Front. Digit. Educ., 2025, 2(2): 23
https://doi.org/10.1007/s44366-025-0060-0

RESEARCH ARTICLE

An Open-Source Large Language Model for Chinese Education Research

Wentao Liu[a], **Hao Hao**[b], **Aimin Zhou**[c]

[a] ECNU's Shanghai Institute for AI Education, East China Normal University, Shanghai 200241, China

[b] Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China

[c] School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

Abstract Open-source large language models (LLMs) research has made significant progress, but most studies predominantly focus on general-purpose English data, which poses challenges for LLM research in Chinese education. To address this, this research first reviewed and synthesized the core technologies of representative open-source LLMs, and designed an advanced 1.5B-parameter LLM tailored for the Chinese education field. Chinese education large language model (CELLM) is trained from scratch, involving two stages, namely, pre-training and instruction fine-tuning. In the pre-training phase, an open-source dataset is utilized for the Chinese education domain. During the instruction fine-tuning stage, the Chinese instruction dataset is developed and open-sourced, comprising over 258,000 data entries. Finally, the results and analysis of CELLM across multiple evaluation datasets are presented, which provides a reference baseline performance for future research. All of the models, data, and codes are open-source to foster community research on LLMs in the Chinese education domain.

Keywords open source, Chinese education large language models, Chinese education research, excelling extension large language model, measuring massive multitask language understanding

1 Introduction

With the introduction of ChatGPT and GPT-4 (OpenAI et al, 2023), research on large language models (LLMs) has garnered significant attention both domestically and internationally. Numerous excellent open-source LLMs have emerged, including Llama 3 (Grattafiori et al., 2024), Qwen2.5 (Qwen Team, 2024a), ChatGLM-4 (Zeng et al., 2024), InternLM2 (Cai et al., 2024), and DeepSeek V2.5 (DeepSeek-AI et al., 2024). These open-source models exhibit two critical characteristics. Firstly, although the final model weights are publicly accessible, the specific training procedures and datasets employed remain undisclosed. For instance, in the study of architectural innovations and algorithmic optimizations of models, Gadre et al. (2024) required transparent training processes to analyze the discrepancies between scaling laws and the actual training and evaluation of language models. Similarly, Kumar et al. (2024) needed transparency in the training process to investigate whether low-precision training and inference impacted the quality and cost of language models. In the realm of safety and ethics of large models, Hu et al. (2024) necessitated transparent data to examine how training data influenced social identity biases. In the investigation of the mechanisms underlying LLMs, Wang et al. (2024) required transparent data to explore whether the remarkable capabilities of LLMs could generalize to unseen tasks or whether they predominantly relied on memorizing vast amounts of pre-training data. These attributes pose significant challenges for advancing research on the application of LLMs (Groeneveld et al., 2024).

Secondly, these open-source models are predominantly tailored for general English-language contexts. The training corpora of existing LLMs are predominantly in English, such as colossal clean crawled corpus (C4) (Pal et al., 2023), Pile (Gao et al., 2020), and BookCorpus (Zhu et al., 2015). Although Chinese corpora, such as WuDaoCorpora 1.0 and THUCNews, have gradually gained attention in LLM training, their scale and diversity remain significantly

Received December 31, 2024; revised February 16, 2025; accepted March 21, 2025

Aimin Zhou (✉)
E-mail: amzhou@cs.ecnu.edu.cn

smaller compared to English corpora. For instance, the total size of WuDaoCorpora's open-source Chinese datasets is 200 GB, whereas the scale of C4's English dataset exceeds 2 TB. Moreover, many open-source LLMs primarily rely on English corpora for training. For example, 89.2% of the training data for Llama 2 is in English. This imbalance results in poorer performance of LLMs in Chinese tasks and sometimes leads to mixed responses of Chinese and English when answering Chinese questions. This poses a challenge to the development of LLMs in Chinese education.

Therefore, an LLM has been designed and developed specifically, which is tailored for the Chinese education domain, named the Chinese education large language model (CELLM). In the initial phase of model development, an in-depth review and comprehensive analysis of the source code of five representative open-source LLMs is conducted. This analysis involves a detailed comparison of the core technologies and methodologies employed in these models. Based on this evaluation, a 1.5 billion parameter LLM is implemented.

The training process for CELLM comprised two distinct stages, pre-training and instruction fine-tuning. During the pre-training stage, this research exclusively utilizes data sourced from the Chinese education domain, as opposed to general-domain English datasets. For the instruction fine-tuning stage, a specialized multiturn dialogue translation framework is developed for LLMs that facilitates the efficient conversion of multi-turn dialogue-based English instruction fine-tuning data into Chinese. The translated data undergoes data cleaning, which results in a collection of 258,000 instruction fine-tuning data entries. Concurrently, a comprehensive set of Chinese instruction fine-tuning datasets is curated, which are integrated with the translated Chinese data to fine tune CELLM's performance further.

Ultimately, this research presents the results and analysis of CELLM across multiple evaluation datasets, including C-Eval (Huang et al., 2024), Chinese massive multitask language understanding (CMMLU) (Li et al., 2023), and massive multitask language understanding (MMLU) (Hendrycks et al., 2021a). These results provide a reference baseline performance for future research.

In addition, representative open-source LLMs are reviewed and the architecture of the proposed CELLM is introduced. Moreover, this research describes the open-source datasets used for training and presents the specialized translation framework. Furthermore, details on the implementation of the model training are provided and the evaluation results of CELLM are presented. Finally, the findings of this study are summarized.

2 Review of Representative Open-Source LLMs

Many open-source LLMs provide only citation formats without accompanying technical papers and detailed reports. Review articles on LLMs (Naveed et al., 2023; Zhao et al., 2023), while introducing these models, often lack a thorough examination of their implementation details. Therefore, five representative open-source LLMs are reviewed and the core technologies are analyzed, including architecture, attention mechanism, positional encode, activation function, and vocabulary size.

Specifically, five prominent open-source models are selected for this research, including Llama 3 (Grattafiori et al., 2024), Qwen2.5 (Qwen Team, 2024a), ChatGLM-4 (Zeng et al., 2024), InternLM2 (Cai et al., 2024), and DeepSeek-V2.5 (DeepSeek-AI, 2024). These models rank highly on the open-source LLM leaderboard and have garnered significant downloads and stars on Hugging Face. Notably, Qwen2.5 serves as a base model for further training of several high-performance LLMs (Qwen Team, 2024a), such as calme-3.2-instruct-78b and QwQ (Qwen Team, 2024b). These models' key technological features are presented in Table 1.

From the architecture aspect as shown in Table 1, almost all LLMs have adopted the architecture of the decoder module introduced by the transformer (Vaswani et al., 2017), along with the autoregressive training approach that sequentially predicts the next

Table 1 Comparison of core technologies in open-source LLMs

LLM	Architecture	Attention mechanism	Positional encode	Activation function	Vocabulary size
Llama 3	Casual-dec	GQA	RoPE	SiLU	128,256
ChatGLM-4	Non-casual-dec	GQA	RoPE	SwiGLU	151,551
Qwen2.5	Casual-dec	GQA	RoPE	SiLU	152,064
InternLM2	Casual-dec	GQA	RoPE	SiLU	92,544
DeepSeek-V2.5	Casual-dec	MHA	RoPE	SiLU	102,400

Notes. LLM: large language model, Casual-dec: casual-decoder, Non-casual-dec: non-casual-decoder, GQA: grouped-query attention, MHA: multi-head attention, RoPE: rotary position encode, SiLU: sigmoid linear unit, SwiGLU: swish gated linear unit.

token. This architecture is referred to as the causal decoder. Among the representative open-source LLMs reviewed in the research, all except ChatGLM-4 employ this architecture (Zeng et al., 2024). While ChatGLM-4 also utilizes the decoder module in its architecture, it differs by adopting a prefix language modelling strategy for pre-training, which can be termed a non-causal decoder. Based on the discussion by Wang et al. (2022), the causal decoder architecture was selected, which demonstrated superior performance during the pre-training phase, for the proposed CELLM model in this research.

From the attention mechanism aspect, most LLMs now utilize grouped-query attention (GQA) rather than multi-head attention (MHA) (Ainslie et al., 2023; Vaswani et al., 2017). The key difference between GQA and MHA lies in the relationship between the number of attention heads corresponding to the number of $\boldsymbol{Q}$ vectors and the number of $\boldsymbol{K}$ and $\boldsymbol{V}$ vectors during the computation of the attention mechanism. When the number of $\boldsymbol{K}$ and $\boldsymbol{V}$ vectors equals the number of attention heads, the model employs the MHA approach. However, MHA requires higher computational costs. It has been proposed that by setting the number of $\boldsymbol{K}$ and $\boldsymbol{V}$ vectors to be more than one but less than the number of query heads, models can achieve performance comparable to MHA while significantly reducing computational complexity (Ainslie et al., 2023). The GQA offers a more efficient alternative to MHA without sacrificing much in terms of performance. To achieve as low a computational cost as possible to facilitate subsequent research, CELLM adopts GQA.

From the positional encoding aspect, all the LLMs utilize rotary positional encoding (RoPE) instead of absolute (Su et al., 2024), relative, and learned positional encoding methods. RoPE is designed to be invariant to rotations in the embedding space, which helps the model better capture the relative positions between tokens. Moreover, RoPE offers flexibility in sequence length, which allows it to be applied to variable-length sequences without modification. This feature is particularly well-suited for the context of LLMs. Therefore, this positional encoding design in CELLM is adopted, which leverages its advantages in capturing relative positions and handles variable-length sequences efficiently.

From the activation function aspect, the data shows that most models adopt sigmoid linear unit (SiLU) (Hendrycks & Gimpel, 2016), whereas swish gated linear unit (SwiGLU) is utilized in ChatGLM-4 (Shazeer, 2020; Zeng et al., 2024). Both activation functions are based on the sigmoid function and have a time complexity of $\boldsymbol{O}(n)$. SwiGLU is a variant of SiLU that introduces a stronger gating mechanism on top of the sigmoid function. Experimental results indicate that SwiGLU exhibits superior performance (Shazeer, 2020). However, the SiLU design is adopted in CELLM, which is used by most models, to ensure more stable model performance.

A small vocabulary size may inadequately represent certain tokens, while an excessively large vocabulary size increases the dimensionality of the input and output layers, which leads to higher computational costs and potentially affects the model's convergence speed. The specific design of the vocabulary is closely tied to the tokenizer. The research's focus is on the internal design of the CELLM model and therefore adopts the tokenizer and vocabulary size used by Qwen2, which is widely utilized in open-source models. It is worth noting that the tokenizer used by Qwen2.5 is similar to that of Qwen2, with the addition of a few special tokens to assist with image input. Since this research does not consider multimodal scenarios, the tokenizer and vocabulary size of Qwen2 are utilized.

Based on the foregoing discussion, the CELLM is designed with key parameters detailed in Table 2.

Table 2 Configuration parameters of the CELLM

Parameter	Value
Vocabulary size	151,936
Layers	64
Hidden size	1,536
Attention heads	12
Key and value heads	4
Max length	2,048

Note. CELLM: Chinese education large language model.

3 Dataset

An overview of the datasets used during the pre-training and instruction fine-tuning stages is provided. Moreover, a specialized multi-turn dialogue translation framework is introduced.

During the pre-training stage, the Chinese-fineweb-edu-v2 dataset is utilized. This dataset is an enhanced version of the original Chinese-fineweb-edu, specifically designed and optimized for natural language processing tasks in the education domain. The Chinese-fineweb-edu-v2 dataset aims to offer researchers and developers a more diverse and broadly applicable corpus of educational resources, thereby facilitating more robust and versatile natural language processing models.

As shown in Table 3, the Chinese-fineweb-edu-v2 primarily consists of four components, including multiple-industry corpus at 25.4%, high-quality and reliable Chinese safety corpus at 18.6%, telecommuni-

Table 3 Datasets and corresponding percentages

Dataset	Percentage (%)
IndustryCorpus2	25.4
CC13	18.6
TeleChat	15.1
ChineseWebText	12.2
WanJuan	11.5
SkyPile	10.3
WuDao	4.7
MiChao	2.1

Notes. IndustryCorpus: multiple-industry corpus, CC13: Chinese safety corpus, TeleChat: telecommunications corpus.

cations corpus at 15.1%, and ChineseWebText at 12.2%. Within the entire dataset, Chinese-language data accounts for 75%.

During the instruction fine-tuning phase, Chinese data is prioritized by collecting a substantial amount of open-source instruction fine-tuning datasets. Then a specialized multi-turn dialogue translation framework is designed and translated over 258,000 English instruction data entries into Chinese.

The framework initially involves designing a prompt template, as shown in Figure 1. The highlighted blue parts in the figure are placeholders that are replaced with corresponding content from the original English instruction data. Although the multi-turn prompt template can also handle single-turn translations, separate prompts are designed for single-turn and multi-turn tasks. This separation is motivated by the observation that single-turn prompts achieve higher success rates when processing single-turn data.

Next, a script is written to automatically convert the text content into JSON data files and perform data cleaning. During the cleaning process, three typical types of errors are identified: First, the unwanted prefixes indicate that the model sometimes prepends unwanted prefixes to the translation results. Second, the incomplete outputs indicate that the model occasionally produces partial translations without completing the full output. Third, format errors indicate that the model's output sometimes has formatting issues that prevent it from being effectively converted into JSON data files.

The string matching and other programmatic methods are implemented to clean the data, which results in the final dataset. Upon comparison, the framework achieved a translation success rate of 97.7% can be extracted. Table 4 provides an overview of the datasets used in the instruction fine-tuning phase, including their datasets, languages, sizes, and brief introductions. Among these, the mathematics aptitude test of heuristics (MATH)-Hard-Chinese and MathInstruct-Chinese are Chinese datasets obtained through our proposed translation framework.

4 Experiments

The training details for CELLM are first provided, followed by its evaluation results on four benchmark datasets, C-Eval (Huang et al., 2024), CMMLU (Li et al., 2023), MMLU (Hendrycks et al., 2021a), and mostly basic Python problems (mbpp) dataset. The performance of CELLM based on these evaluations is analyzed. It is important to note that we have not compared CELLM with other representative open-source models due to significant differences in the quality and quantity of pretraining data. For instance, the pretraining dataset for Qwen2.5 consists of 18 trillion tokens, whereas the proposed model contains only 33 billion tokens.

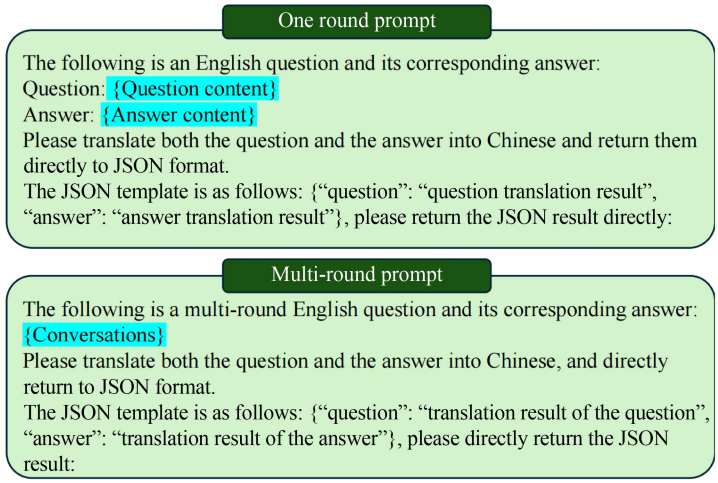

Figure 1 Prompt used in translation framework.

Table 4 Overview of datasets used in the instruction fine-tuning phase

Dataset	Language	Size ($\times 10^3$)	Introduction of datasets in fine-tuning phase
MATH (Hendrycks et al., 2021b)	English	7.5	The MATH dataset consists of problems from mathematics competitions, including the AMC 10, AMC 12, AIME, and more.
Blossom-v3	English or Chinese	5.0	Blossom-v3 is a bilingual dialogue dataset derived from ShareGPT 90,000, specifically designed for multi-turn conversation fine-tuning.
Blossom-wizard-v3	English or Chinese	20.0	Blossom-wizard-v3 is a bilingual instruction dataset derived from WizardLM Evol-Instruct V2, specifically designed for instruction fine-tuning.
Blossom-orca-v3	English or Chinese	20.0	Blossom-orca-v3 is a bilingual instruction dataset derived from OpenOrca, specifically designed for instruction fine-tuning.
Infinity-instruct-C	Chinese	750.0	Infinity-instruct-C collects all Chinese data from Infinity Instruct.
GSM8K (Cobbe et al., 2021)	English	8.8	GSM8K is a dataset of 8,500 high-quality linguistically diverse grade school math word problems.
Evol-Instruct-C	Chinese	50.0	Evol-Instruct-C is a Chinese version of Evol-Instruct (Xu et al., 2023).
Magpie-ultra (Xu et al., 2023)	English	50.0	Magpie-ultra is a synthetically generated dataset using the Llama 3.1 405B-Instruct.
Magpie-Qwen2-Pro-200k-Chinese	Chinese	200.0	Magpie-Qwen2-Pro-200k-Chinese is a synthetically generated Chinese dataset using the Qwen2.
Magpie-Pro-300k-Filtered	English	300.0	Magpie-Pro-300k-Filtered is a synthetically generated dataset using the Llama 3.1 70B-Instruct.
Magpie-MT-Filtered	English	300.0	Magpie-MT-Filtered is a synthetical multi-turn conversation dataset using the Llama 3.1 70B-Instruct.
MATH-Hard-Chinese	Chinese	2.2	MATH-Hard-Chinese is a Chinese translation version of the MATH dataset (Hendrycks et al., 2021b), translated using the proposed framework.
MathInstruct-Chinese	Chinese	256.0	MathInstruct-Chinese is a Chinese translation version of the MAmmoTH (Yue et al., 2024), translated using the proposed framework.

Instead, the research focuses on developing an LLM with full transparency in the training process and data, emphasizing Chinese educational datasets. Meanwhile, more adequate pretraining and instruction fine-tuning in future work can be conducted, aiming to provide a robust comparison with state-of-the-art models of similar parameter scales.

4.1 | Training Details

The training of CELLM consists of two phases, both conducted on an H100 8-graphics processing unit (GPU) machine. To accelerate training, this research utilizes the DeepSpeed distributed framework with the ZerO-2 optimization level (Rasley et al., 2020), and all data formats are set to bfloat16. For both training phases, a cosine decay learning rate schedule is adopted with an initial value of 3×10^{-4} and a warmup ratio of 0.01. The optimizer AdamW is used, configured with a weight decay of 0.10 and an Adam beta2 of 0.95.

In the pretraining phase, 33.6 billion tokens are randomly sampled from the Chinese-fineweb-edu-v2 dataset for training. Specifically, 2,000 training steps are performed, each with a batch size of $8 \times 16 \times 64 = 8{,}192$, where 8 is the number of GPUs, 16 is the batch size per GPU, and 64 is the gradient accumulation steps. The model's training window is set to 2,048 tokens. Therefore, the total number of tokens trained in the pretraining phase was $2{,}000 \times 8{,}192 \times 2{,}048$, approximately equal to 33.6 billion tokens. The total training time for this phase is approximately 106 hours.

In the instruction fine-tuning phase, the model is trained using the datasets, completing a total of 3 epochs, which amounts to approximately 16 billion tokens. The batch size and training window size remain consistent with those in the pretraining phase. The total training time for this phase is approximately 51 hours.

4.2 | Experiment Results

The comprehensive experimental results of CELLM are presented on several benchmarks, including C-Eval (Huang et al., 2024), CMMLU (Li et al., 2023), and MMLU (Hendrycks et al., 2021a). These benchmarks evaluate the model's performance across multiple subjects from elementary school to university level. This setup allows us to assess CELLM's capabilities over a wide range of educational domains, ensuring a thorough evaluation of its effectiveness in various academic contexts. Table 5 summarizes the results of CELLM across various test sets.

CELLM performs relatively weaker on science, technology, engineering, and mathematics (STEM)-related test cases. Specifically, the accuracy on C-Eval-stem is 21.48%, compared to the overall average of 23.95% on C-Eval. On CMMLU-stem, it is 22.55%, compared to the overall average of 24.59% on CMMLU. On MMLU-stem, it is 21.66%, compared to the overall

Table 5 Comprehensive performance of CELLM on AGI-Eval, C-Eval, CMMLU, and MMLU

Case	Metric	Average value
AGI-Eval-Chinese	Average	14.19
AGI-Eval-English	Average	17.53
AGI-Eval-gaokao	Average	14.65
AGI-Eval	Average	15.62
C-Eval-stem	Average	21.48
C-Eval-social-science	Average	26.35
C-Eval-humanities	Average	26.77
C-Eval-other	Average	23.43
C-Eval-hard	Average	18.40
C-Eval	Average	23.95
CMMLU-humanities	Average	26.26
CMMLU-stem	Average	22.55
CMMLU-social-science	Average	24.91
CMMLU-other	Average	24.97
CMMLU-China-specific	Average	26.19
CMMLU	average	24.49
MMLU-humanities	Average	24.29
MMLU-stem	Average	21.66
MMLU-social-science	Average	24.75
MMLU-other	Average	23.80
MMLU	Average	23.40
MMLU-weighted	Weighted-average	23.91

Notes. AGI: artificial general intelligence, CELLM: excel extension large language model, CMMLU: Chinese massive multitask language understanding, MMLU: massive multitask language understanding.

average of 23.40%. Conversely, CELLM demonstrates superior performance in the humanities and social sciences. For instance, the accuracy on C-Eval-social-science and C-Eval-humanities exceeds 26.00%, higher than the overall average of 23.95% on C-Eval. These results indicate that CELLM exhibits relatively lower performance in STEM disciplines, while performing better in literature, humanities, and social sciences. This experimental phenomenon is also commonly observed in other LLMs. The more detailed experimental results for each specific subject in C-Eval (Huang et al., 2024), CMMLU (Li et al., 2023), and MMLU are presented in Tables S1 (Hendrycks et al., 2021), S2, and S3 in the Electronic Supplementary Material (ESM).

The results of CELLM are further presented on the mbpp dataset, which is used to evaluate the model's programming capabilities. The mbpp dataset comprises a total of 500 test problems. It is developed by researchers including those from Google, and serves as a benchmark for assessing and enhancing the performance of programming models on fundamental Python programming tasks.

Referring to Table 6, the findings indicate that CELLM successfully generates and executes code for 3 of these test cases. For the remaining 497 problems, CELLM fails, 3 due to timeout during execution, 472 because the code does not run successfully, and 22 where the output is incorrect. The overall score is 0.6. This suggests that while CELLM demonstrates some capability in generating functional Python code, its overall performance on the mbpp benchmark is limited.

Table 6 Performance of CELLM in mbpp

Metric	Number
Pass	3
Timeout	3
Failed	472
Wrong answer	22

In the ESM, the additional experimental results for CELLM are provided, serving as a performance baseline for future research. These include the results of traditional natural language tasks evaluated on the Chinese language understanding evaluation dataset in Table S4 in ESM, the results of content moderation tasks for edit evaluation on the Wikibench dataset in Table S5 in ESM, the results of scientific question understanding and reasoning tasks on the AI2 reasoning challenge dataset in Table S6 in ESM, the results of Chinese short-text summarization tasks on the LCSTS dataset in Table S7 in ESM, the results of college entrance examination tasks on the gaokao dataset in Table S8 in ESM, and the results of human-centred standardized tests and competitions on the AGI-Eval dataset in Table S9 in ESM.

5 Conclusions

This research presents the development of an LLM specifically tailored for the Chinese education domain, trained from scratch. This research began by reviewing and referencing the architectures of five representative open-source models, which informed the design of the CELLM model architecture. Then the pretraining and instruction fine-tuning datasets are conducted and a multi-round translation framework is proposed to translate 258,000 instruction fine-tuning data points into Chinese.

Detailed training procedures are provided and comprehensive evaluation results of CELLM on 11 benchmark datasets are presented, establishing a reference baseline performance for future research. Moving forward, this research plans to continue releasing more comprehensive training checkpoints of CELLM, along with the results after alignment fine-tuning. The goal is to facilitate and advance research in

the Chinese education domain through the development and availability of CELLM.

Conflict of Interest The authors declare that they have no conflict of interest.

Ethics Statement The authors declare that their Institutional Ethics Committee confirmed that no ethical review was required for this study. Written informed consent for participation was not required because all participants' data was anonymized before the statistical analyses were conducted.

Data Availability Statements The authors confirm that all data generated or analyzed during this study are included in this published article.

Authors Contributions Wentao Liu performed all the experiments and contributed to the writing of parts of the manuscript. Hao Hao contributed to the writing of parts of the manuscript. Aimin Zhou planned the research content and supervised the overall project. All authors approved the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s44366-025-0060-0 and is accessible for authorized users.

References

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023). GQA: Training generalized multi-query transformer models from multi-head checkpoints. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: EMNLP.

Cai, Z., Cao, M. S., Chen, H. J., Chen, K., Chen, K. Y., Chen, X., Chen, X., Chen, Z. H., Chen, Z., Chu, P., & et al. (2024). InternLM2 technical report. *arXiv Preprint*, arXiv:2403.17297.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., & et al. (2021). Training verifiers to solve math word problems. *arXiv Preprint*, arXiv:2110.14168.

DeepSeek-AI, Liu, A. X., Feng, B., Wang, B., Wang, B. X., Liu, B., Zhao, C. G., Dengr, C., Ruan, C., Dai, D., & et al. (2024). DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv Preprint*, arXiv:2405.04434.

Gadre, S. Y., Smyrnis, G., Shankar, V., Gururangan, S., Wortsman, M., Shao, R. L., Mercat, J., Fang, A., Li, J., Keh, S., & et al. (2024). Language models scale reliably with over-training and on downstream tasks. *arXiv Preprint*, arXiv:2403.08540.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., & et al. (2020). The pile: An 800 GB dataset of diverse text for language modeling. *arXiv Preprint*, arXiv:2101.00027.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., & et al. (2024). The Llama 3 herd of models. *arXiv Preprint*, arXiv:2407.21783.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y. Z., & et al. (2024). OLMo: Accelerating the science of language models. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Bangkok: Association for Computational Linguistics, 15789–15809.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021a). Measuring massive multitask language understanding. In: *Proceedings of the 9th International Conference on Learning Representations*. Austria: ICLR.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021b). Measuring mathematical problem solving with the MATH dataset. *arXiv Preprint*, arXiv:2103.03874.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv Preprint*, arXiv:1606.08415.

Hu, T. C., Kyrychenko, Y., Rathje, S., Collier, N., Van Der Linden, S., & Roozenbeek, J. (2024). Generative language models exhibit social identity biases. *Nature Computational Science*, *5*(1), 65–75.

Huang, Y. Z., Bai, Y. Z., Zhu, Z. H., Zhang, J. L., Zhang, J. H., Su, T. J., Liu, J. T., Lv, C. C., Lei, J. Y., Fu, Y., & et al. (2024). C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 2749.

Kumar, T., Ankner, Z., Spector, B. F., Bordelon, B., Muennighoff, N., Paul, M., Pehlevan, C., Ré, C., & Raghunathan, A. (2024). Scaling laws for precision. *arXiv Preprint*, arXiv:2411.04330.

Li, H. N., Zhang, Y. X., Koto, F., Yang, Y. F., Zhao, H., Gong, Y. Y., Duan, N., & Baldwin, T. (2023). CMMLU: Measuring massive multitask language understanding in Chinese. In: *Proceedings of the Association for Computational Linguistics: ACL 2024*. Bangkok: Association for Computational Linguistics, 11260–11285.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv Preprint*, arXiv:2307.06435.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & et al. (2023). GPT-4 technical report. *arXiv Preprint*, arXiv:2407.21783.

Pal, K. K., Kashihara, K., Anantheswaran, U., Kuznia, K. C., Jagtap, S., & Baral, C. (2023). Exploring the limits of transfer learning with unified model in the cybersecurity domain. *arXiv Preprint*, arXiv:2302.10346.

Qwen Team. (2024a, September 19). *Qwen2.5: A party of foundation models*! Available from Qwen website.

Qwen Team. (2024b, November 28). *QwQ: Reflect deeply on the*

boundaries of the unknown. Available from Qwen website.

Rasley, J., Rajbhandari, S., Ruwase, O., & He, Y. X. (2020). DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In: *Proceddings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York: Association for Computing Machinery, 3505–3506.

Shazeer, N. (2020). GLU variants improve transformer. *arXiv Preprint*, arXiv:2002.05202.

Su, J. L., Ahmed, M., Lu, Y., Pan, S. F., Bo, W., & Liu, Y. F. (2024). RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, *568*, 127063.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 6000–6010.

Wang, T., Roberts, A., Hesslow, D., Le Scao, T., Chung, H. W., Beltagy, I., Launay, J., & Raffel, C. (2022). What language model architecture and pretraining objective works best for zero-shot generalization? In: *Proceedings of the 39th International Conference on Machine Learning*. Baltimore: PMLR, 22964–22984.

Wang, X. Y., Antoniades, A., Elazar, Y., Amayuelas, A., Albalak, A., Zhang, K. X., & Wang, W. Y. (2024). Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. *arXiv Preprint*, arXiv:2407.14985.

Xu, C., Sun, Q. F., Zheng, K., Geng, X. B., Zhao, P., Feng, J. Z., Tao, C. Y., & Jiang, D. X. (2023). WizardLM: Empowering large language models to follow complex instructions. *arXiv Preprint*, arXiv:2304.12244.

Yue, X., Qu, X. W., Zhang, G., Fu, Y., Huang, W. H., Sun, H., Su, Y., & Chen, W. H. (2024). MAmmoTH: Building MATH generalist models through hybrid instruction tuning. In: *Proceedings of the Twelfth International Conference on Learning Representations*. Vienna: ICLR.

Zhao, W. X., Zhou, K., Li, J. Y., Tang, T. Y., Wang, X. L., Hou, Y. P., Min, Y. Q., Zhang, B. C., Zhang, J. J., Dong, Z. C., & et al. (2023). A survey of large language models. *arXiv Preprint*, arXiv:2303.18223.

Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., Lai, H., & et al. (2024). ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv Preprint*, arXiv:2406.12793.

Zhu, Y. K., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision*. Washington: IEEE, 19–27.