Reading Comprehension LLM

Anonymous ACL submission

Abstract

We propose Read-ComprehensionLLM, an intelligent system utilizing large language models (LLMs) to provide outstanding ability in Reading Comprehension. We adopt syllogism prompting strategies to construct supervised fine-tuning datasets in the reading comprehension domain named Read-Comprehension50k, including instruction samples of two categories(Reading comprehension Multiple choice questions andReading comprehension short answer questions) .Besides we use LoRA and QLoRA method to fine-tune LLMs . Evaluations conducted on multiple benchmarks demonstrate that our model performs better than baseline models. Further resources are available at https://huggingface.co/datasets/KashiwaByte/DISC-Assignment

1 Introduction

The recent NLP literature has witnessed a tremendous amount of activity in building models that can follow natural language instructions(Mishra et al(2017)., ; Wei et al(2019); Sanh et al(2019); Wang et al(2018), ; Ouyang et al(2019) ; Chung et al(2017) . These developments are powered by two key components: large pretrained language models (LM) and human-written instruction data (e.g., PromptSource and Super-NaturalInstructions , SuperNI for short).

Prior work on instruction finetune need high memory usage, while a new method named LoRA can reduces memory requirements by using a small set of trainable parameters, often termed adapters, while not updating the full model parameters which remain fixed, based on it, another efficient finetuning approach named QLoRA, can backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters. These two approach would be utilized in out work.

Reading comprehension tasks require thorough reading of the given context and step-by-step analysis to arrive at the correct option. This poses a significant challenge for large models, and enhancing reading comprehension abilities can effectively reduce the illusion of large models. Therefore, it becomes imperative to develop a domain-specific large model with reading comprehension capabilities.

To this end,we present our finetuned large language model tailored for effectively solving reading comprehension problems by analyzing and reasoning. We begin by adopting the reading comprehension syllogism prompting strategy to construct supervised fine-tuning datasets in the reading comprehension domain, named Read-Comprehension50k. These datasets are then employed to train Read-ComprehensionLLM with analyzing and reasoning on top of a general domain LLM with 2B parameters.

In order to evaluate the effectiveness Read-ComprehensionLLM, we utilize multiple evaluation benchmarks and experimental results show that Read-ComprehensionLLM outperforms significantly better than the base foundation model in all downstream tasks,In some domains, it even surpasses ChatGPT3.5, which demonstrates the advantage of our work.

2 Related Work

2.1 Traditional NLP models and Limitations

Traditional NLP models have made progress in various reading comprehension scenarios.

However ,The application of NLP models to the reading comprehension sector presents a unique set of challenges. Firstly, reading comprehension tasks require a thorough understanding of long texts. Secondly, reading comprehension tasks require a certain level of logical reasoning ability in order to derive answers from questions and context.Finally, many NLP models show poor adaptability, being designed for single-task performance and lacking cross-task generalization(Mishra et al., 2022) These challenges underscore the need for future research to develop more robust and adaptable NLP models for the ever-evolving reading comprehension sector.

2.2 Large Language Models for Reading Comprehension

The proposal of LLM-based dialogue systems like ChatGPT OpenAI (2023a), GPT-4 OpenAI(OpenAI et al., 2024), Alpaca Taori et al. (2023) have subverted previous dialogue systems Zhang et al. (2019); Chen et al. (2022b, a). These systems are famous for their zero-shot generalization ability Zhao et al. (2023). One of the key technologies is instruction-tuning Wei et al. (2021). Fine-tuning pre-trained LLM through diverse instruction data to obtain the desired behavior pattern has become a common way to domainize LLM Bao et al. (2023); Yue et al. (2023). However,General domain LLMs reveal serious problems of hallucination by generating irrelevant content to the specific case.

2.3 Methods to enhance performence of Large Language Models

Prompt engineering, which involves designing effective prompts for large language models to guide their responses and improve performance in specific tasks or domains. Domain-specific instruction dataset construction, focusing on creating datasets that are tailored to specific domains or tasks, including task design and data partitioning strategies.

Zero-Shot Prompting is an important innovation in the field of Large Language Models (LLMs). Introduced by Radford et al. (2019), this technique allows us to guide the model to perform new tasks in the absence of large-scale specialized training data by using cleverly designed prompts.

Few-Shot Prompting was proposed by Brown et al. (2020) and, compared to Zero-Shot Prompting, it helps the model learn specific tasks by providing a small number of input-output examples. The paper describes that through carefully selected high-quality examples, the model's performance in executing complex tasks can be significantly improved, especially in cases where no examples are available at all.

To overcome the limitations of Large Language Models (LLMs) in handling complex reasoning tasks, Wei et al(Wei et al., 2023) proposed an innovative approach called CoT. This technique introduces a special prompting strategy aimed at facilitating a more continuous and step-by-step thinking process in the model. In comparison to traditional prompting methods, the primary contribution of CoT lies in its ability to more effectively prompt LLMs to generate structured and deeply considered answers.

SFT (Shanoff et al., 2022) is an instruction fine-tuning dataset that provides explicit guidance for training language models. The primary characteristic of the SFT dataset is its collection in real-world environments, with a focus on instructions that align with human understanding and generation.

Some efficient fine-tuning strategies such as LoRA(Hu et al., 2021) (Layer-wise Adaptive Rate Scaling) and QLoRA(Dettmers et al., 2023) (Quantized Layer-wise Adaptive Rate Scaling), which aim to optimize the fine-tuning process for large language models, can reduces memory requirements by using a small set of trainable parameters

3 Method

3.1 Read-Comprehension50k Datasets

To train Read-ComprehensionLLM, we construct a high-quality supervised fine-tuning dataset, Read-Comprehension50k with two subsets, namely CosmosQA25K and TriviaQA25K. The former aims to enhance the capabilities of LLMs in multiple choice questions and standardized outputs, while the latter seeks to bolster the abilities of LLMs in responding to long-text short answer questions. The core of dataset construction involves creating <instruction, input, output> triplets based on prompt and the content of the original dataset.

Dataset	Samples	Input Token
Multiple choice	25k	230
Short answer	25k	350
Total	50k	300

Table 1: Data statistics of the Read-Comprehension50k dataset.

3.1.1 CosmosQA25k

The CosmosQA(Huang et al., 2019) dataset comprises over 35,000 questions and more than 16,000 article paragraphs, sourced from diverse fields such as Wikipedia, news, fiction, and history. Each question is accompanied by one correct answer and also includes three incorrect answers as distractors.

This design makes Cosmos QA a dataset suited for Multiple-Choice Question Answering (MCQA) tasks.

Utilizing ChatGPT, we designed a prompt and refined it according to the principles of Prompt Engineering to serve as the instruction. We then consolidated the context, question, answer0, answer1, answer2, and answer3 into a single JSON pair as the input. The label of the correct answer is used as the output.

The crafted prompt is as follows: "As a reading comprehension expert, you will receive context, question, and four answer options. Please understand the given context first and then output the label of the correct option as the answer to the question based on the context."

3.1.2 TriviaQA25K

TriviaQA(Joshi et al., 2017) is a challenging reading comprehension dataset that contains over 650,000 question-answer-evidence triplets. TriviaQA includes 95,000 question-answer pairs authored by trivia enthusiasts, accompanied by independently collected evidence documents.However, the original TriviaQA dataset is not suitable for use as an instructional dataset. Therefore, we have reformatted its data structure to align with the format of the Stanford Question Answering Dataset(Rajpurkar et al., 2016) (SQuAD). This adjustment involves structuring the data to better facilitate the development and evaluation of models on question answering tasks.

Leveraging ChatGPT and the principles of Prompt Engineering, we designed a prompt to serve as the instruction for processing the data. We then paired <context, question> as the input and designated the answer as the output.

The crafted prompt is as follows: "As a reading comprehension expert, you will receive context and a question. Please understand the given context first and then output the answer to the question based on the context."

3.2 LLM Finetuning

We utilized MiniCPM-2B(Hu et al., 2024) and ChatGLM3-6B(Du et al., 2022) as the base models for fine-tuning, applying the LoRA method to the former and the QLoRA method to the latter. For both fine-tuning processes, we used a 3090 GPU and leveraged the DeepSpeed framework Rasley et al.(2020). to accelerate training.

3.2.1 LoRA Finetune for MiniCPM-2B

The hyperparameters setting of this training process are as follows: global batch size of 4, learning rate of 1e-4,LoRA rank of 8, dropout parameters of 0.1 , 3 epochs training stage, maximum target length of 512 tokens. The training process was carried out on an 3090 GPU and the training cost is further reduced with the help of deepspeed Rasley et al.(2020).

3.2.2 QLoRA Finetune for ChatGLM3-6B

We used the Xtuner(Contributors, 2023) framework for QLoRA fine-tuning, with the following hyperparameter settings: global batch size of 1, bit quantization of 4,learning rate of 2e-4, LoRA rank of 64, dropout parameters of 0.1,3 epochs training stage, maximum source length of 512 tokens, The training process was carried out on an 3090 GPU and the training utilized deepspeed Rasley et al.(2020).

4 Experiment

4.1 Evaluation Setup

To evaluate the overall performance of the fine-tuned model on reading comprehension tasks, we primarily tested it on two types of tasks: Multiple Choice questions and Short answer questions.

4.1.1 Multiple Choice Questions task

For Multiple-Choice questions, we conducted experiments using the official testing link and test set provided by CosmosQA. We conducted multiple rounds of testing using the methods of CoT, ZeroShot, and FewShot.

4.1.2 Short Answer Questions task

For Short answer questions, we partitioned a 7k SQuAD-formatted TriviaQA dataset for testing,and modified the official validation code to align with our format while retaining core metrics (Exact and F1).

4.2 Main Results

In this section, we present evaluation results of our model on above two tasks in the reading comprehension domain.

4.2.1 Multiple Choice Questions task

We conducted both ablation studies and comparative research on the fine-tuned models to assess their performance and identify the impact of different modifications.

Model	Score
Fewshot MiniCPM	0.3251
Fewshot LoRA MiniCPM	0.7773
Fewshot CoT LoRA MiniCPM	0.7790
CoT LoRA MiniCPM	0.8211
ZH LoRA MiniCPM	0.8215
LoRA MiniCPM	0.8291

Table 2: Ablation Studies

In ablation studies, we observed that the basic MiniCPM-2B model essentially lacks the capability for reading comprehension and selection. Under ZeroShot conditions, it is utterly incapable of completing tasks, and even with FewShot, its performance is only slightly better than pure randomness.

After LoRA fine-tuning, the MiniCPM-2B was able to achieve respectable results, ranking Top 36 on the evaluation leaderboard. When using Chinese prompts, the performance of the LoRA fine-tuned model showed only a minor decrease, reaching Top 42. This demonstrates that MiniCPM-2B possesses multilingual capabilities and can also be applied to Chinese reading comprehension tasks.

It appears that smaller models have lower receptivity to prompts and FewShot learning. Experiments indicate that incorporating FewShot and CoT tends to degrade performance.

Model	Score
ChatGPT3.5	0.7233
LoRA MiniCPM	0.8291
QLoRA Chatglm3	0.8416

Table 3: Comparative Experiments

In the comparative experiments, we contrasted the fine-tuned MiniCPM-2B with fine-tuned ChatGLM3-6B and ChatGPT3.5, The results demonstrated that small-parameter models, after undergoing instruction-based fine-tuning, were able to surpass the performance of ChatGPT3.5 in specific tasks. This highlights the potential of smaller models to achieve competitive results in targeted applications when effectively fine-tuned.

4.2.2 Short Answer Questions task

In the Short Answer questions task, we repeatedly tested eight scenarios including FewShot and ZeroShot LoRA fine-tuned MiniCPM-2B, the original MiniCPM-2B model, QLoRA fine-tuned ChatGLM3-6B, and ChatGLM3-6B itself.

Unfortunately, none of these configurations yielded practically usable results. This outcome suggests that despite the fine-tuning efforts, the models may still face challenges in handling the complexity or specific requirements of short answer question tasks, indicating a need for further model optimization or exploring alternative approaches.

We speculate that the reason for the training failures is that the maxline setting during model fine-tuning was too small, preventing the models from effectively handling reading comprehension tasks with long contexts. This limitation likely restricted the models' ability to process and understand the full scope of the provided texts, thereby impacting their performance on tasks requiring detailed comprehension of lengthy passages. Adjusting the maxline parameter to accommodate longer contexts could potentially improve model performance in future experiments.

Additionally, we tested ChatGPT3.5 and the results indicated that it performed well. On a test set of 1,000 entries, it achieved an Exact Match score of 0.157 and an F1 score of 0.377, with only 73 instances of misunderstanding. This performance highlights the model's effectiveness in grasping and responding to short answer questions, suggesting a robust comprehension capability compared to the earlier tested models.

5 Conclusion

In this paper, we constructed an instruction fine-tuning dataset specifically for the reading comprehension domain and developed two domain-fine-tuned models using LoRA and QLoRA methods. Our evaluation results demonstrate the effectiveness of our models on the Multiple Choice Questions task. Additionally, we identified significant limitations in small-parameter models when dealing with long-text reading comprehension challenges. These findings highlight the importance of model parameter scale in handling complex reading tasks and suggest avenues for further research and optimization in model training strategies.

References

XTuner Contributors. 2023. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning.

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.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and

Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.