MRCEval: A Comprehensive, Challenging and Accessible Machine Reading Comprehension Benchmark

Anonymous ACL submission

Abstract

Machine Reading Comprehension (MRC) is an essential task of evaluating natural language understanding. Previous MRC datasets focus on the specific skill of reading comprehension, lacking the requirements of a comprehensive MRC benchmark to assess Large Language Models (LLMs) thoroughly. To fill this gap, we first introduce a novel taxonomy to classify the needed capabilities for RC, then based on the taxonomy, we automatically build an MRC benchmark **MRCEval**, which employs powerful LLMs as sample generators and selection judges. MRCEval is a comprehensive, challenging and accessible benchmark, which consists of three main tasks and 13 sub-tasks with a total of 2.2K high-quality multi-choice questions. We perform an extensive evaluation of 28 widely used open-source and proprietary models, highlighting that MRC continues to present significant challenges even in the era of LLMs. Project is available at github. [1]

1 Introduction

With the advancement of Large Language Models (LLMs), such as o3-mini (OpenAI, 2025) and DeepSeek-R1 (Guo et al., 2025), their remarkable language understanding and generation capabilities continue to impress AI communication. Machine Reading Comprehension (MRC), which requires machine reading and comprehending the given passage, then answering the questions correctly, is the fundamental evaluation of natural language understanding (Hirschman et al., 1999).

To facilitate the reading comprehension (RC) capability of the machine, a great number of datasets are proposed (Rajpurkar et al., 2016; Yang et al., 2018; Trivedi et al., 2022; Yao et al., 2023; Parmar et al., 2024). However, all these datasets only focus on a specific RC skill, and there is a lack of a unified benchmark to evaluate the RC challenges of

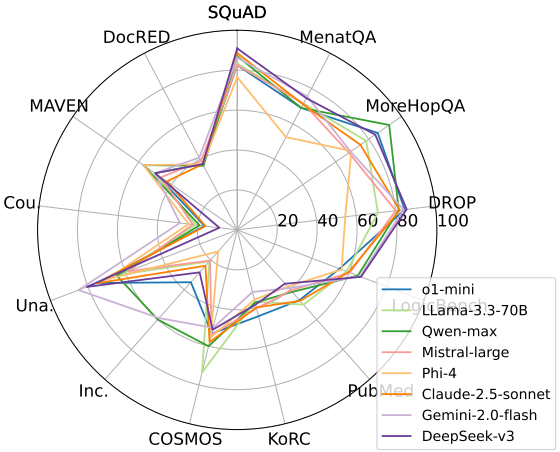

Figure 1: Performance on MRCEval Benchmark of representative models.

LLMs. Current MRC taxonomies are fine-grained, complicated, and not suitable for creating a comprehensive but accessible MRC benchmark. On the other hand, there are many new issues appeared with LLMs, such as hallucination (Ji et al., 2023) and knowledge conflict (Xu et al., 2024). These issues make LLMs unable to accurately understand the given context and answer the question incorrectly, making RC more challenging for LLMs.

To address these issues, we first introduce a novel taxonomy for MRC. Drawing inspiration from the machine's question-answering process of text comprehension (Lehnert and Lehnert, 1978), we summarize the needed RC skills for LLMs into three levels: context comprehension, external knowledge comprehension and reasoning. As McCarthy (1990) said, machines first understand the facts in the passage, then grasp the expression of the general information about the world that could allow getting the answers to the questions by formal reasoning from the facts and the general information. These levels correspond to the accurate comprehension of facts information, the acquisition of external knowledge, and the integration of

[1] https://github.com/anonymous

facts and expertise for reasoning.

Based on the proposed taxonomy, we introduce MRCEval, a comprehensive, challenging and accessible MRC benchmark designed to assess RC capabilities of LLMs. MRCEval comprises three main tasks with 13 sub-tasks, and a total of 2.2K high-quality multi-choice questions. It is built employing GPT-4o (Hurst et al., 2024) as the generator, and three light-weight models as judges to generate the high-quality and challenging samples.

We conduct an extensive evaluation on 28 representative open-source and closed-source models. Figure 1 summarizes the performance of popular competitive models on the 13 sub-tasks in MRCEval. It reveals that MRC still remains challenging, even the most competitive models like o1-mini and Gemini-2.0-flash still perform badly on MRCEval, despite their strong performance on standard benchmarks. As far as we know, MRCEval is the first comprehensive, challenging and accessible benchmark tailored for MRC, contributing to the advancement of natural language understanding in LLMs.

2 Related Work

MRC datasets. Numerous datasets have been proposed in the past decade. Cloze-test format (SQuAD (Rajpurkar et al., 2016)), free-answer format (NarrativeQA (Kočiský et al., 2018)), arithmetic (DROP (Dua et al., 2019)), commonsense (OpenBookQA (Mihaylov et al., 2018)), world knowledge (Natural Questions (Kwiatkowski et al., 2019)), reasoning (HotpotQA (Yang et al., 2018)), logical reasoning (ReClor (Yu et al., 2020)), multi-hop reasoning (MorehopQA (Schnitzler et al., 2024)), temporal reasoning (TORQUE (Ning et al., 2020)), medical (MedMCQA (Pal et al., 2022)) and science (ScienceQA (Lu et al., 2022)).

Benchmarking MRC. Researchers summarize RC skills in different aspects. Chen (2018) first defines the MRC task depending on the answer type: cloze style, multiple choice, span prediction, and free-form answer. Then Schlegel et al. (2020) analyze modern MRC gold standards and propose a qualitative annotation schema to evaluate popular MRC datasets. Sugawara et al. (2021) further provide a theoretical basis for the design of MRC datasets based on psychology as well as psychometrics and summarize it in terms of the prerequisites for benchmarking MRC. Based on these research, Rogers et al. (2023) propose an alternative taxonomy for a wider range of RC skills.

Task	Dataset	instances
Context comprehension	-	**870**
Facts understanding		
-entity	SQuAD	132
-relation	DocRED	110
-event	MAVEN	28
Context faithful		
-counterfactual	FaithEval	200
-unanswerable	FaithEval	200
-inconsistent	FaithEval	200
External knowledge comprehension	-	**600**
Commonsense knowledge	COSMOS	200
World knowledge	KoRC	200
Domain knowledge	PubMed	200
Reasoning	-	**784**
Logical reasoning	LogicBench	184
Arithmetic reasoning	DROP	200
Multi-hop reasoning	MoreHopQA	200
Temporal reasoning	MenatQA	200
Overall		**2254**

Table 1: MRCEval tasks division.

3 MRCEval Benchmark

3.1 Taxonomy

Building on the three levels of machine text comprehension (McCarthy, 1990), we define three key aspects of MRC: **Context Comprehension**, **External Knowledge Comprehension**, and **Reasoning**.

Context comprehension. Focuses on facts understanding and models' context-faithful capability (Ming et al., 2024). First, models should understand the facts, which include entities, relations and events related facts in the text. Then overcome the hallucination from their parameters to be faithful to the given context.

Externation knowledge comprehension. Focuses on external knowledge acquisition and application (Wang et al., 2021). Models are supposed to incorporate the external knowledge outside the given text, which is from the real world, namely the world general knowledge, commonsense knowledge, and the specific domain knowledge to comprehend the passage and the question.

Reasoning. Focuses on deep context comprehension and inference, which is an essential capability for complex problem-solving (Qiao et al., 2023). In the MRC task, we classify reasoning into logical reasoning, arithmetic reasoning, multi-hop reasoning and temporal reasoning.

3.2 Benchmark Construction

Based on the proposed taxonomy, we construct the MRCEval benchmark.

Source datasets. For each sub-task, we collect representative datasets, including SQuAD (Rajpurkar et al., 2016), DocRED (Yao et al., 2019), and MAVEN (Wang et al., 2020), FaithEval (Ming et al., 2024), KoRC (Yao et al., 2023), COSMOS (Huang et al., 2019), PubMed (Jin et al., 2019), LogicBench (Parmar et al., 2024), DROP (Dua et al., 2019), MoreHopQA (Schnitzler et al., 2024) and MenatQA (Wei et al., 2023). Due to access rights, we only use their development sets.

Multi-choice samples construction. For multi-choice format datasets COSMOS, LogicBench, and counterfactual part of FaithEval, we have retained the original data. For some question-answering datasets SQuAD, KoRC, DROP, MenatQA, and unanswerable part of FaithEval, we set the answer as the correct choice and prompt GPT-4o (Hurst et al., 2024) to generate three incorrect choices. For others, we use the automated method to construct three incorrect choices. As for DocRED and MAVEN, since they have no questions, we prompt GPT-4o to generate a facts-related question and choices for each passage as a sample.

LLMs as judges. To select challenging samples for LLMs, we adopt a voting strategy employing three light-weight LLMs as judges: LLama-3-8B-Instruct (Dubey et al., 2024), Qwen-2.5-7B-Instruct (Yang et al., 2024), and GPT-4o-mini (Hurst et al., 2024). For each sample, if at least one of the judges answers incorrectly, we put the sample as the candidate. Then for each sub-task, we randomly select 200 candidates to build the final benchmark.

Statistic. As Table 1, MRCEval is an English benchmark, which consists of general topics with three main tasks: context comprehension, external knowledge comprehension, reasoning, and 13 sub-tasks. Context comprehension includes facts understanding (entity, relation and event facts) and context-faithful (counterfactual, unanswerable and inconsistent). External knowledge comprehension includes world knowledge, commonsense knowledge and domain knowledge. Reasoning includes logical reasoning, arithmetic reasoning, multi-hop reasoning and temporal reasoning. In sum, MRCEval has 2254 multi-choice samples, and each sub-task has nearly 200 samples.

4 Evaluation

Models. We evaluate extensive popular open-source and closed-source models. For open-source models, we consider their instruction-tuned models, including LLama-3.1-8B-Instruct, LLama-3.3-70B-Instruct (Dubey et al., 2024), Qwen-2.5-14B-Instruct (Yang et al., 2024), Mistral-7B-Instruct-v0.3, Mistral-Nemo-Instruct-2407, Mistral-8x7B-Instruct-v0.1 (Jiang et al., 2023), Gemma-2-9B-it, Gemma-2-27B-it (Team et al., 2024), Phi-3-mini-4k-Instruct, Phi-3-medium-4k-Instruct, Phi-4 (Abdin et al., 2024), Command-R-7B-12-2024 (Cohere, 2024), DeepSeek-R1-Distll-LLama-8B, DeepSeek-R1-Distll-Qwen-14B (Guo et al., 2025), DeepSeek-v3 (Liu et al., 2024). For closed-source models, we access them through their official API, including GPT-3.5-turbo, GPT-4-turbo, GPT-4o, o1-mini, o3-mini, Gemini-1.5-flash, Gemini-1.5-pro, Gemini-2.0-flash, Gemini-2.0-flash-lite-preview-02-05, Claude-3.5-haiku-20241022, Claude-3.5-sonnet-20241022, Mistral-large and Qwen-max-2025-01-25.

Settings. All models use greedy sampling or temperature of 0.0, except for DeepSeek series, which follows their official settings with temperature of 0.60 and top-p of 0.95. For all tasks, we append the instruction to the beginning of each sample: *You are an expert in reading comprehension. Read the passage and select one of the most appropriate options to answer the question.* We report accuracy as the metric from a single run result.

5 Results and Analysis

5.1 Overall Performance

LLMs are good at facts extraction, while bad at context-faithful. LLMs have great performance at entity-facts understanding, which demonstrates that they can comprehend simple entity facts and are good at extracting entity answers directly from the passage. Stronger models can recognize which questions cannot be answered, they know to answer the question based on the given text, rather than their trained parameters. These two aspects confirm that LLMs have a good capability for simple information extraction. As for more complicated relation or event facts, models perform worse. Even the most competitive models, like Gemini-2.0-flash or Qwen-max, are still struggling with them. Large commercial models are better at inconsistent tasks but worse at counterfactual tasks, smaller open-source models do the opposite. This is because models with more parameters remember more facts and can easily fit them into memory, while smaller models are better at reasoning itself.

Models	Context						External Knowledge			Reasoning				Overall			
	Facts Understanding			*Context Faithful*			Com.	Wor.	Dom.	Log.	Ari.	Mul.	Tem.	Con.	Kno.	Rea.	**Avg.**
	Ent.	Rel.	Eve.	Cou.	Una.	Inc.											
Open-source Models																	
Mistral-7B-Instruct-v0.3	57.5	30.9	50.0	**52.0**	27.5	3.0	45.5	30.5	36.5	36.9	31.5	22.5	31.5	33.2	37.5	30.4	33.4
Mistral-8x7B-Instruct-v0.1	59.0	31.8	**57.1**	45.0	39.5	8.5	54.0	28.0	39.5	46.1	41.5	31.5	43.0	36.2	40.5	40.4	38.8
Mistral-Nemo-Instruct-2407	68.1	33.6	39.2	39.5	42.5	5.5	48.5	32.0	40.0	44.0	37.0	31.0	57.9	35.9	40.1	42.4	39.3
Phi-3-mini-4k-Instruct	71.2	31.8	**57.1**	29.5	33.5	8.5	47.5	41.5	25.5	47.8	43.0	63.5	42.0	33.1	38.1	49.1	40.0
Phi-3-medium-4k-Instruct	76.5	**40.0**	46.4	13.5	45.0	9.5	54.5	24.5	31.5	43.4	62.0	53.0	45.5	33.7	36.8	51.1	40.6
LLama-3.1-8B-Instruct	62.1	31.8	46.4	47.0	38.0	33.5	59.0	37.0	38.5	42.4	34.5	37.0	43.0	42.2	44.8	39.2	41.8
Gemma-2-9B-it	77.2	32.7	39.2	32.0	56.9	16.0	53.5	35.5	30.5	51.6	52.5	50.0	52.5	41.2	39.8	51.6	44.4
Phi4	76.5	37.2	**57.1**	24.5	70.0	14.4	55.5	36.0	38.0	55.9	54.5	69.5	52.5	43.2	43.1	58.1	48.4
Command-R-7B-12-2024	79.5	35.4	53.5	43.0	70.0	37.0	35.0	**43.5**	38.5	61.9	46.0	34.0	68.0	52.7	39.0	52.2	48.9
Gemma-2-27B-it	82.5	36.3	35.7	30.5	61.5	26.0	53.0	39.0	36.5	47.8	50.0	69.5	63.5	45.4	42.8	57.9	49.0
DeepSeek-R1-Distll-LLama-8B	72.7	34.5	**57.1**	51.0	44.0	28.0	42.5	30.5	38.5	47.8	77.5	57.4	67.0	45.5	37.1	62.7	49.2
Qwen-2.5-14B-Instruct	84.8	33.6	50.0	24.5	76.5	9.5	58.0	35.5	34.0	58.7	67.0	70.0	66.0	44.1	42.5	65.6	51.2
LLama-3.3-70B-Instruct	83.3	37.3	50.0	21.0	63.5	21.0	**73.5**	37.0	**50.0**	58.7	71.0	78.5	71.5	43.2	**53.5**	70.2	55.3
DeepSeek-R1-Distll-Qwen-14B	79.5	35.4	46.4	36.0	69.5	25.0	45.0	35.0	40.0	53.8	**85.5**	77.5	**74.5**	**48.0**	40.0	73.2	54.6
DeepSeek-v3	**90.9**	37.2	50.0	9.0	**80.5**	**28.4**	51.5	38.5	36.0	**66.3**	**85.5**	**84.0**	**74.5**	47.2	42.0	**77.8**	**56.4**
Closed-source Models																	
Claude-3.5-haiku-20241022	75.7	36.3	53.5	25.0	48.5	8.0	51.0	31.0	34.0	47.8	43.5	30.0	44.5	36.5	38.6	41.3	38.7
GPT-3.5-turbo	67.4	31.8	50.0	12.5	59.0	21.5	50.0	34.0	30.0	47.3	41.5	46.0	48.5	37.2	38.0	45.8	40.4
Gemini-1.5-flash	84.0	34.5	46.4	28.9	60.0	9.0	53.0	33.5	47.0	59.2	62.0	78.0	69.0	41.1	44.5	67.2	51.1
Gemini-2.0-flash-lite-preview-02-05	84.8	34.5	39.2	**30.5**	84.0	23.5	46.0	20.5	33.5	58.1	78.0	58.5	68.5	50.2	33.3	65.9	51.1
Mistral-large	82.5	39.0	46.4	22.5	58.5	20.5	54.0	40.5	40.5	60.3	80.0	68.0	71.5	42.2	45.0	70.1	52.7
Gemini-1.5-pro	88.6	36.3	42.8	23.5	66.5	9.5	56.4	43.5	39.0	65.2	81.0	89.0	69.5	42.2	46.3	76.4	55.2
Claude-3.5-sonnet-20241022	88.6	37.2	42.8	16.0	75.5	24.0	57.9	40.0	47.5	59.7	82.0	75.0	71.0	46.0	48.5	72.1	55.8
GPT-4o	**90.9**	38.2	46.4	12.0	75.0	35.5	58.5	47.5	41.5	58.2	66.0	81.5	71.5	48.3	49.2	69.5	55.9
GPT-4-turbo	89.4	39.1	42.9	15.0	76.5	35.5	58.5	**49.5**	41.5	59.2	64.0	81.5	72.0	49.1	49.8	69.4	56.3
o1-mini	82.6	37.3	50.0	16.5	75.0	35.0	53.5	44.5	47.0	54.9	84.5	85.5	69.0	47.9	48.3	73.9	57.1
o3-mini	87.9	38.4	53.4	28.0	75.6	29.5	54.5	48.0	**49.0**	59.6	**86.0**	83.5	71.0	49.4	**51.5**	75.8	59.0
Gemini-2.0-flash	86.3	**40.9**	50.0	28.9	**85.0**	59.0	50.5	32.0	39.0	**65.7**	84.0	69.0	**75.5**	**59.6**	40.5	73.7	59.3
Qwen-max-2025-01-25	87.1	36.3	**57.1**	19.0	64.5	**60.0**	**60.0**	37.5	41.5	64.6	81.5	**92.5**	69.0	52.6	46.3	**77.1**	**59.4**

Table 2: Performance of open-source and closed-source models in all tasks of MRCEval. The highest results are denoted in bold respectively.

External knowledge still remains a challenge. Both large and small models perform almost equally poorly in commonsense knowledge and world knowledge comprehension, which indicates that increasing the parameter scales has little effect on the understanding and application of general knowledge. However, as for domain knowledge acquisition and application, LLama-3.3 with 70B parameters performs better than LLama-3.1 with 8B, which demonstrates that larger models have a stronger ability to learn new knowledge.

Large-scale models are good reasoners in MRC. Larger models perform well in reasoning tasks of MRC, even in complicated, more-hop reasoning tasks. Due to the recent research focus on model reasoning (Wei et al., 2022), there has been a greater emphasis on reasoning when training large models, so LLMs are better at reasoning in reading comprehension than the other two aspects, especially the recently released reasoning models like o1-mini, o3-mini, and DeepSeek-R1 series.

5.2 Error Analysis

To assess which aspects are more challenging for all the LLMs, we collect the proportion of samples for each sub-task in which all the models predict incorrectly. We consider six models, LLama-3.3-70B-Instruct, Qwen-2.5-14B-Instruct, Gemini-2.0-flash, GPT-4o, o1-mini and Claude-3.5-sonnet-20241022. As Figure 2, We find that the models in relation and counterfactual tasks have the most common prediction errors, which indicates that these two aspects are the common weaknesses for all models. While in other tasks, such as entity understanding and multi-hop reasoning, models have different agreements, meaning that these aspects are not common weaknesses of all the models.

6 Conclusion

In this work, we propose a novel MRC taxonomy and build a comprehensive, challenging and accessible MRC benchmark based on it. In constructing MRCEval, we employ LLMs as generators for multi-choice sample construction, and judges for challenging samples selection. Extensive studies demonstrate that MRC is still a challenging task for almost all LLMs, especially in relation or event facts understanding and context-faithful. We aim for this work to inspire further advancements in natural language understanding of LLMs.

Limitations

MRCEval is an automated construction comprehensive benchmark, it covers a great number of data from other datasets. While we have taken various factors into account, there are a few limitations. First, since we're building on existing datasets, we do not perform refined manual de-noising, but we did perform automated quality detection filtering on the origin data. Secondly, the process of parsing the answers is not completely rigorous. On the one hand, models will output some non-standard responses to a small number of samples, on the other hand, the business models will refuse to answer some questions due to security, ethics and other factors. We do our best to parse the answers from all the responses, but inevitably a small percentage of the sample fails to parse the answers. After analysis, we found that it only accounted for a small part, so it would not have a great impact on the experimental results.

Ethical Considerations

We address several potential ethical considerations in relation to this work: (1) Intellectual property: This work utilizes several widely adopted MRC datasets, and we fully adhere to their respective licensing agreements. MRCEval will be shared under the CC BY-SA 4.0 license. (2) Intended Use and risk mitigation: The purpose of this work is to present MRCEval, a benchmark designed to evaluate the capabilities of LLMs on MRC tasks. During the sample selection process, we performed sensitive information filtering on the samples that were rejected by GPT-4o-mini. While we cannot completely rule out the possibility of omissions, we trust in the sensitive information filtering capabilities of GPT-4o-mini. (3) AI assistance: GPT-4o was employed to assist in verifying the grammar of the writing.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Danqi Chen. 2018. *Neural reading comprehension and beyond*. Stanford University.

Cohere. 2024. Command r.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 325–332, College Park, Maryland, USA. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

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.

Wendy G Lehnert and WG Lehnert. 1978. *The process of question answering: A computer simulation of cognition*. L. Erlbaum Associates.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

John McCarthy. 1990. An example for natural language understanding and the ai problems it raises. *Formalizing Common Sense: Papers by John McCarthy*, 355:825.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". *arXiv preprint arXiv:2410.03727*.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

OpenAI. 2025. Openai o3-mini.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand. Association for Computational Linguistics.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

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

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. 2024. Morehopqa: More than multi-hop reasoning. *arXiv preprint arXiv:2406.13397*.

Saku Sugawara, Pontus Stenetorp, and Akiko Aizawa. 2021. Benchmarking machine reading comprehension: A psychological perspective. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1592–1612, Online. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Chenhao Wang, Yubo Chen, Zhipeng Xue, Yang Zhou, and Jun Zhao. 2021. Cognet: Bridging linguistic knowledge, world knowledge and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16114–16116.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. MenatQA: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1434–1447, Singapore. Association for Computational Linguistics.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Zijun Yao, Yantao Liu, Xin Lv, Shulin Cao, Jifan Yu, Juanzi Li, and Lei Hou. 2023. KoRC: Knowledge oriented reading comprehension benchmark for deep text understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11689–11707, Toronto, Canada. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

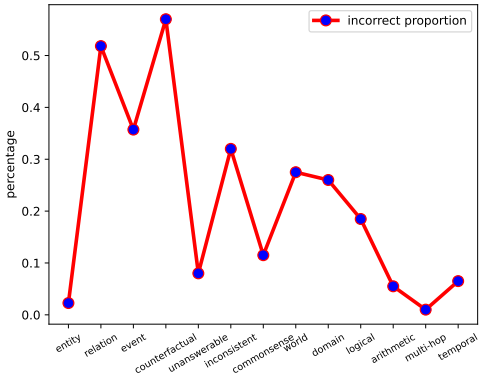

Figure 2: The proportion of incorrect samples for each sub-task.

Appendices

A Experiment Details

Experimental setup. During our experiment, we use 8 NVIDIA RTX 3090 to run open-source models and We access closed-source models by calling their official API. The cost of calling the API during the evaluation is approximately $800 - 1000. All these takes around one month GPU hours in total to complete all experiments.

Model sizes. In this work, we evaluate a total of 31 popular and latest models, including open-source and proprietary models. A summary of model sizes from different model families is shown in Table 3.

B Error Study

As Figure 2, the results of proportion of samples for each sub-task in which all the models predict incorrectly.

Model Name	Size
LLama Family(Dubey et al., 2024)	
LLama-3-8B-Instruct	8B
LLama-3.1-8B-Instruct	8B
LLama-3.3-70B-Instruct	70B
Qwen Family(Yang et al., 2024)	
Qwen-2.5-7B-Instruct	7B
Qwen-2.5-14B-Instruct	14B
Qwen-max-2025-01-25	unknown
Mistral Family(Jiang et al., 2023)	
Mistral-7B-Instruct-v0.3	7B
Mistral-Nemo-Instruct-2407	12B
Mistral-8x7B-Instruct-v0.1	47B
Mistral-large	unknown
Gemma Family(Team et al., 2024)	
Gemma-2-9B-it	9B
Gemma-2-27B-it	27B
Phi Family(Abdin et al., 2024)	
Phi-3-mini-4k-Instruct	3.8B
Phi-3-medium-4k-Instruct	14B
Phi-4	14B
Cohere(Cohere, 2024)	
Command-R-7B-12-2024	7B
DeepSeek(Guo et al., 2025)	
DeepSeek-R1-Distll-LLama-8B	8B
DeepSeek-R1-Distll-Qwen-14B	14B
DDeepSeek-v3	671B
OpenAI	
GPT-3.5-turbo	unknown
GPT-4-turbo	unknown
GPT-4o	unknown
GPT-4o-mini	unknown
o1-mini (low reasoning effort)	unknown
o3-mini (low reasoning effort)	unknown
Gemini	
Gemini-1.5-flash	unknown
Gemini-1.5-pro	unknown
Gemini-2.0-flash	unknown
Gemini-2.0-flash-lite-preview-02-05	unknown
Anthropic	
Claude-3.5-haiku-20241022	unknown
Claude-3.5-sonnet-20241022	unknown

Table 3: Model size across different model families.