ISSN: 2424-8460(O)
2251-2608(P)

Original Research Article

A comparative study of model architectures on multilingual comprehension capabilities

Yangchen Ji

Sinopec Research Institute of Petroleum Processing, Liaocheng, Shandong, 100083, China

Abstract: This study investigates the differences in Chinese and English comprehension capabilities between the open-source large language models Llama3-8B (Meta) and GLM-4-9B (Zhipu AI) using a controlled variable approach. Employing a unified BPE tokenizer and cleaned monolingual corpora (850M tokens for Chinese, 1.1B tokens for English), both models were pretrained under identical hyperparameters (learning rate: 3e-5, batch size: 32, epochs: 10). Performance was evaluated primarily using the F1 score on text classification tasks. Results indicate that GLM-4-9B significantly outperformed Llama3-8B on Chinese tasks (F1: 92.3% vs. 89.7%), while both models exhibited comparable performance on English tasks (F1: 94.1% vs. 93.8%). This suggests that the Autoregressive Blank Infilling objective in GLM's architecture may be better suited to the syntactic characteristics of Chinese. The study notes that limitations in dataset scale and hyperparameter optimization depth warrant further validation of these conclusions.

Keywords: Large language models; Monolingual training; Model evaluation

1. Introduction

The widespread adoption of the Transformer architecture[1] has led to high-performance LLMs like ChatGPT and GLM. Despite shared underlying principles, design variations—including tokenization strategies (e.g., BPE vs. WordPiece), attention mechanisms (Grouped-Query Attention vs. standard Multi-Head Attention), and training objectives (autoregressive prediction vs. blank infilling)—may impact multilingual comprehension. Prior research indicates that tokenizer coverage significantly affects performance on low-resource languages[2], yet implicit biases for major languages (e.g., Chinese/English) remain underexplored, with studies often focusing on cross-lingual transfer[3]. This work hypothesizes that differences in tokenization and training objectives lead to divergent comprehension capabilities across languages with distinct grammatical structures. To test this, we select Llama3-8B and GLM-4-9B—open-source models of comparable parameter scale—and quantify their multilingual performance under controlled training data and evaluation tasks.

2. Research methodology

2.1. Model Selection and comparison

Commercial LLMs are typically closed-source, limiting reproducibility. Thus, this study prioritizes open-source alternatives.

Llama3-8B[4] is an 8-billion parameter, instruction-tuned generative text model from Meta's Llama 3 series. Optimized for dialogue, it outperforms many open-source chat models on standard benchmarks. Technically, it employs an optimized Transformer architecture with an autoregressive objective, a 128K-token tokenizer, 8K-token sequence training, and Grouped-Query Attention (GQA)[5] for inference efficiency.

GLM-4-9B[6] is the open-source variant of Zhipu AI's GLM-4 series, pretrained on 10 trillion tokens (pri-

ISSN: 2424-8460(O)
2251-2608(P)

marily Chinese/English, plus 24 other languages). It supports 26 languages and features architectural improvements, including an Autoregressive Blank Infilling objective[7], multi-stage post-training, and optimized design.

Both models have similar parameter counts, release timelines, and accessible documentation (GitHub/HuggingFace), making them suitable for controlled comparison.

2.2. Dataset selection and processing

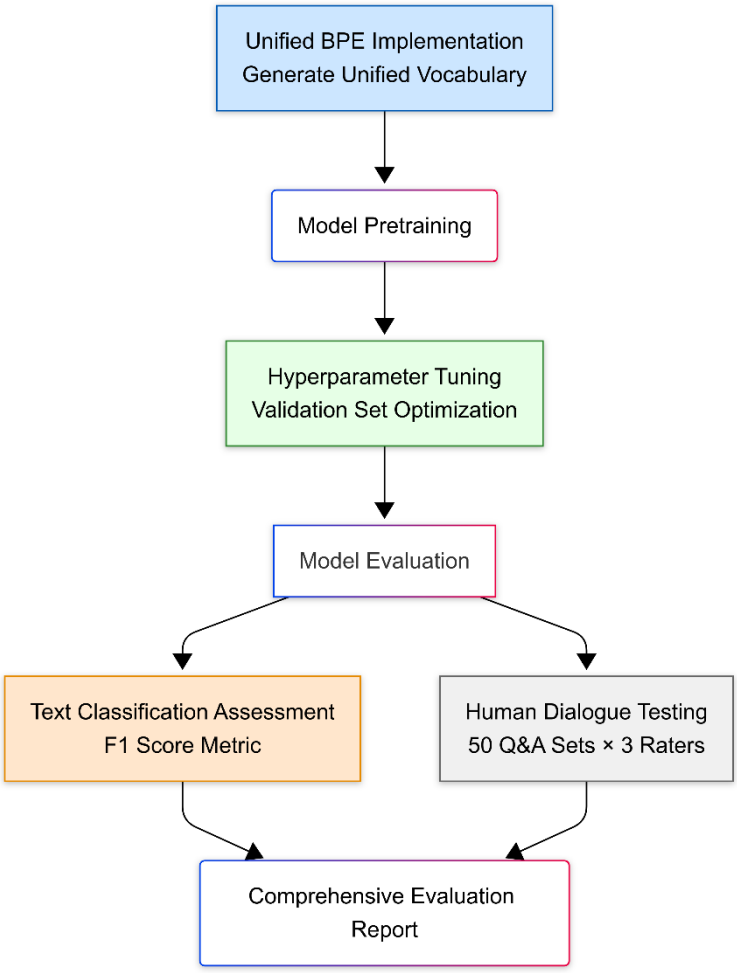
Datasets were sourced from HuggingFace: ClueCorpus (Chinese) and C4 (English). To mitigate noise from low-quality data, rigorous cleaning was applied, including Simplified-Traditional conversion (OpenCC), case normalization, and regex-based removal of non-standard symbols. Post-cleaning, the Chinese corpus comprised ~7 GB (850M tokens), and the English corpus ~15 GB (1.1B tokens). Cleaned data was stratified-sampled (7:2:1) into training, validation, and test sets, preserving class distribution.

2.3. Experimental design

To isolate architectural effects, both models used a unified BPE tokenizer[8]. A custom tokenizer was trained using SentencePiece[9] and merged with cl100k_base.

During pretraining, hyperparameters were tuned via validation-set performance. Optimal settings (learning rate: 3e-5, batch size: 32, dropout: 0.1) were adopted uniformly. Evaluation involved text classification (primary metric: F1 score) and human dialogue assessment (50 Q&A pairs rated by 3 annotators on a 5-point scale for coherence).

ISSN: 2424-8460(O)
2251-2608(P)



2.4. Research platform

Training/validation utilized an NVIDIA RTX 4090 GPU (24 GB VRAM), sufficient for both models.

3. Results

Table 1. Test results.

Model	Chinese F1 (%)	English F1 (%)	Dialogue Score (Avg.)
Llama3-8B	89.7±0.4	93.8±0.3	4.2
GLM-4-9B	92.3±0.3	94.1±0.2	4.5

Results (**Table 1**) show GLM-4-9B's significant advantage in Chinese comprehension, attributed to its blank infilling objective aligning better with Chinese syntactic features (short sentences, high context-dependency). Performance parity on English tasks likely stems from larger English corpora and convergent model generalization.

4. Conclusion

GLM-4-9B demonstrates superior Chinese comprehension, while both models perform similarly on English. Limitations include constrained dataset size, suboptimal hyperparameter tuning, and focus on high-resource languages. Future work should extend to low-resource languages (e.g., Thai, Swahili) and multimodal inputs to

explore syntax-semantics interactions.

About the author

JI Yangchen (born October 1996), male, Han nationality, from Liaocheng, Shandong Province. He is a master's graduate student at the Petrochemical Processing Research Institute of Sinopec, specializing in research within the field of applied chemistry.

References

[1] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. arXiv August 2, 2023. https://doi.org/10.48550/arXiv.1706.03762.

[2] Limisiewicz, T.; Balhar, J.; Mareček, D. Tokenization Impacts Multilingual Language Modeling: Assessing Vocabulary Allocation and Overlap Across Languages. arXiv May 26, 2023. https://doi.org/10.48550/arXiv.2305.17179.

[3] Lample, G.; Conneau, A. Cross-Lingual Language Model Pretraining. arXiv January 22, 2019. https://doi.org/10.48550/arXiv.1901.07291.

[4] AI@Meta. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[5] Ainslie, J.; Lee-Thorp, J.; Jong, M. de; Zemlyanskiy, Y.; Lebrón, F.; Sanghai, S. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. arXiv December 23, 2023. https://doi.org/10.48550/arXiv.2305.13245.

[6] GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Sun, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; Wang, Z. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv July 30, 2024. https://doi.org/10.48550/arXiv.2406.12793.

[7] Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; Tang, J. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. arXiv March 17, 2022. https://doi.org/10.48550/arXiv.2103.10360.

[8] Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. arXiv June 10, 2016. https://doi.org/10.48550/arXiv.1508.07909.

[9] Kudo, T.; Richardson, J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. arXiv August 19, 2018. https://doi.org/10.48550/arXiv.1808.06226.