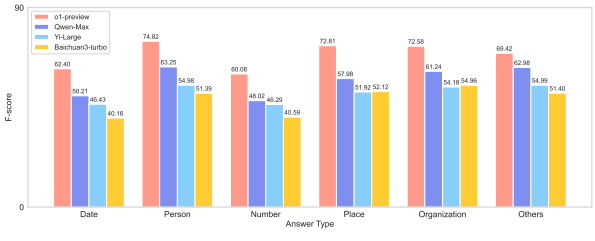


Figure 11: Model Performance on Different Answer Types.

E Analysis on the results of subtopics

As mentioned in Section 3.2, the benchmark covers a total of 99 subtopics, which can comprehensively detect the knowledge level of the model in various domains. Figure 12 illustrates the performance

Model	CO	NA	IN	CGA	F-score
GPT-4o	59.3	1.3	39.3	60.1	59.7
w. RAG (Google's API)	78.4	4.2	13.2	85.6	81.8
w. RAG (Baidu's API)	79.2	5.7	15.1	84.0	81.5
Qwen2.5-72B	48.4	1.8	44.5	52.1	50.2
w. RAG (Google's API)	74.9	4.2	13.9	84.3	79.4
w. RAG (Baidu's API)	77.5	6.7	15.8	81.5	79.2
Qwen2.5-32B	38.8	5.6	50.1	43.6	41.1
w. RAG (Google's API)	74.1	7.2	12.8	85.2	79.3
w. RAG (Baidu's API)	72.9	10.7	15.8	82.1	77.2
Qwen2.5-7B	26.6	4.6	63.5	29.5	27.9
w. RAG (Google's API)	71.1	6.0	16.9	80.8	75.7
w. RAG (Baidu's API)	68.3	11.2	19.5	77.8	72.7
Qwen2.5-3B	16.2	7.6	71.0	18.6	17.3
w. RAG (Google's API)	68.8	4.4	20.9	76.7	72.5
w. RAG (Baidu's API)	65.4	7.7	26.4	71.2	68.2
GPT-4o mini	37.6	0.9	61.5	37.9	37.8
w. RAG (Google's API)	76.8	2.0	17.0	81.9	79.3
w. RAG (Baidu's API)	75.0	2.9	22.1	77.3	76.1

Table 4: Performance Comparison of Models with and without RAG using Different APIs.

comparison between the o1 model and seven notable Chinese community models within several common domains. Firstly, from an overall perspective, the o1-preview model exhibits the most comprehensive performance across these domains, with the Doubao model following closely. In contrast, the Moonshot model demonstrates the weakest overall performance. Secondly, when examining specific domains, a significant disparity emerges between the Chinese community models and the o1 model in areas such as Computer Science and Medicine. However, this gap is minimal in domains like Education and Economics. Notably, in Education, some Chinese community models outperform the o1-preview, highlighting their potential for achieving success in specific vertical domains. Lastly, when examining specific models, the Moonshot model is notably weaker in Mathematics, Law, and Entertainment, while the Baichuan model also underperforms in Entertainment. The Yi-Large model excels in Education, and the o1 model maintains the strongest performance across other domains. Evaluating the performance of the models across diverse domains within the benchmark dataset enables users to identify the most suitable model for their specific needs.

F Performance Comparison of Different Retrieval APIs

we conduct additional experiments using Baidu as the search engine within the RAG framework. The experimental results are as follows in Table 4.

Our findings indicate that the F-score performance of most models using Baidu's API is lower compared to those using Google's API. Further analysis reveals that the retrieval contexts from Baidu's API exhibit relatively higher noises, suggesting the need for more effective filtering strategies. Additionally, we observed that when the model is larger, the performance degradation is smaller. This suggests that larger models are more robust in handling noisy or inconsistent input data.

G Chinese SimpleQA Examples

We present several examples of our dataset in Figure 13.

H Evaluation Details

For evaluating the chat models, we use the official chat template. For the pre-trained models, we employ the few-shot learning approach, where a small set of high-quality question-answer examples is provided to guide the model in understanding the task and generating accurate, concise responses.

For evaluating models with RAG, the model is given relevant reference materials related to the user's question. The system prompt instructs the model to prioritize information from the retrieved information, or, if no relevant information is found, to rely on its own knowledge to generate the answer. The prompt is illustrated in Figure 14.

I Distribution of Subtopics

Figure 15 shows the distribution of the 99 categories in the SimpleQA dataset, including the count and the average question token length.

Figure 16 illustrates the difficulty distribution for questions across 99 distinct categories, measured by the number of incorrect responses from models. Firstly, many categories exhibit narrow and symmetrical violin shapes, indicating a relatively uniform distribution of question difficulty. This suggests that within these categories, questions are consistently challenging, providing a balanced assessment environment for model evaluation. Such consistency is crucial for ensuring fairness in comparisons across different models, as it minimizes biases introduced by disproportionately easy or overly difficult questions. Secondly, the overall layout shows the diversity of the dataset, while most categories maintain a reasonable difficulty distribution. This balance enables the identification of strengths and weaknesses in model performance without being affected by extreme outliers or irregular distributions.

J Models Performance Across Difference Topics

primary_category	secondary_category	question	reference answer
中华文化 (Chinese Culture)	中医(Traditional Chinese Medicine)	创立了调气活血的“衡法”治则的是哪一位中医学家?	颜德
		Which Chinese medicine scientist created the "Heng method" treatment principle for regulating Qi and activating blood circulation?	Yan De
	民俗(Folklore)	西迁节主要是哪个少数民族的节日?	锡伯族
		Which ethnic minority primarily celebrates the Xiqian Festival?	Xibe ethnic group
人文与社会科学 (Humanities)	政治(Political Science)	2021年国际宗教自由峰是在美国哪个城市举行?	华盛顿
		In which U.S. city was the 2021 International Religious Freedom Summit held?	Washington, D.C.
	教育学(Education)	先行组织者概念是哪位美国教育心理学家提出的?	大卫·奥苏伯尔
		Which American educational psychologist proposed the concept of advance organizers?	David Ausubel
自然与自然科学 (Natural Science)	资讯科学 (Information Science)	《中国图书馆分类法》第五版中规定U6表示哪一类?	水路运输
		What category does U6 represent in the 5th edition of the Chinese Library Classification?	Water Transport
	数学(Mathematics)	画法几何这门学科主要是由哪位法国数学家提出的?	加斯帕尔·蒙日
		Which French mathematician is credited with the development of descriptive geometry?	Gaspard Monge
生活、艺术与文化(Life, Art, and Culture)	动画(Animation)	《喜羊羊与灰太狼》是哪家公司制作的原创动画作品?	广东原创动力文化传播有限公司
		Which company produced the original animated series "Pleasant Goat and Big Big Wolf"?	Guangdong Original Power Culture Communication Co., Ltd.
	雕塑(Sculpture)	著名的华尔街铜牛雕塑是由哪位意大利雕塑家创作的?	阿图罗·迪·莫迪卡 (Arturo Di Modica)
		Who created the famous Wall Street Bull sculpture?	Arturo Di Modica
社会(Society)	传统(Tradition)	根据《礼记·内则》，如果女子未许嫁，她应在多少岁进行笄礼?	20
		According to the "Liji: Neize," at what age should an unmarried woman undergo the Ji ceremony?	20
	城市(City)	特拉斯卡拉城是哪个国家的城市?	墨西哥
		In which country is the city of Tlaxcala located?	Mexico
工程、技术与应用科学 (Engineering, Technology, and Applied Sciences)	计算机科学 (Computer Science)	1.7.0版本的pytorch可以兼容的最新cuda版本是多少?	11
		What is the latest compatible CUDA version for PyTorch 1.7.0?	11
	生物工程 (Bioengineering)	哪位中国科学家首先研制成功了转基因鱼?	朱作言
		Which Chinese scientist was the first to successfully develop genetically modified fish?	Zhu Zuoyan

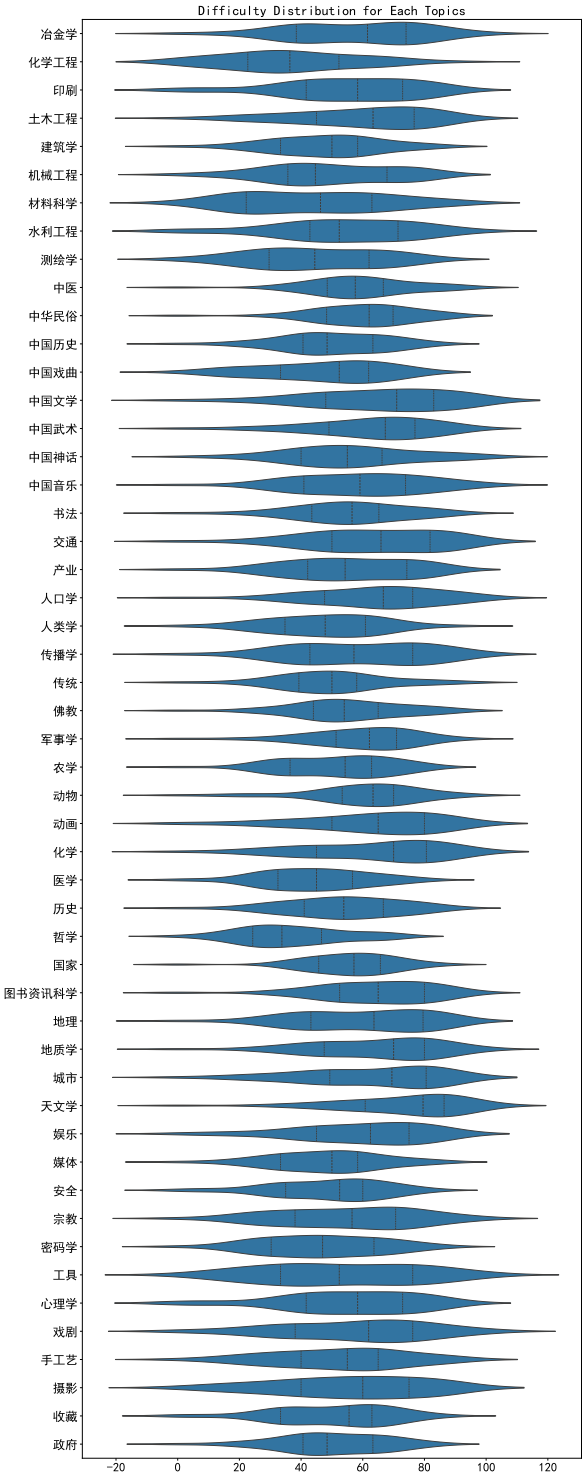
Figure 13: Some examples of Chinese SimpleQA.

请结合检索材料准确地回答用户的问题，如果检索材料有相关知识，优先参考检索材料，若无则依靠自身知识作答。

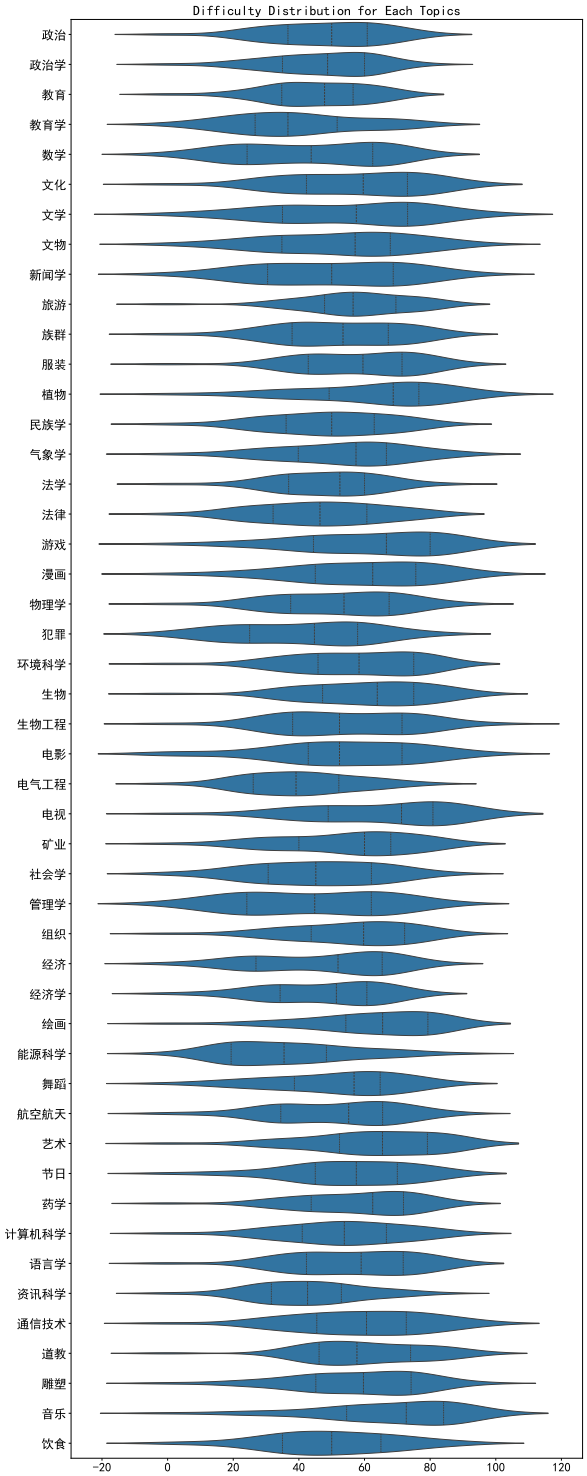
Figure 14: The prompt of RAG.

primary_category	secondary_category	count	question avg. token	primary_category	secondary_category	count	question avg. token
Chinese Culture	Traditional Chinese Medicine (中医)	33	32.1	Society	Industry (产业)	35	28.2
	Chinese Folklore (中华民俗)	30	36.2		Tradition (传统)	26	32.6
	Chinese History (中国历史)	48	30.9		Country (国家)	35	25.8
	Chinese Opera (中国戏曲)	21	30.9		City (城市)	36	26.4
	Chinese Literature (中国文学)	51	29.0		Media (媒体)	24	24.0
	Chinese Martial Arts (中国武术)	26	29.3		Safety (安全)	20	32.3
	Chinese Mythology (中国神话)	20	38.6		Religion (宗教)	27	26.6
	Chinese Music (中国音乐)	22	36.5		Government (政府)	32	31.2
	Calligraphy (书法)	23	31.0		Politics (政治)	31	35.4
	Buddhism (佛教)	26	27.9		Education (教育)	23	28.5
	Taoism (道教)	26	33.7		Culture (文化)	25	27.5
	Demography (人口学)	21	32.5		Ethnic Groups (族群)	29	29.3
	Anthropology (人类学)	20	33.9		Law (法律)	28	31.6
	Communication Studies (传播学)	21	28.3		Crime (犯罪)	19	37.5
	Military Studies (军事学)	37	36.6		Organization (组织)	37	27.2
	History (历史)	50	37.1		Economy (经济)	26	27.3
	Philosophy (哲学)	36	34.4		Agriculture (农学)	35	25.0
	Library and Information Science (图书资讯科学)	29	24.2	Natural Science	Animals (动物)	30	24.4
	Psychology (心理学)	17	32.0		Chemistry (化学)	40	24.3
	Political Science (政治学)	40	29.6		Medicine (医学)	30	32.9
Humanities	Education (教育学)	30	31.3		Geography (地理)	44	26.6
	Literature (文学)	40	32.1		Geology (地质学)	40	25.5
	Journalism (新闻学)	31	25.6		Astronomy (天文学)	44	22.6
	Ethnology (民族学)	27	30.3		Cryptography (密码学)	33	25.9
	Law (法学)	39	32.4		Mathematics (数学)	32	29.1
	Sociology (社会学)	31	28.3		Botany (植物)	32	21.7
	Management (管理学)	23	27.1		Meteorology (气象学)	27	29.1
	Economics (经济学)	35	30.4		Physics (物理学)	41	29.6
	Arts (艺术)	42	32.3		Biology (生物)	36	28.3
	Linguistics (语言学)	40	25.6		Pharmacy (药学)	32	24.7
	Transportation (交通)	22	28.1		Information Science (资讯科学)	34	28.7
	Metallurgy (冶金学)	26	24.8	Life, Art, and Culture	Animation (动画)	40	29.4
	Chemical Engineering (化学工程)	20	31.4		Entertainment (娱乐)	41	30.1
	Printing (印刷)	24	23.6		Tools (工具)	22	28.5
	Civil Engineering (土木工程)	30	28.2		Drama (戏剧)	21	28.5
	Architecture (建筑学)	24	30.8		Handicrafts (手工艺)	20	31.3
	Mechanical Engineering (机械工程)	27	28.4		Photography (摄影)	20	34.0
	Materials Science (材料科学)	26	26.4		Collectibles (收藏)	27	25.3
Engineering, Technology, and Applied Sciences	Hydraulic Engineering (水利工程)	21	30.8		Cultural Relics (文物)	28	30.3
	Surveying (测绘学)	27	24.6		Tourism (旅游)	22	32.1
	Environmental Science (环境科学)	24	32.4		Fashion (服装)	18	25.6
	Bioengineering (生物工程)	21	40.8		Gaming (游戏)	45	30.2
	Electrical Engineering (电气工程)	23	29.9		Comics (漫画)	42	27.8
	Mining (矿业)	25	26.0		Movies (电影)	45	31.5
	Energy Science (能源科学)	31	28.9		Television (电视)	48	28.3
	Aerospace (航空航天)	29	25.9		Painting (绘画)	29	35.8
	Computer Science (计算机科学)	48	24.4		Dance (舞蹈)	20	29.8
	Communication Technology (通信技术)	33	25.8		Festivals (节日)	18	30.2
					Sculpture (雕塑)	30	33.8
					Music (音乐)	45	32.7
					Food and Drink (饮食)	20	24.5

Figure 15: Statistics of All Categories in the Dataset: Counts and Question Token Number Distribution.

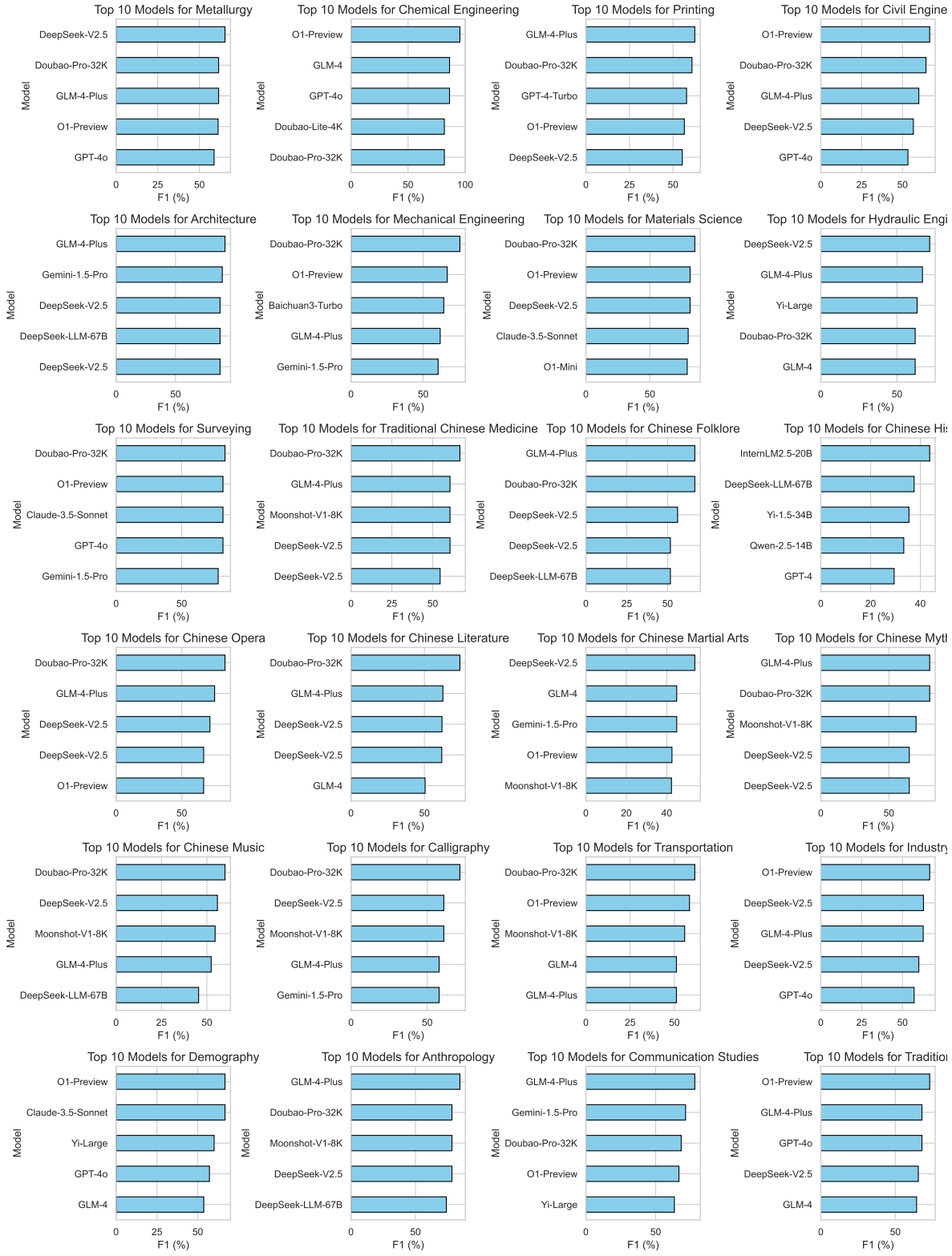


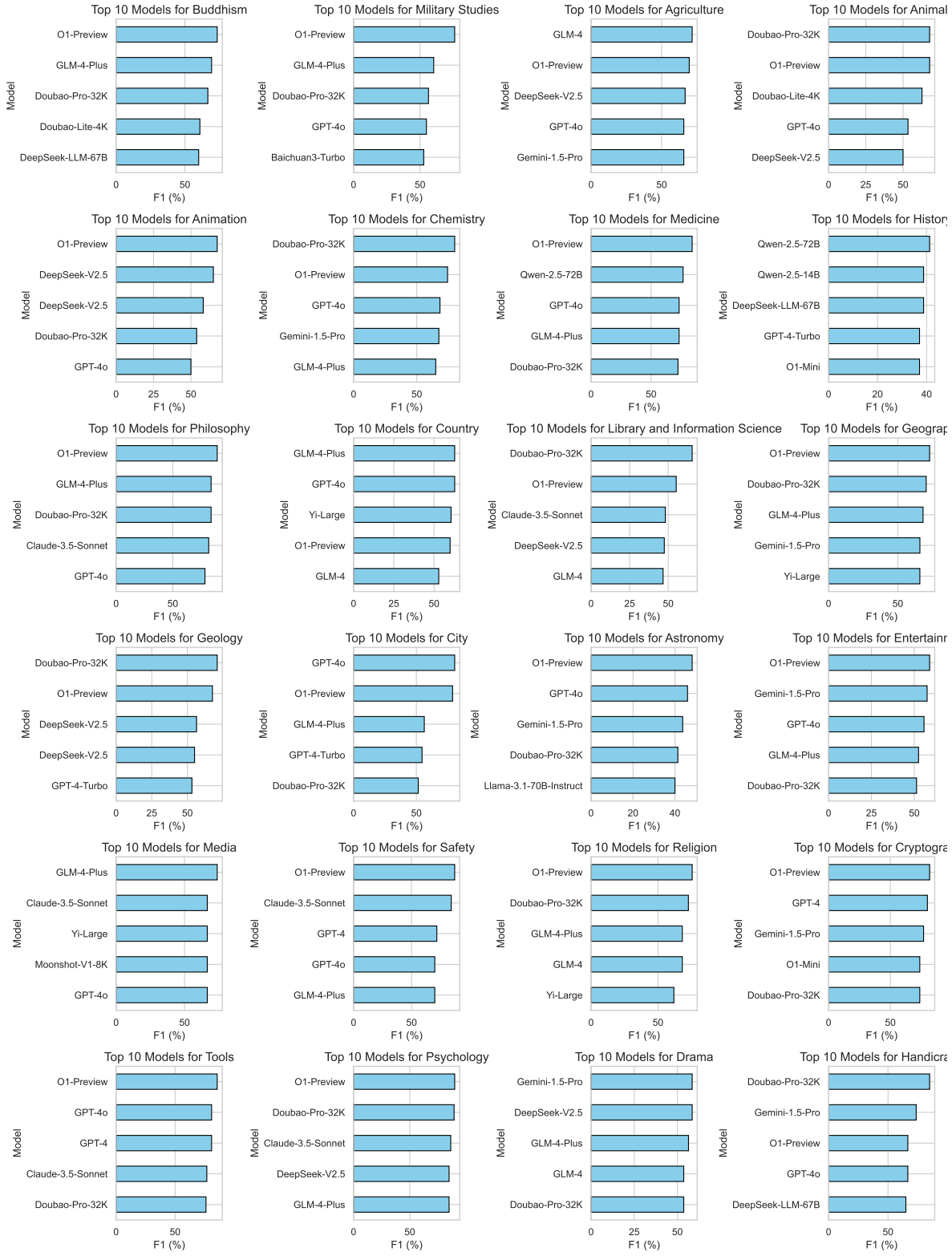
(a) Part 1

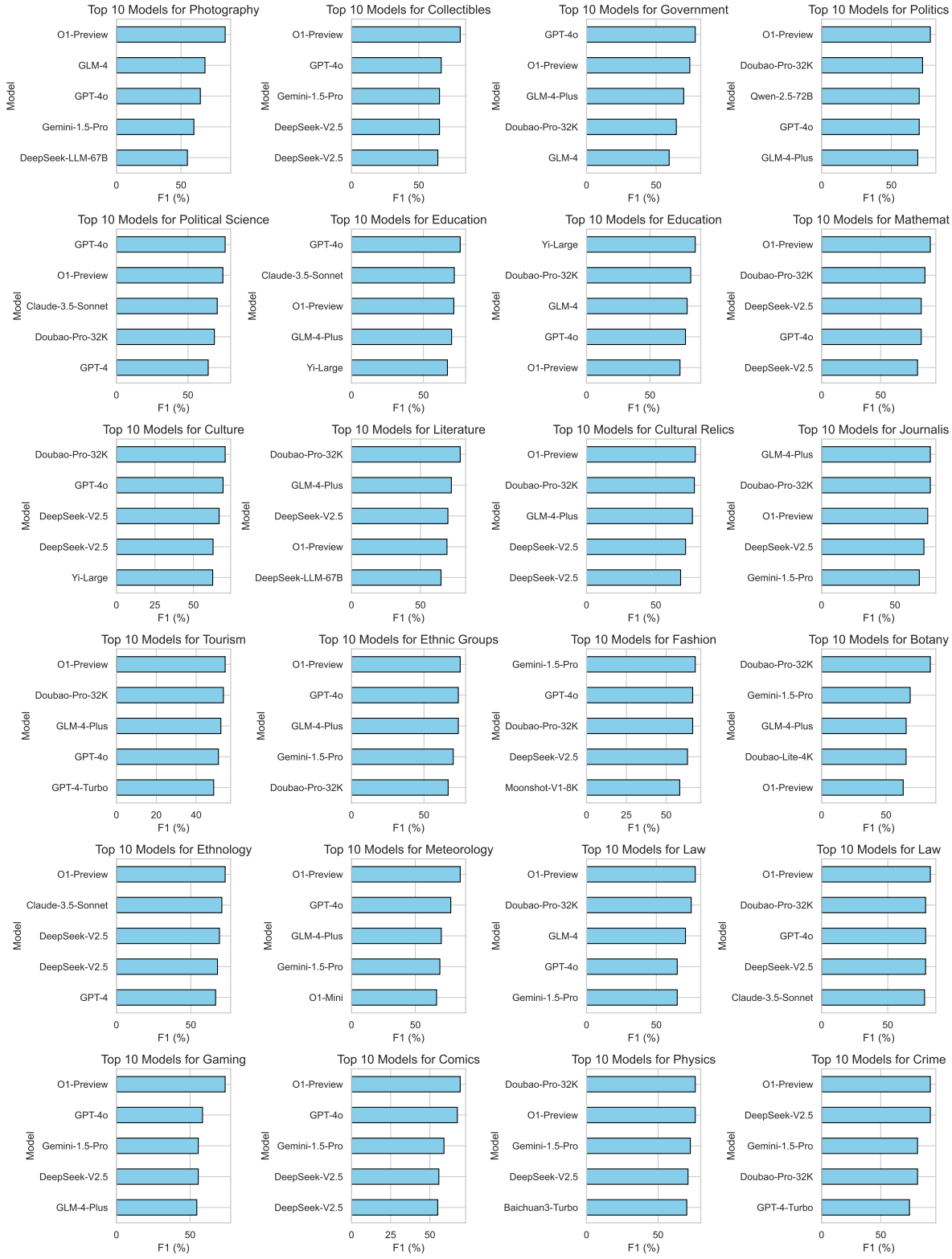


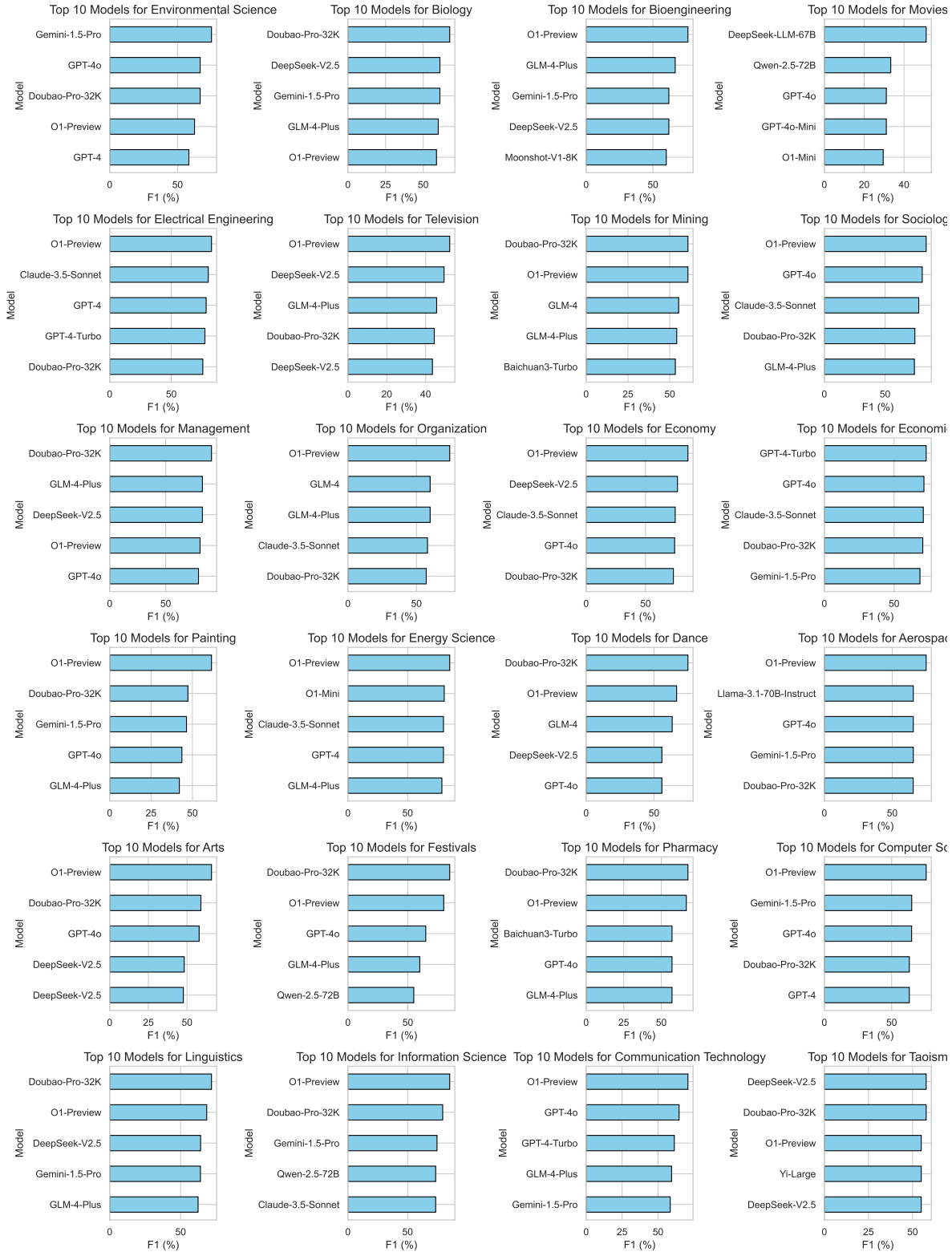
(b) Part 2

Figure 16: Distribution of question difficulty across 99 categories in the dataset. Each violin represents the spread of difficulty within a category, measured by the number of models that failed to answer the question correctly. Narrow and symmetrical violins indicate more uniform difficulty distributions, while wider violins or those with elongated tails suggest greater variability in question difficulty.


Top 10 Models for Metallurgy
DeepSeek-V2.5
Doubao-Pro-32K
GLM-4-Plus
O1-Preview
GPT-4o
Top 10 Models for Chemical Engineering
O1-Preview
GLM-4
GPT-4o
Doubao-Lite-4K
Doubao-Pro-32K
Top 10 Models for Printing
GLM-4-Plus
Doubao-Pro-32K
GPT-4-Turbo
O1-Preview
DeepSeek-V2.5
Top 10 Models for Civil Engine
O1-Preview
Doubao-Pro-32K
GLM-4-Plus
DeepSeek-V2.5
GPT-4o
Top 10 Models for Architecture
GLM-4-Plus
Gemini-1.5-Pro
DeepSeek-V2.5
DeepSeek-LLM-67B
DeepSeek-V2.5
Top 10 Models for Mechanical Engineering
Doubao-Pro-32K
O1-Preview
Baichuan3-Turbo
GLM-4-Plus
Gemini-1.5-Pro
Top 10 Models for Materials Science
Doubao-Pro-32K
O1-Preview
DeepSeek-V2.5
Claude-3.5-Sonnet
O1-Mini
Top 10 Models for Hydraulic Engi
DeepSeek-V2.5
GLM-4-Plus
Yi-Large
Doubao-Pro-32K
GLM-4
Top 10 Models for Surveying
Doubao-Pro-32K
O1-Preview
Claude-3.5-Sonnet
GPT-4o
Gemini-1.5-Pro
Top 10 Models for Traditional Chinese Medicine
Doubao-Pro-32K
GLM-4-Plus
Moonshot-V1-8K
DeepSeek-V2.5
DeepSeek-V2.5
Top 10 Models for Chinese Folklore
GLM-4-Plus
Doubao-Pro-32K
DeepSeek-V2.5
DeepSeek-V2.5
DeepSeek-LLM-67B
Top 10 Models for Chinese Hi
InternLM2.5-20B
DeepSeek-LLM-67B
Yi-1.5-34B
Qwen-2.5-14B
GPT-4
Top 10 Models for Chinese Opera
Doubao-Pro-32K
GLM-4-Plus
DeepSeek-V2.5
DeepSeek-V2.5
O1-Preview
Top 10 Models for Chinese Literature
Doubao-Pro-32K
GLM-4-Plus
DeepSeek-V2.5
DeepSeek-V2.5
GLM-4
Top 10 Models for Chinese Martial Arts
DeepSeek-V2.5
GLM-4
Gemini-1.5-Pro
O1-Preview
Moonshot-V1-8K
Top 10 Models for Chinese Mytl
GLM-4-Plus
Doubao-Pro-32K
Moonshot-V1-8K
DeepSeek-V2.5
DeepSeek-V2.5
Top 10 Models for Chinese Music
Doubao-Pro-32K
DeepSeek-V2.5
Moonshot-V1-8K
GLM-4-Plus
DeepSeek-LLM-67B
Top 10 Models for Calligraphy
Doubao-Pro-32K
DeepSeek-V2.5
Moonshot-V1-8K
GLM-4-Plus
Gemini-1.5-Pro
Top 10 Models for Transportation
Doubao-Pro-32K
O1-Preview
Moonshot-V1-8K
GLM-4
GLM-4-Plus
Top 10 Models for Industry
O1-Preview
DeepSeek-V2.5
GLM-4-Plus
DeepSeek-V2.5
GPT-4o
Top 10 Models for Demography
O1-Preview
Claude-3.5-Sonnet
Yi-Large
GPT-4o
GLM-4
Top 10 Models for Anthropology
GLM-4-Plus
Doubao-Pro-32K
Moonshot-V1-8K
DeepSeek-V2.5
DeepSeek-LLM-67B
Top 10 Models for Communication Studies
GLM-4-Plus
Gemini-1.5-Pro
Doubao-Pro-32K
O1-Preview
Yi-Large
Top 10 Models for Traditio
O1-Preview
GLM-4-Plus
GPT-4o
DeepSeek-V2.5
GLM-4
Model
F1 (%)


Top 10 Models for Buddhism
O1-Preview
GLM-4-Plus
Doubao-Pro-32K
Doubao-Lite-4K
DeepSeek-LLM-67B
Top 10 Models for Military Studies
O1-Preview
GLM-4-Plus
Doubao-Pro-32K
GPT-4o
Baichuan3-Turbo
Top 10 Models for Agriculture
GLM-4
O1-Preview
DeepSeek-V2.5
GPT-4o
Gemini-1.5-Pro
Top 10 Models for Animal
Doubao-Pro-32K
O1-Preview
Doubao-Lite-4K
GPT-4o
DeepSeek-V2.5
Top 10 Models for Animation
O1-Preview
DeepSeek-V2.5
DeepSeek-V2.5
Doubao-Pro-32K
GPT-4o
Top 10 Models for Chemistry
Doubao-Pro-32K
O1-Preview
GPT-4o
Gemini-1.5-Pro
GLM-4-Plus
Top 10 Models for Medicine
O1-Preview
Qwen-2.5-72B
GPT-4o
GLM-4-Plus
Doubao-Pro-32K
Top 10 Models for Histor
Qwen-2.5-72B
Qwen-2.5-14B
DeepSeek-LLM-67B
GPT-4-Turbo
O1-Mini
Top 10 Models for Philosophy
O1-Preview
GLM-4-Plus
Doubao-Pro-32K
Claude-3.5-Sonnet
GPT-4o
Top 10 Models for Country
GLM-4-Plus
GPT-4o
Yi-Large
O1-Preview
GLM-4
Top 10 Models for Library and Information Science
Doubao-Pro-32K
O1-Preview
Claude-3.5-Sonnet
DeepSeek-V2.5
GLM-4
Top 10 Models for Geograp
O1-Preview
Doubao-Pro-32K
GLM-4-Plus
Gemini-1.5-Pro
Yi-Large
Top 10 Models for Geology
Doubao-Pro-32K
O1-Preview
DeepSeek-V2.5
DeepSeek-V2.5
GPT-4-Turbo
Top 10 Models for City
GPT-4o
O1-Preview
GLM-4-Plus
GPT-4-Turbo
Doubao-Pro-32K
Top 10 Models for Astronomy
O1-Preview
GPT-4o
Gemini-1.5-Pro
Doubao-Pro-32K
Llama-3.1-70B-Instruct
Top 10 Models for Entertainr
O1-Preview
Gemini-1.5-Pro
GPT-4o
GLM-4-Plus
Doubao-Pro-32K
Top 10 Models for Media
GLM-4-Plus
Claude-3.5-Sonnet
Yi-Large
Moonshot-V1-8K
GPT-4o
Top 10 Models for Safety
O1-Preview
Claude-3.5-Sonnet
GPT-4
GPT-4o
GLM-4-Plus
Top 10 Models for Religion
O1-Preview
Doubao-Pro-32K
GLM-4-Plus
GLM-4
Yi-Large
Top 10 Models for Cryptogra
O1-Preview
GPT-4
Gemini-1.5-Pro
O1-Mini
Doubao-Pro-32K
Top 10 Models for Tools
O1-Preview
GPT-4o
GPT-4
Claude-3.5-Sonnet
Doubao-Pro-32K
Top 10 Models for Psychology
O1-Preview
Doubao-Pro-32K
Claude-3.5-Sonnet
DeepSeek-V2.5
GLM-4-Plus
Top 10 Models for Drama
Gemini-1.5-Pro
DeepSeek-V2.5
GLM-4-Plus
GLM-4
Doubao-Pro-32K
Top 10 Models for Handicra
Doubao-Pro-32K
Gemini-1.5-Pro
O1-Preview
GPT-4o
DeepSeek-LLM-67B
Model
F1 (%)


Top 10 Models for Photography
O1-Preview
GLM-4
GPT-4o
Gemini-1.5-Pro
DeepSeek-LLM-67B
Top 10 Models for Collectibles
O1-Preview
GPT-4o
Gemini-1.5-Pro
DeepSeek-V2.5
DeepSeek-V2.5
Top 10 Models for Government
GPT-4o
O1-Preview
GLM-4-Plus
Doubao-Pro-32K
GLM-4
Top 10 Models for Politics
O1-Preview
Doubao-Pro-32K
Qwen-2.5-72B
GPT-4o
GLM-4-Plus
Top 10 Models for Political Science
GPT-4o
O1-Preview
Claude-3.5-Sonnet
Doubao-Pro-32K
GPT-4
Top 10 Models for Education
GPT-4o
Claude-3.5-Sonnet
O1-Preview
GLM-4-Plus
Yi-Large
Top 10 Models for Education
Yi-Large
Doubao-Pro-32K
GLM-4
GPT-4o
O1-Preview
Top 10 Models for Mathemat
O1-Preview
Doubao-Pro-32K
DeepSeek-V2.5
GPT-4o
DeepSeek-V2.5
Top 10 Models for Culture
Doubao-Pro-32K
GPT-4o
DeepSeek-V2.5
DeepSeek-V2.5
Yi-Large
Top 10 Models for Literature
Doubao-Pro-32K
GLM-4-Plus
DeepSeek-V2.5
O1-Preview
DeepSeek-LLM-67B
Top 10 Models for Cultural Relics
O1-Preview
Doubao-Pro-32K
GLM-4-Plus
DeepSeek-V2.5
DeepSeek-V2.5
Top 10 Models for Journalis
GLM-4-Plus
Doubao-Pro-32K
O1-Preview
DeepSeek-V2.5
Gemini-1.5-Pro
Top 10 Models for Tourism
O1-Preview
Doubao-Pro-32K
GLM-4-Plus
GPT-4o
GPT-4-Turbo
Top 10 Models for Ethnic Groups
O1-Preview
GPT-4o
GLM-4-Plus
Gemini-1.5-Pro
Doubao-Pro-32K
Top 10 Models for Fashion
Gemini-1.5-Pro
GPT-4o
Doubao-Pro-32K
DeepSeek-V2.5
Moonshot-V1-8K
Top 10 Models for Botany
Doubao-Pro-32K
Gemini-1.5-Pro
GLM-4-Plus
Doubao-Lite-4K
O1-Preview
Top 10 Models for Ethnology
O1-Preview
Claude-3.5-Sonnet
DeepSeek-V2.5
DeepSeek-V2.5
GPT-4
Top 10 Models for Meteorology
O1-Preview
GPT-4o
GLM-4-Plus
Gemini-1.5-Pro
O1-Mini
Top 10 Models for Law
O1-Preview
Doubao-Pro-32K
GLM-4
GPT-4o
Gemini-1.5-Pro
Top 10 Models for Law
O1-Preview
Doubao-Pro-32K
GPT-4o
DeepSeek-V2.5
Claude-3.5-Sonnet
Top 10 Models for Gaming
O1-Preview
GPT-4o
Gemini-1.5-Pro
DeepSeek-V2.5
GLM-4-Plus
Top 10 Models for Comics
O1-Preview
GPT-4o
Gemini-1.5-Pro
DeepSeek-V2.5
DeepSeek-V2.5
Top 10 Models for Physics
Doubao-Pro-32K
O1-Preview
Gemini-1.5-Pro
DeepSeek-V2.5
Baichuan3-Turbo
Top 10 Models for Crime
O1-Preview
DeepSeek-V2.5
Gemini-1.5-Pro
Doubao-Pro-32K
GPT-4-Turbo
Model
F1 (%)


Top 10 Models for Environmental Science
Gemini-1.5-Pro
GPT-4o
Doubao-Pro-32K
O1-Preview
GPT-4
Top 10 Models for Biology
Doubao-Pro-32K
DeepSeek-V2.5
Gemini-1.5-Pro
GLM-4-Plus
O1-Preview
Top 10 Models for Bioengineering
O1-Preview
GLM-4-Plus
Gemini-1.5-Pro
DeepSeek-V2.5
Moonshot-V1-8K
Top 10 Models for Movies
DeepSeek-LLM-67B
Qwen-2.5-72B
GPT-4o
GPT-4o-Mini
O1-Mini
Top 10 Models for Electrical Engineering
O1-Preview
Claude-3.5-Sonnet
GPT-4
GPT-4-Turbo
Doubao-Pro-32K
Top 10 Models for Television
O1-Preview
DeepSeek-V2.5
GLM-4-Plus
Doubao-Pro-32K
DeepSeek-V2.5
Top 10 Models for Mining
Doubao-Pro-32K
O1-Preview
GLM-4
GLM-4-Plus
Baichuan3-Turbo
Top 10 Models for Sociolog
O1-Preview
GPT-4o
Claude-3.5-Sonnet
Doubao-Pro-32K
GLM-4-Plus
Top 10 Models for Management
Doubao-Pro-32K
GLM-4-Plus
DeepSeek-V2.5
O1-Preview
GPT-4o
Top 10 Models for Organization
O1-Preview
GLM-4
GLM-4-Plus
Claude-3.5-Sonnet
Doubao-Pro-32K
Top 10 Models for Economy
O1-Preview
DeepSeek-V2.5
Claude-3.5-Sonnet
GPT-4o
Doubao-Pro-32K
Top 10 Models for Economi
GPT-4-Turbo
GPT-4o
Claude-3.5-Sonnet
Doubao-Pro-32K
Gemini-1.5-Pro
Top 10 Models for Painting
O1-Preview
Doubao-Pro-32K
Gemini-1.5-Pro
GPT-4o
GLM-4-Plus
Top 10 Models for Energy Science
O1-Preview
O1-Mini
Claude-3.5-Sonnet
GPT-4
GLM-4-Plus
Top 10 Models for Dance
Doubao-Pro-32K
O1-Preview
GLM-4
DeepSeek-V2.5
GPT-4o
Top 10 Models for Aerospac
O1-Preview
Llama-3.1-70B-Instruct
GPT-4o
Gemini-1.5-Pro
Doubao-Pro-32K
Top 10 Models for Arts
O1-Preview
Doubao-Pro-32K
GPT-4o
DeepSeek-V2.5
DeepSeek-V2.5
Top 10 Models for Festivals
Doubao-Pro-32K
O1-Preview
GPT-4o
GLM-4-Plus
Qwen-2.5-72B
Top 10 Models for Pharmacy
Doubao-Pro-32K
O1-Preview
Baichuan3-Turbo
GPT-4o
GLM-4-Plus
Top 10 Models for Computer Sc
O1-Preview
Gemini-1.5-Pro
GPT-4o
Doubao-Pro-32K
GPT-4
Top 10 Models for Linguistics
Doubao-Pro-32K
O1-Preview
DeepSeek-V2.5
Gemini-1.5-Pro
GLM-4-Plus
Top 10 Models for Information Science
O1-Preview
Doubao-Pro-32K
Gemini-1.5-Pro
Qwen-2.5-72B
Claude-3.5-Sonnet
Top 10 Models for Communication Technology
O1-Preview
GPT-4o
GPT-4-Turbo
GLM-4-Plus
Gemini-1.5-Pro
Top 10 Models for Taoism
DeepSeek-V2.5
Doubao-Pro-32K
O1-Preview
Yi-Large
DeepSeek-V2.5
Model
F1 (%)

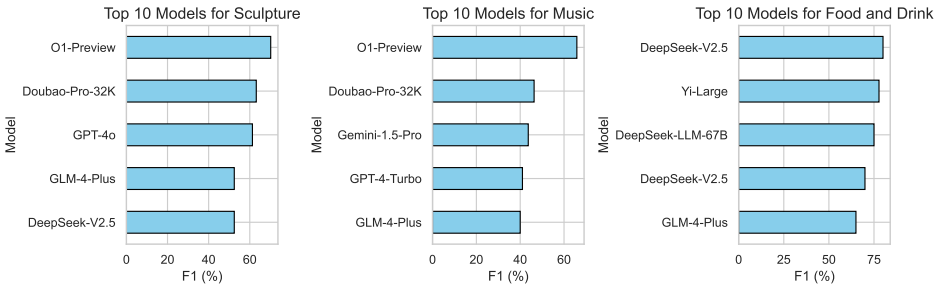


Figure 17: Bar charts showcasing the top 10 models ranked by F1 score across diverse topics.

Figure 17 shows the best performing models in each topic, which are sorted by F1 score. It can be seen that the strongest models in different fields may not be the same. The results emphasize that model performance varies with the field, which can help researchers gain a deeper understanding of the applicability of each model to a specific field.

K Generation and validation of question-answer pairs

The generation and validation of question-answer pairs both use OpenAI's gpt-4o-0806. The specific prompts are shown in Figures 18, 19, and 20.

L Analysis of Model Calibration

现在需要你根据给定的文档生成一个事实类问题和对应的标准答案，需要满足下列要求:
1. 生成的问题必须关联到客观世界的知识，例如可以询问“2024年诺贝尔物理学奖的获得者是谁？”不得构造涉及个人观点或感受相关的主观问题，如“你如何看待xxx？”。
2. 所提出的问题应该有且只有一个明确且无争议的实体作为答案，且问题表述中不应存在任何形式的模糊性或歧义。例如，避免提问“巴拉克和米歇尔·奥巴马在哪里会面？”因为无法确定是指哪一次会面；同样不要问“白民国人身体的特点是什么？”因为这个问题过于模糊，没有明确的答案。“周汝昌最为人熟知的著作是哪个？”也是不合格问题，因为“最熟知”可能是有争议的。
3. 问题的答案应当是时间不变的，不会随着时间的推移而改变。例如，“美国现任总统是谁？”就不是一个合适的问题，因为总统身份会随选举结果改变。
4. 问题应该具有一定的难度，以体现出一定的挑战性。例如：电影《脱衣舞娘》是由同名小说改编的，该小说的作者是谁?
5. 如果问题的答案为英文人名，请给出中文翻译后的名字和括号里带上英文原名，格式如：雅各布·福格（Jakob Fugger）。
6. 生成的问题需要与给定的类目相关
请将生成的问题和答案以JSON格式返回，具体格式如下：
{"question": "这里填写生成的问题", "answer": "这里填写对应的标准答案"}
###以下是一些示例###
示例一
类目：娱乐
文档内容：2022年国际足联世界杯为第22届国际足联世界杯，于2022年11月20日至12月18日在卡塔尔举行[2][3]，成为全球爆发防疫后首个终结限制的大型国际体育盛事。考量到气候因素，本届世界杯亦是首次于11月至12月北半球秋季[注 1]举行之世界杯。决赛于卡塔尔的卢赛尔体育场举行，由阿根廷队对阵卫冕冠军法国队。双方先于比赛正规时间踢至加时赛以3–3赛和，后在点球大战中阿根廷以4–2击败法国，赢得了此届世界杯，也是阿根廷继1986年世界杯后，相隔36年再度于世界杯夺冠，继巴西，意大利及德国后第四支三次冠军的球队，也成为继巴西后第二支在亚洲夺冠的南美洲球队。克罗地亚则以2比1击败该年黑马摩洛哥赢得季军[4][5]。
返回结果：{"question": "2022年世界杯决赛点球大战中阿根廷队是以多少击败法国队", "answer": "4–2"}
示例二
类目：政治
文档内容：中国与世界贸易组织（英语：China and the World Trade Organization）指中华人民共和国与世界贸易组织的关系。在部长级会议达成协议后，中国于2001年12月11日成为世界贸易组织成员。[1][2]在承认这一点之前，双方进行了漫长的谈判，并且需要对中国经济进行重大改革。世贸组织的成员资格一直存在争议，对其它国家产生了重大的经济和政治影响(也被称为“中国冲击”)，对世贸组织框架与中国经济模式之间的不匹配也存在争议。[3][4]评估和执行合规已成为中美贸易关系中的问题，包括中国的不合规行为如何为本国经济创造利益。[5][6]
返回结果：{"question": "中国是哪一年正式成为世界贸易组织成员？", "answer": "2001"}
###
让我们开始吧！

Figure 18: The prompt for generating question-answer pairs.

你是一个数据质量检查员，现在需要你检查下面生成的问题是否满足以下要求：
1.生成的问题必须对客观世界的知识的提问，例如可以询问“2024年诺贝尔物理学奖的获得者是谁？”不得构造涉及个人观点或感受相关的主观问题，如“你如何看待xxx？”。
2. 问题应该有且只有一个明确且无争议的实体作为答案，且问题表述中不应存在任何形式的模糊性或歧义。例如，避免提问“巴拉克和米歇尔·奥巴马在哪里会面？”因为无法确定是指哪一次会面；同样不要问“白民国人身体的特点是什么？”因为这个问题过于模糊，没有明确的答案。注意如果回答是多个实体也不满足要求，例如：“软体动物、腕足动物及被囊动物”
3. 问题的答案应当是时间不变的，不会随着时间的推移而改变。例如，“美国现任总统是谁？”就不是一个合适的问题，因为总统身份会随选举结果改变。
如果问题不合格则解释并输出“【否】”， 如果问题合格则直接输出“【是】”
以下是一些示例：
问题：《黄帝内经》中, 援物比类思维方式包括哪些核心概念?
评价：该问题不是只有一个确切答案，【否】

问题：建筑理论主要研究什么内容?
评价：该问题不具体，回答不是只有一个实体，【否】

问题：成立了程派高氏八卦掌的高义盛原籍是哪里?
评价：回答范围不明确，不清楚是回答到城市还是省份，【否】

问题：自由恋爱主义最初的目标是将哪些事务与国家分离?
评价：该问题不是只有一个答案，【否】

问题：汉十高速公路连接的武汉市和哪两个城市?
评价：【是】
如果问题不合格则输出原因并最后输出“【否】”， 如果问题合格则直接输出“【是】”，注意如果认为问题不合格需要输出原因
让我们开始吧！

Figure 19: The prompt for validating criteria.

请结合检索材料判断给定的问题是否满足以下要求以及答案是否正确：
1. 问题应该有且只有一个明确且无争议的实体作为答案，且问题表述中不应存在任何形式的模糊性或歧义。例如，避免提问“巴拉克和米歇尔·奥巴马在哪里会面？”因为无法确定是指哪一次会面；同样不要问“白民国人身体的特点是什么？”因为这个问题过于模糊，没有明确的答案。注意如果回答是多个实体也不满足要求，例如：“软体动物、腕足动物及被囊动物”
2. 问题的答案应当是时间不变的，不会随着时间的推移而改变。例如，“美国现任总统是谁？”就不是一个合适的问题，因为总统身份会随选举结果改变。
3. 检索材料应该能够验证给定答案的正确性，并且这些材料中不应存在其他与之相矛盾的答案。如果检索材料中存在不同的答案，则该问题被视为无法验证或答案不明确，不满足要求。
请依据上述标准，审查并确保提出的问题及其答案符合要求，如果不满足上面任一要求，则输出原因，最后输出“【不合格】”， 否则输出检索材料对应的片段，最后输出“【合格】”

Figure 20: The prompt for verifying factual accuracy based on RAG.

请阅读以下问题：
{question}
请基于此问题提供你的最佳答案，并用0到100的分数表示你对该答案的信心（置信度）。请以如下的JSON格式给出回复：
{
"answer": "你的答案",
"confidence_score": 你的置信度
}

Figure 21: The prompt for guiding the model to output confidence.