Simorgh at SemEval-2026 task 7: Region-Aware Hybrid Retrieval for Low-Resource Cultural Reasoning in Multilingual Question Answering

Hadi Bayrami Asl Tekanlou[1] **Mahdi Bakhtiyarzadeh**[2] **Jafar Razmara**[1]
[1,2]University of Tabriz, Tabriz, Iran
h.bayrami1403@ms.tabrizu.ac.ir
m.bakhtiyarzadeh1403@ms.tabrizu.ac.ir
razmara@tabrizu.ac.ir

Abstract

Although Large Language Models (LLMs) demonstrate excellent capabilities and performance for general reasoning tasks within the general public domain, they may face challenges with culturally grounded knowledge within languages with limited digital and textual data. In this paper, we investigate culturally grounded multiple-choice question answering with the BLEnD benchmark, which consists of a multilingual corpus of 30 languages and covers various socio-cultural domains, such as cuisine, sports, family, etc. We propose a region-aware hybrid retrieval approach that combines BM25 lexical matching and dense semantic similarity with regional weighting heuristics to improve the relevance of the answer. The retrieved documents are used to construct a structured prompt for the Qwen3-14B quantized model with logit-based deterministic answer selection. The experimental results show improvements to cross-lingual stability with the hybrid retrieval approach over pure parametric inference for culturally grounded question answering. However, there are still notable performance gaps between languages with more and less training data. This shows that the limitations of the retrieval augmentation approach are not entirely overcome by the training data imbalance problem.

1 Introduction

Large Language Models (LLMs) have shown impressive capabilities in delivering outstanding results across a broad range of natural language processing tasks (Schick et al., 2023; Brown et al., 2020). When asked to answer questions based on Western culture, their answers are highly accurate. However, if they are asked to answer questions related to everyday cultural knowledge from an Eastern background or other less-represented regions, their performance is not as satisfactory (Tao et al., 2024; Naous et al., 2024). If these models are asked questions related to everyday cultural knowledge, their limitations become quite apparent. For example, if they are asked to answer questions like *"What do Australians need to do to prevent fires in summer?"* their limitations become quite apparent. The reason behind this is that these models are trained on large-scale corpora like Wikipedia and Common Crawl, which are dominated by information from specific regions, languages, and perspectives. The distribution of culturally specific everyday knowledge in these models is inherently unbalanced (Navigli et al., 2023; Naous et al., 2024). As a result, these models are likely to produce answers that are incomplete, too general, or even incorrect (McIntosh et al., 2024). The unbalanced distribution of culturally specific everyday knowledge in these models can result in hallucinations, leading to stereotypical answers (Myung et al., 2025). The limitations of these models in question-answering systems need to be evaluated in relation to everyday knowledge.

This research focuses on the answering of multiple-choice questions culturally based with large language models. The following are the main items we will accomplish:

- Create a hybrid retrieval method that incorporates BM25 (what the BM25 algorithm does is attempts to match relevant documents to the query) matched to the query lexically and matched semantically .
- Implement a region-based enhancement on the retrieval process (where we prioritize evidence that is culturally relevant to the question).
- Utilize retrieval based augmentation to add culturally appropriate evidence retrieved to

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 624–629

the structured prompt for culturally situated reasoning.

- Apply a logit-based approach for determining the answer choice (A-D) so we can provide deterministic and efficient selection of multiple-choice answers.

2 Related Work

2.1 Cultural Commonsense Knowledge in Large Language Models

Even though Large Language Models (LLMs) perform well on general types of commonsense reasoning, they do not perform as well when evaluated based on culture-specific commonsense reasoning. The differences between cultures often result in very different performance on the general commonsense tasks provided to LLM's. For example, while both general commonsense and cultural context influence performance of LLMs on commonsense reasoning, the language used in the query modifies the accuracy of the LLM on tasks related to only culture (Shen et al., 2024). These differences in performance indicate that LLM's exhibit built-in biases towards a particular culture, resulting from an unbalanced amount of training data that is heavily weighted toward dominant cultural representations as well as completely monolingual, English-language datasets (Navigli et al., 2023). The same types of biases exist with respect to moral values. For example, monolingual English LLMs do not adequately account for the fine-grained differences between countries with respect to moral concepts such as homosexuality or divorce, nor do they have a good grasp of the overall patterns of global moral diversity found in datasets such as the World Values Survey or PEW Surveys. The fine-tuning of LLMs using representative data will increase the accuracy of moral inferences across multiple countries but decrease accuracy with respect to norms within the United States (Ramezani and Xu, 2023).

2.2 Performance on Eastern and Region-Specific Cultural Knowledge

Evaluations of LLMs (Large Language Models) confirm that they perform less well with everyday knowledge of Eastern cultures. In a case study, the Korean medical licensing examination (Korean National Licensing Examination for Korean Medical Doctors, K-NLEKMD) had a 66.18% successful score for GPT-4. While this was above the cutoff of 60%, LLMs lost accuracy on localized topics (e.g. public health law, internal medicine, acupuncture) when compared to non-localized topics (compared TKM-specialized versus non-TKM-specialized).Improvements in prompting (e.g., annotating the prompts with Chinese terms, self-consistency) were an important factor in the LLM's successful score; reflecting how LLMs continue to encounter challenges when adapting to other languages and cultures (Jang et al., 2023). n the context of Mandarin Chinese, ChatGPT-3.5 had a passing score of 153.5/300 in the Postgraduate Examination for Clinical Medicine (passed at the 20th percentile), but ChatGPT-3.5's performance on the open-ended exam questions (31.5% accuracy) failed; results were 42% in common, 37% in multi-choice, and 17% in case analysis, although there was a 90% average agreement and high insight from the generated results. These results suggest that the recall/diagnosis performance of the LLM was better than that of the intervention-based performance and suggest that the LLM faces challenges adapting to other languages (Yu et al., 2024). Research comparing English to Chinese has shown vast differences between these two languages as well as how LLM's are challenged greatly with Chinese language formality and structure, or grammar, due to cultural differences (Wang et al., 2024).

2.3 Bias Origins and Mitigation Strategies

selection bias of the training data, I.e., it may reflect a variety of social biases (e.g., those related to gender, age, ethnicity, religion, and language) (Navigli et al., 2023). An imbalanced corpus can also contribute to under-representation of low-resource languages and cultures. Examples of initiatives intended to mitigate this problem include NusaWrites, which creates a culturally relevant corpus using paragraphs written by native speakers in each of the 12 low-resource languages spoken in Indonesia, showing more lexical diversity than scraping or translation methods of obtaining data (Cahyawijaya et al., 2023). LLMs present opportunities for researchers in the humanities to use low-resource languages for research; however, researchers face difficulties due to the scarcity of data, adaptability, and cultural sensitivity of the data collected. Researchers require a customized model and collaboration with interdisciplinary teams in order to overcome these obstacles (Zhong et al., 2026).

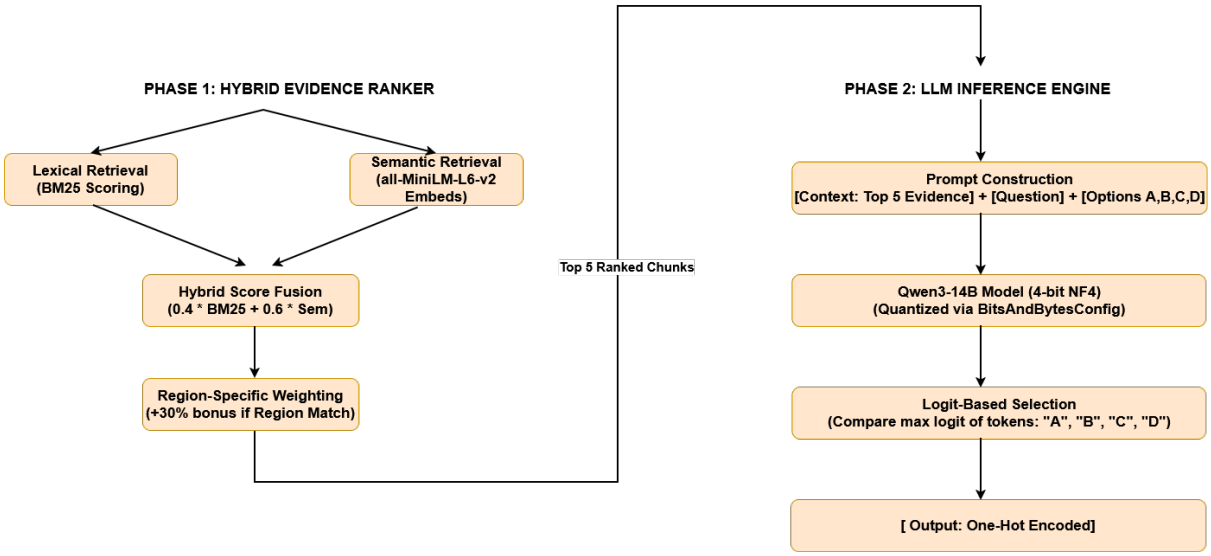

Figure 1: Overview of the proposed system architecture.

3 Methodology

We designed a two step pipeline for the BLEnD dataset consisting of:

- a Hybrid Evidence Ranker
- a Quantized LLM Inference Engine

to address both the cultural rationale of the dataset and the Multiple Choice Question (MCQ) objectives. The overall system architecture is illustrated in Figure 1. It consists of a hybrid evidence ranking module followed by a quantized LLM-based inference engine for deterministic answer selection.

3.1 Hybrid Evidence Ranking

To determine cultural context relevance, our retrieval system provides a combination of traditional lexical matching and dense semantic similarity.

- Using the standard BM25 algorithm (Robertson and Zaragoza, 2009) with word-boundary tokenization, lexical matching captures exact matches of keywords between the cultural question and document evidence.
- To define contextual nuance and semantic meaning, dense embeddings were produced using the *all-MiniLM-L6-v2* SentenceTransformer model and the cosine similarity between the question and the picture evidence was calculated.
- BM25 and semantic scores were combined to obtain an overall score (40% BM25, 60% semantic). For further refinement, a heuristic of 30% weight was added to the combined score if there was a regional reference in the evidence evidence text. The final ranking score for each document is calculated according. $\text{FinalScore} = (0.4 \cdot \text{BM25} + 0.6 \cdot \text{Semantic}) \cdot (1 + \text{RegionBonus})$
- Output: The documents are sorted based on this final score, and the top 5 highest-ranking evidence chunks are extracted and passed to the inference stage.

3.2 LLM-Based Cultural Reasoning

For the MCQ selection, we utilized a large language model to synthesize the retrieved evidence and determine the correct answer.

- Modeling Setup: The Qwen3-14B (Team, 2025) model was utilized. To reduce VRAM usage and efficiently batch process models, we loaded the model with 4-bit NormalFloat (nf4) quantization with double quant and using float16 for compute.
- Prompt Construction: The model is prompted as a "cultural reasoning model." The model's 5 best documents are inserted directly above the question and choices in the context window, with choices labelled A, B, C and D. If there is no evidence returned, the prompt defaults to using the model's parametric knowledge.

- Prediction Selection based on Logits: Instead of using generated free text, which can have a lot of formatting errors or be very wordy, we used a very strict logarithmic scoring system. The next-token prediction logit values are extracted only for the four letters A, B, C, and D. The choice with the highest probability mass (highest logit) is selected as the final choice.

- Prediction Processing and Exporting: Predictions are processed in batches of 16 for maximum use of GPU. The final predictions will be converted to one-hot encoding format for the four choices.

4 Results

4.1 Dataset

For evaluations, we use the BLEnD benchmark (Myung et al., 2024), developed as part of SemEval-2026 Task 7 (Ousidhoum et al., 2026), which focuses on evaluating everyday knowledge across diverse languages and cultural contexts. It comprises approximately 52,600 question–answer pairs spanning 16 diverse countries/regions and 30 languages, including several low-resource languages such as Amharic, Assamese, Azerbaijani, Hausa, and Sundanese. This diversity ensures a rigorous test of a model's ability to reason across both Western-centric and underrepresented cultural contexts. The questions are hand-crafted by native speakers and categorized into six core sociocultural domains: Food, Sports, Family, Education, Holidays & Leisure, and Work-Life. Figure 2 presents the language-wise percentage distribution of question–answer pairs in the benchmark. Although the dataset is multilingual, the representation is not uniform across languages, which may influence cross-lingual evaluation dynamics.

4.2 Quantitative Results

To quantify the impact of region-aware hybrid retrieval on culturally situated reasoning, we perform a comprehensive evaluation on the BLEnD benchmark. The most accurate answer (of the four) will be identified through an exact match to determine how accurately the cultural appropriateness of a model's response can be without generating free-text output. As illustrated in Figure 2 2, a significant amount of variance exists in performance levels among different languages; that is, high-resource languages demonstrate much greater accuracy than many low-resource languages. The

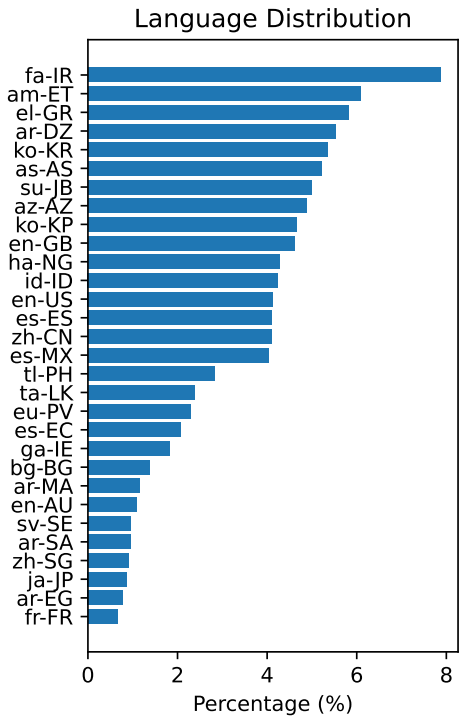

Figure 2: Percentage distribution of question–answer pairs across languages in the BLEnD benchmark.

continued performance disparity between the two groups of languages relates to the disparity of training data available for the models; thus, retrieval-based augmentation affects the degree of performance disparity to a certain extent. However, the overall performance difference is smaller than typically observed in analyses of pure parametric (i.e., non-hybrid) LLMs, suggesting that the hybrid ranking approach may help mitigate some of the existing disparities in cultural knowledge representation.

5 Limitation and Conclusion

Figure 3's quantitative evaluation results indicate that, although the model performs adequately from an accuracy perspective, there is still a substantial lack of success in low-resource languages (e.g., low accuracy). Although the hybrid retrieval mechanism proposed in this study alleviates some of the cultural knowledge gaps present in certain languages, the performance gap remains in languages that were not as well represented during large-scale pretraining (e.g., Amharic, Hausa, and Sundanese) based on the BLEnD benchmark. Therefore, it appears that retrieval augmentation alone will not adequately address some of the imbalances created by using large-scale pretraining corpora for

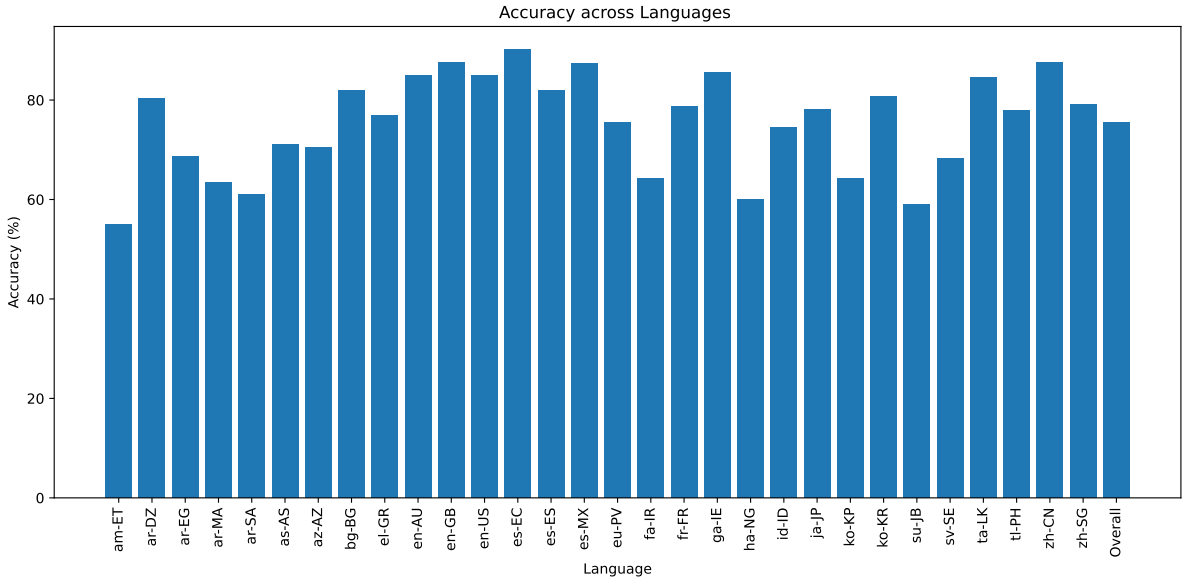

Figure 3: Test-set performance comparison across languages.

creating a machine translation system. One limiting factor is how well the search engine depends on only external document collection. If there are not enough relevant cultural documents available, or if they do not match the query, the retrieval module will be more likely to retrieve irrelevant documents than it will to provide a basis for using meaningful contextual evidence. In addition, the region-weighting heuristic may not work properly in situations where the cultural reference is not stated explicitly, but only found on the surface of a region. Therefore, it may not be effective in many cases because it evaluates cultural documents based on their general location rather than on their specific nature. A second limitation is that a logit-based deterministic answer selection strategy limits the reasoning to next-token probability distributions and does not model multi-step deliberation directly, even though this provides for robustness of format and computation efficiency. As a result, more nuanced, culturally-grounded reasoning involving deeper inferences may be missed. Afterwards, 4 bit quantizing allows quick and effective uses of synthetic intelligence, but can result in small losses of representational precision, which can impact the ability of the AI to perform well on languages that are very linguistically complex or grammatically rich. In this paper, we proposed an interesting hybrid retrieval and logit-based inference structure that was created using regionally contextualized data to answer multiple-choice questions with cultural background. Results from our experiments using the BLEnD benchmark indicated that actually incorporating lexical matching, semantic similarity, and regionally prioritized ranking into a hybrid retrieval and logit-based inference model helped stabilize cross-cultural reasoning performance. Nevertheless, the ongoing performance gaps between low resource languages remind us that dissimilar training sets have an ongoing negative impact on the performance of AI systems; and there is a continuing need to use more culturally diverse training and retrieval methods on the development and improvement of AI systems. It is essential for future work to focus on developing dynamic region models, creating a culturally informed re-ranking mechanism, and developing fine-tuning strategies that target low-resource linguistic communities. By enhancing multilingual evidence bases and using structured cultural knowledge graphs to bolster the cross-cultural evaluation of LLMs, we can improve the integrity and fairness of the evaluations.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Dongyeop Jang, Tae-Rim Yun, Choong-Yeol Lee, Young-Kyu Kwon, and Chang-Eop Kim. 2023. Gpt-4 can pass the korean national licensing examination for korean medicine doctors. *PLOS Digital Health*, 2(12):e0000416.

Timothy R. McIntosh, Tong Liu, Teo Susnjak, Paul Watters, Alex Ng, and Malka N. Halgamuge. 2024. A culturally sensitive test to evaluate nuanced gpt hallucination. *IEEE Transactions on Artificial Intelligence*, 5(6):2739–2751.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2025. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Nedjma Ousidhoum, Junho Myung, Carla Perez-Almendros, Jiho Jin, Amr Keleg, Meriem Beloucif, Yi Zhou, Rodrigo Agerri, Vladimir Araujo, Naomi Baes, James Barry, Joanne Boisson, Nancy F. Chen, Christine de Kock, Aleksandra Edwards, Joseba Fernandez de Landa, Mohamed Fazli Imam, Huda Hakami, Shu-Kai Hsieh, and 11 others. 2026. SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shiyu Wang, Qian Ouyang, and Bing Wang. 2024. Comparative evaluation of commercial large language models on promptbench: An english and chinese perspective.

Peng Yu, Changchang Fang, Xiaolin Liu, Wanying Fu, Jitao Ling, Zhiwei Yan, Yuan Jiang, Zhengyu Cao, Maoxiong Wu, Zhiteng Chen, Wengen Zhu, Yuling Zhang, Ayiguli Abudukeremu, Yue Wang, Xiao Liu, and Jingfeng Wang. 2024. Performance of chatgpt on the chinese postgraduate examination for clinical medicine: Survey study. *JMIR Med Educ*, 10:e48514.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Weihang You, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2026. Opportunities and challenges of large language models for low-resource languages in humanities research. *Preprint*, arXiv:2412.04497.