From National Curricula to Cultural Awareness: Constructing Open-Ended Culture-Specific Question Answering Dataset

Anonymous ACL submission

Abstract

Large language models (LLMs) achieve strong performance on many tasks, but their progress remains uneven across languages and cultures, often reflecting values latent in English-centric training data. To enable practical cultural alignment, we propose a scalable approach that leverages national social studies curricula as a foundation for culture-aware supervision. We introduce **CuCu**, an automated multi-agent LLM framework that transforms national textbook curricula into open-ended, culture-specific question–answer pairs for supervised fine-tuning (SFT). Applying **CuCu** to the Korean national social studies curriculum, we construct **KCaQA**, comprising 34.1k open-ended QA pairs. Our analyses and training experiments suggest that **KCaQA** covers culture-specific topics and produces responses grounded in local sociocultural contexts. [1]

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance on complex tasks, yet this progress has been uneven across languages and cultures. As training corpora are dominated by English-centric resources (Qin et al., 2025; Zhu et al., 2024; Blasi et al., 2022), LLMs often internalize values and cultural framing latent in these resources. This yields culturally hollow outputs or inappropriate behaviors when applied to underrepresented regions (Tao et al., 2024; AlKhamissi et al., 2024; Held et al., 2023). As LLMs are deployed in culture-sensitive domains, this skew becomes consequential, leading to reduced performance, biased judgments, and limited cultural sensitivity in local sociocultural contexts (Romanou et al., 2025; Kwok et al., 2024; Cecilia Liu et al., 2024; Foroutan et al., 2022).

Post-training (*e.g.*, supervised fine-tuning and preference optimization) offers a practical path toward culture-adapted or sovereign LLMs (Xu et al., 2025; Guo et al., 2025; Masoud et al., 2025). However, obtaining high-quality supervision grounded in local civic norms, historical narratives, and social institutions remains a major bottleneck (Laiyk et al., 2025). Existing multilingual post-training datasets are typically general-purpose (*e.g.*, daily life, STEM, reasoning) and often produced via translation of English resources (Li et al., 2023), which provides linguistic coverage but weak cultural grounding (Rachamalla et al., 2025).

In this paper, we propose using national curricula as a structured prior for generating culture-grounded supervision at scale. Unlike ad hoc prompts or translated datasets, national curricula are carefully designed by domain experts and, in many countries, publicly published by governments or educational authorities; they explicitly enumerate what students are expected to learn and thereby provide both coverage (which topics a society prioritizes) and constraints (how those topics should be framed in relation to civic norms and institutions). This structure offers a principled backbone for constructing open-ended, culture-specific question answering for supervised fine-tuning (SFT).

We introduce **CuCu** (from national **Cu**rricula to **Cu**ltural Awareness, pronounced *kkukku*), a multi-agent LLM framework that leverages textbook curricula to automatically generate open-ended, culture-specific question–answer pairs (Figure 1). We apply **CuCu** to the Korean national social studies curriculum to produce **KCaQA** (**K**orean **C**ulturally-**a**ware **QA**, pronounced *kkakka*), comprising 34.1k open-ended, culture-specific QA pairs on Korean culture in four languages. By grounding the generation process in national standards, **CuCu** ensures that the resulting data incorporates culture-specific knowledge as well as normative perspectives required for genuine cultural awareness.

[1] **CuCu** and **KCaQA** are available at https://anonymous.4open.science/r/cucu/.

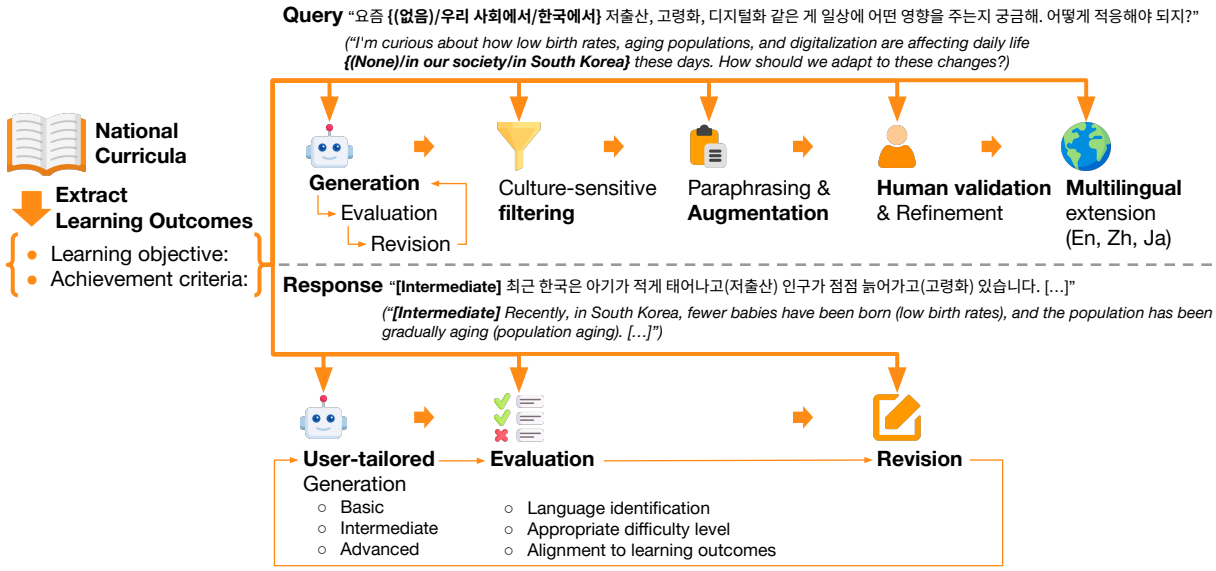

Figure 1: Overview of **CuCu**. **CuCu** constructs an open-ended culture-specific QA pairs using national curricula.

2 Related Work

2.1 AI-assisted Data Construction

The construction of natural language data has traditionally been conducted manually, but the recent disclosure of generative language models has enabled practitioners to obtain such resources with the help of machines. Especially for the pre-training and post-training data of language models, Gunasekar et al. (2023) has demonstrated the significance of synthetic textbook-like materials as a sufficient and necessary source of learning. Following this light, Cosmopedia (Ben Allal et al., 2024) investigated diverse depths in data synthesis, considering a wide range of domains and audiences. These approaches led to multiple data generation attempts in alignment (Wang et al., 2024), reducing hallucination (Jones et al., 2024), and further synthesizing frameworks (Maini et al., 2025).

2.2 Culture-specific NLP Data

Culture-specific NLP data has long been a topic of interest among diversity and inclusion interest groups. While culture incorporates a wide range of dimensions (Spencer-Oatey and Franklin, 2012), prior works have primarily focused on ensuring or benchmarking language models' capability to understand professional knowledge, such as customs (Li et al., 2024; Myung et al., 2024; Chiu et al., 2025), norms (Rao et al., 2025), perceptions (Shi et al., 2024), and common grounds (Tanwar et al., 2025) shared among cultural groups. Furthermore, these data primarily consist of closed-ended and short-form QAs, with human-authored and validated items, which differ from natural user queries and make scaling challenging.

Especially for Korean, prior studies have evaluated cultural alignment as specific tasks, focusing on cultural heritage (Kim et al., 2022), historical knowledge (Son et al., 2024), factuality (Ko et al., 2025), or social bias (Jin et al., 2024; Lee et al., 2023). While Lee et al. (2024) investigated social values and common knowledge regarding Korean among LLMs, the queries are close-ended and does not necessarily account for LLM-targeted questions. In this light, we aim to introduce a scalable framework and the following dataset to address open-ended, culture-specific questions based on social consensus grounded in national curricula.

3 CuCu and KCaQA

We introduce **CuCu** (from national **Cu**rricula to **Cu**ltural Awareness), a multi-agent LLM framework that constructs an open-ended, culture-specific QA dataset using national curricula. Figure 1 describes an overview pipeline of **CuCu**. Details on **CuCu** are described in Appendix B.

Query Generation. We use `Solar Pro 2` (Upstage, 2025) and `GPT-4o` (OpenAI et al., 2024) to generate or revise query texts and to evaluate query texts in each LLM agent, respectively.

1. **Initial query generation:** We generate an initial query in Korean based on a learning outcome (*i.e.*, a pair of a learning objective and an achievement criterion). We (1) generate a query,

(2) evaluate whether it fully reflects the learning outcome, and (3) revise based on the feedback, repeating this loop up to five times.

2. **Culture-sensitive filtering:** We filter out queries that involve general, culture-agnostic knowledge, further improving their naturalness and language use.

3. **Paraphrasing & Augmentation:** We paraphrase the initial query into two additional queries that target the same learning outcomes but differ in style and tone. Each resulting query is expanded into three variants: (1) without mentioning any culture or country, (2) implicitly referring to "our/my" culture or country, and (3) explicitly stating the name of culture or country.

4. **Human validation & Refinement:** One of the authors, who is a Korean L1 speaker born and raised in Korea, educated under the Korean national curricula and thus familiar with its norms and culture, manually validates and refines queries generated by multi-LLM agents.

5. **Multilingual extension:** We finally extend the queries into three additional languages (*i.e.*, English, Chinese, and Japanese) by translating the ones explicitly stating the culture/country name.

Response Generation. We use four open-source LLMs (*i.e.*, `GPT-OSS` (OpenAI et al., 2025), `Qwen3`, `Qwen3-Next` (Yang et al., 2025), and `DeepSeek R1` (DeepSeek-AI et al., 2025)) for response.

1. **User-tailored response generation:** For each query, we produce three variants targeting different audience bands—*Basic*, *Intermediate*, and *Advanced*—to cover a range of readability and assumed background knowledge.

2. **Response evaluation:** We evaluate each candidate for (1) language consistency with the query, (2) appropriateness of difficulty in terms of style/readability and content depth, and (3) alignment with the corresponding learning outcomes.

3. **Response revision:** We revise the response and repeat this process up to five times.

To showcase **CuCu**, we use the 2022 Revised Elementary, Middle, and High School National Curriculum of Social Studies [2], published by the Ministry of Education, Republic of Korea, as a source data. We extract 357 atomic pairs of learning objectives and achievement criteria and filter out 158 learning outcomes, generating 2,844 queries (*i.e.*, 3 paraphrasing × (3 Korean variants + 3 translations) × 158 learning outcomes). In total, we construct **KCaQA** (**K**orean **C**ulturally-**a**ware **QA**), comprising 34,128 QA pairs (*i.e.*, three different responses per 2,844 queries) that reflect Korean cultural knowledge and perspective. An example instance of **KCaQA** is provided in Appendix C. We provide further details on our design choice and discuss wider adaptability in Appendices D.1 and D.2.

4 Discussions

4.1 KCaQA Quantitative Analysis

Culture-specific Query. We analyze whether **KCaQA** queries cover culture-specific topics using topic modeling. In particular, we compare the Korean explicit queries in **KCaQA** with Korean LIMA (Zhou et al., 2023) dataset [3], a high-quality, general-domain QA dataset for LLM alignment training. The resulting distributions diverge sharply, with a JSD of 0.807 between **KCaQA** and Korean LIMA; a 1k permutation test confirms the difference is significant ($p < 0.01$). A qualitative inspection of the top-10 most skewed topics suggests that **KCaQA** mainly covers Korean sociocultural content (*e.g.*, Korean geography, economy, history, climate, and politics).

User-tailored Response. We examine whether **KCaQA** responses match their intended difficulty level (*i.e.*, *Basic*, *Intermediate*, and *Advanced*) using readability statistics, as reported in Table 2. In general, **KCaQA** responses become longer and syntactically denser with increasing target level, while lexical diversity also increases. Further analyses on both query and response of **KCaQA** are provided in Appendix D.3.

4.2 KCaQA Quality Estimation

We validate the data quality of **KCaQA** by LLM-as-a-judge (Zheng et al., 2023) and human inspection. The evaluation criteria with detailed rubrics are described in Appendix E.

- **Language selection**: If both query and response are monolingual in the target language;

[2] https://ncic.re.kr/

[3] Translated version of LIMA dataset. Accessible at https://huggingface.co/datasets/taeshahn/ko-lima

	CLIcK	BLEnD	CulturalBench	KorNAT
Qwen 3 14B	0.587	0.842	0.585	0.679
+ 1. CultureBank	0.611	0.859	0.607	0.655
+ 2. Korean LIMA	0.615	0.853	0.611	0.682
+ 3. National curriculum	0.633	0.866	0.615	0.721
+ 4. **CuCu** without source data	0.636	0.870	0.624	0.736
+ 5. **KCaQA** (Ours)	**0.662**	**0.891**	**0.634**	**0.744**

Table 1: Experimental results of LLM post-training (SFT) using **CuCu** and **KCaQA** tailored for Korean culture. We use CLIcK, BLEnD, and CulturalBench for cultural knowledge evaluation and KorNAT for cultural alignment evaluation. We extract and use the samples targeting Korean culture for BLEnD and CulturalBench.

- **Cultural appropriateness**: If both query and response adhere to the desired stance regarding culture-sensitive concepts;
- **Language use**: If both query and response use appropriate and natural style and expressions.

We employ `GPT-5.2` (OpenAI, 2025) to automatically assess the quality of **KCaQA**. Specifically, the model outputs a binary decision for *language selection* (0/1) and assigns Likert-scale scores (1–10) for *cultural appropriateness* and *language use*. **KCaQA** achieves 0.91 accuracy on language selection, with average scores of 8.56 and 7.78 for cultural appropriateness and language use, respectively. These results suggest that **KCaQA** produced by **CuCu** largely satisfies the intended language constraints and stance requirements for culture-specific concepts. In addition, we manually inspect all Korean instances and a randomly sampled set of multilingual instances. We observe that, in most cases, queries capture the core topic of the learning objective, and responses align with the corresponding criteria. The full results, including inter-annotator agreements and failure analysis, are discussed in Appendix D.4.

4.3 LLM Training using KCaQA

We train `Qwen 3 14B` on five datasets for Korean cultural alignment: (1) CultureBank (Shi et al., 2024), a multilingual, multicultural dataset sourced from Tiktok and Reddit, (2) Korean LIMA, a general-purpose, open-domain QA dataset, and three ablation variants of **CuCu** (*i.e.*, (3) national curricula, (4) **CuCu** without source data, and (5) **KCaQA**). We use CLIcK (Kim et al., 2024) cultural questions, BLEnD (Myung et al., 2024) multiple-choice questions, CulturalBench (Chiu et al., 2025)-Hard for cultural knowledge evaluation, as well as KorNAT (Lee et al., 2024) for cultural alignment evaluation, all tailored for Korean culture. Experimental settings are described in Appendix D.5

Table 1 shows that the model trained with KCaQA yields the strongest gains on both cultural knowledge and alignment. While training on the raw curriculum improves performance, it is weaker than training on **KCaQA**. It implies that converting curricula into structured, open-ended supervision provides non-trivial, transferable gains. Furthermore, **CuCu** without source grounding data also improves cultural knowledge and alignment, indicating that the framework design itself contributes. However, using CultureBank does not substantially enhance cultural knowledge or alignment, likely because it is sourced from Reddit or TikTok and primarily focuses on how people from other cultures perceive cultural differences and culture shock rather than grounded civic norms.

5 Conclusion

We propose a scalable recipe for culture-grounded post-training supervision by leveraging national social studies curricula as an expert-structured prior. We introduce **CuCu**, an automated multi-agent LLM pipeline that converts learning outcomes into open-ended, culture-specific QA pairs, and apply it to the Korean national social studies curriculum to generate **KCaQA** comprising 34.1k QA pairs across four languages. Through quantitative and qualitative analyses, we show that **KCaQA** captures curriculum-specified, culture-specific topics and yields answers that are grounded in Korean sociocultural context across languages. Through rigorous training experiments, we empirically validate the effectiveness of **CuCu** and **KCaQA** for cultural alignment. This study provides a practical and reproducible path to constructing culturally grounded supervision for sovereign or culture-adapted LLMs.

Limitation

This study incorporates several conceptual and methodological limitations that require acknowledgment.

First, the concepts of language, linguistic region, cultural region, ethnicity, and nation-state each carry distinct meanings, yet our approach examines cultural reflection primarily through the lens of national curricula. This methodology proves feasible for South Korean, where its compact nation-state highly aligns with the cultural boundary. However, this approach faces challenges when applied to nation-state in multiple geocultural regions with varying cultural contexts.

Second, our analysis focuses exclusively on Seoul Korean as used in the Republic of Korea (ROK) and references the standard curriculum of Republic of Korea. This scope excludes Pyongyang Korean as employed in the Democratic People's Republic of Korea (DPRK) and does not consider North Korea's educational standards. While we attempted to select queries and responses that would minimize controversial elements, certain aspects may not align with North Korean linguistic or educational norms.

Third, our use of national curricula as primary sources for query and response generation, while ensuring verified and socially accepted content, does not fully capture the range of geocultural characteristics expressed through language. Curricula represent institutionalized knowledge rather than the complete spectrum of cultural expression. Nevertheless, our framework maintains sufficient flexibility to accommodate alternative source materials that may more comprehensively represent linguistic and cultural phenomena when such resources become available.

Broader Impact Statement

Our dataset construction process mainly consists of automated synthesis and validation, which is transparently shared to ensure reproducibility. Additionally, our study does not involve the recruitment or engagement of human subjects, thereby minimizing subjective influence. While we conduct some manual inspection for qualitative analysis, we adopt the same predefined guidelines as those provided to LLMs.

We regard the national curricula of the Republic of Korea as a reliable data source, which may be subject to bias. While the 2022 reform reflects some public will for curricular revision (Cho et al., 2023; Lee and Eun, 2024), it does not necessarily guarantee that curricula are presented in a fully unbiased manner, with all potential concerns addressed. This implies a potential amplification of the bias for those who use our framework, queries, or answers in their practice. Furthermore, multiple nations, including the Republic of Korea [4], have ongoing discussions about bias or revisionism in some national curricula. Therefore, practitioners from multiple national or cultural regions should caution that the national curricula to be adopted may propagate unresolved controversies, either at the domestic or international level.

We use ChatGPT, Gemini, Claude, and Copilot for writing and coding assistance.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Cosmopedia.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. 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 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

[4] https://en.wikipedia.org/wiki/Bias_in_curricula

Linguistics (Volume 1: Long Papers), pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Chul-Ki Cho, HyeSook Kim, and Soyoung Lee. 2023. Analysis of the issues that emerged in the revision of the national social studies curriculum in South Korea: Text mining and semantic network analysis of the comments at the public hearing on YouTube. *Journal of Education and E-Learning Research*, 10(3):463–473.

Eunjung Cho, Won Ik Cho, and Soomin Seo. 2025. Hermit kingdom through the lens of multiple perspectives: A case study of LLM hallucination on North Korea. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3353–3371, Abu Dhabi, UAE. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32854–32883, Suzhou, China. Association for Computational Linguistics.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in NLP. *arXiv preprint arXiv:2311.08391*.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Erik Jones, Hamid Palangi, Clarisse Simões Ribeiro, Varun Chandrasekaran, Subhabrata Mukherjee, Arindam Mitra, Ahmed Hassan Awadallah, and Ece Kamar. 2024. Teaching language models to hallucinate less with synthetic tasks. In *The Twelfth International Conference on Learning Representations*.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Gyeongmin Kim, Jinsung Kim, Junyoung Son, and Heuiseok Lim. 2022. KoCHET: A Korean cultural heritage corpus for entity-related tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3496–3505, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Donghyeon Ko, Yeguk Jin, Kyubyung Chae, Byungwook Lee, Chansong Jo, Sookyo In, Jaehong Lee, Taesup Kim, and Donghyun Kwak. 2025. KoSimpleQA: A Korean factuality benchmark with an analysis of reasoning LLMs. *arXiv preprint arXiv:2510.18368*.

Louis Kwok, Michal Bravansky, and Lewis Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. In *First Conference on Language Modeling*.

Nurkhan Laiyk, Daniil Orel, Rituraj Joshi, Maiya Goloburda, Yuxia Wang, Preslav Nakov, and Fajri Koto. 2025. Instruction tuning on public government and cultural data for low-resource language: a case study in Kazakh. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14509–14538, Vienna, Austria. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand. Association for Computational Linguistics.

Jungwoo Lee and Ji-Yong Eun. 2024. Presence of social minorities in high school social studies textbooks in the post-democratic era in South Korea. *Cogent Education*, 11(1):2406580.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. CulturePark: Boosting cross-cultural understanding in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 65183–65216. Curran Associates, Inc.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-X: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Pratyush Maini, Vineeth Dorna, Parth Doshi, Aldo Carranza, Fan Pan, Jack Urbanek, Paul Burstein, Alex Fang, Alvin Deng, Amro Abbas, and 1 others. 2025. BeyondWeb: Lessons from scaling synthetic data for trillion-scale pretraining. *arXiv preprint arXiv:2508.10975*.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

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. 2024. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

OpenAI. 2025. GPT-5 System Card.

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, and 262 others. 2024. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Suzi Park and Hyopil Shin. 2021. Kr-sbert: A pre-trained korean-specific sentence-bert model. https://github.com/snunlp/KR-SBERT.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Neel Prabhanjan Rachamalla, Aravind Konakalla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. Pragyaan: Designing and curating high-quality cultural post-training datasets for Indian languages. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 285–321, Suzhuo, China. Association for Computational Linguistics.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024. HAE-RAE bench: Evaluation of Korean knowledge in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7993–8007, Torino, Italia. ELRA and ICCL.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Helen Spencer-Oatey and Peter Franklin. 2012. What is culture. *A compilation of quotations. GlobalPAD Core Concepts*, 1(22):1–21.

Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang Wang, and Tanmoy Chakraborty. 2025. Do you know about my nation? investigating multilingual language models' cultural literacy through factual knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14967–14990, Suzhou, China. 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):pgae346.

Upstage. 2025. Solar Pro 2.

Zifeng Wang, Chun-Liang Li, Vincent Perot, Long Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. 2024. CodecLM: Aligning language models with tailored synthetic data. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3712–3729, Mexico City, Mexico. Association for Computational Linguistics.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6859–6877, Albuquerque, New Mexico. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. 2024. Multilingual large language models: A systematic survey. *arXiv preprint arXiv:2411.11072*.

Appendix

A Reproducibility Statement

For open-source LLM inference, we use 4 H200 GPUs with 564 GB memory and 4 RTX 8000 GPUs with 188 GB memory. For proprietary LLM inference, we use the official API implementation of each model. We specifically use the versions of `gpt-4o-2024-08-06` and `gpt-5.2-2025-12-11` for `GPT-4o` and `GPT-5.2`, respectively. We use the following LLMs for response generation via Hugging Face:

- `openai/gpt-oss-120b`,
- `Qwen/Qwen3-235B-A22B`,
- `Qwen/Qwen3-Next-80B-A3B-Thinking`,
- `deepseek-ai/DeepSeek-R1`.

In Section 4.1, we use Korean Sentence BERT (Park and Shin, 2021) for query analysis to fit a single BERTopic (Grootendorst, 2022) model on the two corpora, representing each instance as a document that concatenates the query and its paired response. For response analysis, tokens are obtained with `wordfreq` (Speer, 2022), a unified multilingual tokenizer. The rare-token ratio is computed only over in-vocabulary tokens (OOV excluded), using a language-specific threshold set to the bottom 25% of Zipf-frequency values estimated from all responses in that language. In Section 4.3, we use 4 NVIDIA H200 141GB GPUs with CUDA version 12.8 to train and evaluate the models.

B Implementation Details on CuCu

B.1 Query Generation

Initial Query Generation. We use `GPT-4o` as an evaluation agent to assess whether a query (1) reflects the learning objective and (2) incorporates key ideas from the achievement standards. The agent scores alignment on a 5-point Likert scale and provides short revision feedback.

Culture-sensitive Filtering. We use `GPT-4o` as an evaluation agent to filter out queries that involve general, culture-agnostic knowledge according to the following criteria: whether a query (1) aligns with the Korean cultural and social context, (2) reflects social consensus institutionalized in Korean society, (3) is not a generic, universal question, and (4) is meaningfully grounded in the Korean context. This ensures that queries elicit perspectives shaped by local institutions and normative values, rather than subjective stereotypes. For instance, the following questions are rejected and accepted, respectively, through the filtering process:

Example questions within the filtering process

- **Reject (General/Culture-agnostic):** "*What are the general causes of colonialism?*"
- **Accept (Korea-specific context):** "*How did Korean society and culture change during the Japanese colonial period, and what movements did people engage in? Give me a quick rundown of daily life changes and the key players behind the movements.*"
  - Reason: The Japanese colonial period is more than a mere historical fact; it is a core narrative prioritized in the curriculum as the foundation of South Korea's constitutional spirit and civic consciousness.

Augmentation & Multilingual Extension. We vary the linguistic framing of already culture-grounded queries into three settings: (1) no-country, (2) implicit, and (3) explicit. The motivation is as follows: even when a user does not explicitly mention the country name"Korea" in a Korean language query, a culturally aligned model should correctly infer the intended sociocultural context as Korean. Importantly, the "no-country" variant does not remove cultural grounding. It removes only explicit lexical markers, while the underlying learning outcome and assumed context remain Korean. This distinction becomes especially important in the multilingual extension stage. As the query language changes (*e.g.*, from Korean to English, Chinese, or Japanese), implicit cultural assumptions tied to Korean no longer hold. Therefore, in multilingual extension, we translate only the explicitly culture-marked variant, ensuring that the Korean sociocultural grounding remains clear in non-Korean languages.

Human Validation & Refinement. This process mainly includes (1) revising queries to be natural colloquial questions, (2) inspecting if the variants imply the same meaning regardless of the presence of culture or country-related expressions, and (3) assuring that queries adhere to the learning objective and the achievement criteria, not handling topics that are overly generic or too professional.

C KCaQA Samples

Example instance extracted from Korean social studies national curriculum

- Learning objectives: [4사03-01] 최근 사회 변화의 양상과 특징을 파악하고, 그로 인해 나타난 생활모습의 변화를 탐색한다.
  (*Translation: [4SA03-01] Identify the patterns and characteristics of recent social changes and explore changes in daily life that have emerged as a result.*)

- Achievement criteria: [4사03-01]은 사례를 통해 저출산, 고령화, 지능정보화 등 최근 사회 변화가 가져온 현상과 특징을 파악하고, 그로 인해 우리의 생활 모습이 어떻게 변화되는지 분석하도록 설정한 것이다. 이를 통하여 사회 변화에 대응할 수 있는 역량을 기르는 데 초점을 맞추도록 한다.
  (*Translation: [4SA03-01] is designed to have students understand the phenomena and characteristics brought about by recent social changes—such as low birth rates, population aging, and intelligent informatization—through examples, and to analyze how these changes affect our daily lives. Through this process, the focus is on fostering the capacity to respond to social changes.*)

- Query:
  - Korean (No country): 요즘 저출산, 고령화, 디지털화 같은 게 일상에 어떤 영향을 주는지 궁금해. 어떻게 적응해야 되지?
  - Korean (Implicit): 요즘 우리 사회에서 저출산, 고령화, 디지털화가 일상에 어떤 영향을 주는지 궁금해. 어떻게 적응해야 될까?
  - Korean (Explicit): 요즘 한국에서 저출산, 고령화, 디지털화가 일상에 어떤 영향을 주는지 궁금해. 어떻게 적응해야 되지?
  - English: I'm curious about how low birth rates, aging populations, and digitalization are affecting daily life in South Korea these days. How should we adapt to these changes?
  - Chinese: 最近好奇韩国的低出生率，老龄化和数字化对日常生活有什么影响？应该怎么适应这些变化呢？
  - Japanese: 最近，韓国の低出生率，高齢化，デジタル化が日常生活にどんな影響を与えているのか気になります。こうした変化にどう適応していけばいいですか？

- Response:
  - Basic: 우리나라에 아기들이 적어지고 어른이 많아졌어요. 그래서 가족이 함께 시간을 보내는 게 더 중요해요. 스마트폰과 컴퓨터를 잘 사용하면 생활이 편해져요. 모두가 서로 도와가며 살아가면 좋아요.
    (*Translation: In our country, there are fewer babies being born and more adults. So it's more important for families to spend time together. If we use smartphones and computers well, life becomes more convenient. It's good when everyone lives by helping each other.*)
  - Intermediate: 최근 한국은 아기가 적게 태어나고(저출산) 인구가 점점 늙어가고(고령화) 있습니다. 동시에 스마트폰, 인터넷 등 디지털 기술이 일상 곳곳에 스며들어(디지털화) 생활이 편리해졌지만 새로운 적응이 필요합니다. 작은 가정에서는 아이를 돌볼 사람이 부족해 보육 서비스 이용이 늘고, 많은 가정이 조부모와 함께 사는 다세대 형태가 늘어납니다. 디지털화 덕분에 온라인 쇼핑, 원격 근무, 비대면 의료 등 새로운 생활 방식이 보편화되었습니다. 이런 변화에 적응하려면 디지털 기기 사용법을 배우고, 세대 간에 서로 도와주는 문화가 중요합니다.
    (*Translation: Recently, Korea has seen fewer babies being born (low birth rate) and the population is gradually getting older (aging). At the same time, digital technologies such as smartphones and the internet have become part of everyday life (digitalization), making life more convenient but also requiring people to adapt in new ways. In smaller households, there often aren't enough caregivers for children, so the use of childcare services is increasing, and more families are living in multi-generational households with grandparents. Thanks to digitalization, new lifestyles such as online shopping, remote work, and contactless healthcare have become widespread. To adapt to these changes, it is important to learn how to use digital devices and to build a culture where different generations support one another.*)
  - Advanced: 2023년 우리나라 합계출산율은 0.78명으로 세계 최저 수준이며, 65세 이상

인구 비중은 16.5%에 달해 고령화가 급속히 진행되고 있습니다. 동시에 인터넷 보급률은 96%에 육박하고, 인공지능 · 빅데이터 기반 서비스가 일상에 스며들면서 디지털 전환이 가속화되고 있습니다. 이러한 인구 · 기술 변화는 노동시장 구조조정, 연금 · 보건복지 재정 압박, 소비 패턴의 세대별 차별화 등을 초래하고 있습니다. 대응 방안으로는 평생학습 체계 강화와 디지털 격차 해소를 위한 국가 · 지방 차원의 교육 프로그램 확대가 필요합니다. 또한, 세대 간 교류를 촉진하는 커뮤니티 센터와 스마트 시티 인프라를 활용해 고령층의 사회참여와 안전을 지원해야 합니다. 기업은 유연근무제와 가족친화 정책을 도입해 출산 · 육아 부담을 경감하고, 정부는 출산 장려금 · 보육 서비스 질 향상을 통해 저출산 문제를 완화할 수 있습니다. 이러한 다각적 접근이 개인과 사회가 함께 변화에 적응하는 기반이 될 것입니다.
(*Translation: In 2023, Korea's total fertility rate was 0.78, one of the lowest in the world, and the share of the population aged 65 and older reached 16.5%, showing that aging is progressing rapidly. At the same time, internet penetration is close to 96%, and digital transformation is accelerating as services based on artificial intelligence and big data spread into daily life. These demographic and technological shifts are causing labor-market restructuring, increasing financial pressure on pensions and health/welfare systems, and widening generational differences in consumption patterns. As responses, it is necessary to strengthen lifelong learning systems and expand education programs at the national and local levels to reduce the digital divide. In addition, community centers that promote intergenerational exchange, along with smart-city infrastructure, should be used to support older adults' social participation and safety. Companies can reduce the burdens of childbirth and childcare by introducing flexible work arrangements and family-friendly policies, while the government can ease the low birth rate problem through childbirth incentives and improving the quality of childcare services. This multifaceted approach will become the foundation for individuals and society to adapt to change together.*)

D Additional Discussions

D.1 Design Choices of CuCu and KCaQA

CuCu. During the response generation in **CuCu**, we produce three response variants—Basic, Intermediate, and Advanced—based on the following dimensions.

Dimensions for three response variants

- **Basic:**
  - Audience: elementary school students and young children
  - Language: very simple words, short sentences, and concrete ideas
  - Style: use everyday examples and simple comparisons; one idea per sentence
  - Content: provide only the most basic and essential meaning of the topic
  - Length: around 2–3 short sentences
- **Intermediate:**
  - Audience: high school students and young adults
  - Language: clear, structured explanations similar to the national curriculum and textbooks
  - Style: show balanced reasoning and connect personal experience with broader social context
  - Content: when relevant, incorporate the Learning Objective and Achievement Standards
  - Length: around 4–6 sentences
- **Advanced:**
  - Audience: adults, university students, graduate-level learners, and professionals
  - Language: analytical, evidence-based, and conceptually deeper
  - Style: culturally grounded perspectives common among well-educated Korean adults
  - Content: offer contextual analysis, balanced arguments, and practical implications; when relevant, mention history, institutions, data, or research
  - Length: around 6–12 sentences, possibly in two short paragraphs

For query generation, we intentionally separate

the model that generates or revises query texts from the one that evaluates them, to reduce its bias toward its own style. Here, we use `Solar Pro 2` with the MIT license for the query generation and revision, while we adopt `GPT-4o` for query evaluation. For response generation, we use open-source LLMs (Apache or MIT licenses), as the generated data may be used for further model training.

KCaQA. The up-to-date Korea national curricula released in 2022 incorporate the following subjects: *Korean language*, *Mathematics*, *Social Studies*, *Science*, *Moral Education*, *Technology & Home Economics*, *Music*, *Fine Arts*, *Physical Education*, *Computer & Information*, *Classical Chinese*, *Second Foreign Language*, and *Creative Experiential Activities*. Among them, we use the *Social Studies* curriculum only to build **KCaQA**, as other subjects largely cover culture-agnostic, general knowledge. For example, *Moral Education* primarily does not target Korean normative values but rather offers general social philosophy and theory. In contrast, *Social Studies* encompasses various sub-contents closely related to Korean culture: Korean/Asian/World History, Korean/International Geography, Society & Culture, Common Social Studies, Politics, Economics, Law & Society, and Understanding International Relations.

D.2 Adaptability of CuCu and KCaQA

CuCu While we implement and evaluate our framework only for the Republic of Korea and the Korean language, **CuCu** generalizes to multiple nation-states in various geo-cultural regions. National curricula are typically published by governments or educational authorities and are publicly available—*e.g.*, United Kingdom [5], India [6], Bangladesh [7], and Australia [8]. Nonetheless, we urge practitioners to consider that (1) a nation-state may not follow a single uniform curriculum, with policies and perspectives varying across subgroups; (2) constraining outputs to a specific language may inadvertently shift the framing toward perspectives prevalent regions where that language is widely used (Cho et al., 2025); and (3) **CuCu** may inherit and potentially amplify any bias and revisionism latent in the curriculum without a thorough review.

[5] https://www.gov.uk/government/collections/national-curriculum

[6] http://www.ncert.nic.in/rightside/links/nc_framework.html

[7] https://nctb.gov.bd/

[8] https://www.australiancurriculum.edu.au/

KCaQA. We examine the sample-level adaptability of **KCaQA** to facilitate its extension to other languages and cultures. Following Jin et al. (2024), we categorize the queries associated with each learning outcome into three types: (1) *Simply Transferred*, which requires culture-sensitive translation only; (2) *Target Modified*, whose proper nouns and expressions must be modified to the target culture; and (3) *Sample Removed*, which cannot be meaningfully adapted. Each sample is annotated by three authors who are Korean L1 speakers, resulting in almost perfect inter-annotator agreement (Fleiss' Kappa (Fleiss, 1971) = 0.86). The final labels are determined via majority voting; for samples without a clear majority, all annotators jointly discuss the case to reach a consensus. Consequently, among 158 learning outcomes, 118 are classified as *Simply Transferred*, 35 as *Target Modified*, and 5 as *Sample Removed*. All *Sample Removed* learning outcomes are relevant to North Korea, which has a unique political, historical, and geographical relationship to South Korea. This distribution suggests that most queries and learning outcomes can be adapted to other cultures with minimal effort, highlighting the scalability of **KCaQA** across languages and cultures.

D.3 KCaQA Quantitative Analysis

Culture-specific Query. JSD ranges from 0 (identical distributions) to 1 (maximal divergence). A value of 0.807 in Section 4.1 indicates strong distributional separation between **KCaQA** and Korean LIMA in the shared topic space. To better interpret JSD, we set up two comparison query datasets (*i.e.*, KoAlpaca v1.1 [9] and **KCaQA**-filtered) that are culturally agnostic or general and provide qualitative examples for each case. KoAlpaca v1.1 consists of user-written queries from NAVER Knowledge iN [10], a community-driven Q&A platform in South Korea. It primarily covers general-domain knowledge acquisition and advising scenarios. **KCaQA**-filtered indicates queries generated from national social studies curricula that are filtered out because they involve general, culture-agnostic knowledge. We under sample the datasets with a random seed of 42 to match the data size for fair comparison. We then perform topic modeling and compute JSD compared to Korean LIMA, a standard, well-known general-purpose QA dataset. While **KCaQA** and Ko-LIMA yields a JSD of 0.807, implying a

[9] https://huggingface.co/datasets/beomi/KoAlpaca-v1.1a

[10] https://kin.naver.com/

substantial topic divergence between two datasets, KoAlpaca v1.1 and **KCaQA**-filtered report JSD of 0.201 and 0.343, respectively, indicating relatively similar topic coverage. Through qualitative inspection, we observe that Ko-LIMA and KoAlpaca v1.1 is dominated by general-purpose daily life topics (*e.g.*, STEM, recipes, and advice), while KoAlpaca v1.1 involves several queries about Korea. **KCaQA**-filtered primarily covers general knowledge on social studies (*e.g.*, politics, climate, and economy) without any focus on Korean cultural context. The following example queries showcase each dataset.

Example questions within the filtering process

- **KCaQA**
  - "*I'm curious about how low birth rates, aging populations, and digitalization are affecting daily life in South Korea these days. How should we adapt to these changes?*"
  - "*In South Korea, what do the National Assembly, the executive branch, and the judiciary each do? Why is the separation of powers important for protecting citizens' rights in South Korea?*"
  - "*Can you briefly explain why Dokdo belongs to South Korea, both geographically and historically?*"
- Ko-LIMA
  - "*Does Java casting introduce overhead? Or the compiler just resolves everything and there is no cost at run time? Is this a general things, or there are different cases?*"
  - "*Write me an official resignation letter.*"
  - "*This is my first trip to Shanghai. Please help me plan a classic 3-day tour in Shanghai.*"
- KoAlpaca v1.1
  - "*What is the Pygmalion effect? Can you explain it with a specific example?*"
  - "*Why is Korea so good at archery? Is there any specific reason?*"
  - "*Which bank would you recommend as a primary bank, and why?*"
- KCaQA-filtered
  - "*What are the pros and cons of democracy?*"
  - "*What can I do as an individual to help tackle climate change and support sustainable development?*"
  - "*During inflation, what's the best place to invest among stocks, bonds, and funds? Could you break down how each one works and what risks they carry?*"

User-tailored Response. Table 2 reports readability statistics across languages in terms of response length and lexical diversity, measured by the number of tokens, the number of sentences, tokens per sentence, and the rare token ratio.

Lang.	Lvl.	# tokens	# sentences	TPS	RTR
Ko	B	49.6	3.3	15.5	0.14
	I	113.9	4.8	23.9	0.23
	A	178.2	6.0	30.2	0.30
En	B	27.7	3.5	8.2	0.24
	I	63.6	5.0	12.8	0.25
	A	103.5	6.2	17.1	0.31
Zh	B	79.0	3.4	23.8	0.24
	I	191.5	4.8	40.4	0.26
	A	313.7	6.0	53.6	0.27
Ja	B	27.6	3.4	8.4	0.15
	I	66.5	4.9	13.8	0.21
	A	112.7	6.3	18.6	0.57

Table 2: Readability statistics of **KCaQA** responses across languages and target levels. B, I, and A indicate basic, intermediate, and advanced levels. TPS and RTR denote token per sentence and rare token ratio, respectively.

D.4 KCaQA Quality Estimation

LLM-as-a-Judge. Table 3 reports the full results of LLM-as-a-Judge for quality estimation of **KCaQA**. We further validate these results with a human evaluation. Specifically, we randomly sample 10 QA instances for each setting and each response level in Korean and English, yielding 120 samples in total. One of the authors, who is a native Korean speaker and fluent in English, annotates the samples using the same rubric as the LLM-as-a-Judge (Appendix E). We then compute inter-rater reliability between LLM-as-a-Judge and human annotator. The results show substantial to almost perfect agreement, with Cohen's Kappa (Cohen, 1960) of 0.98, 0.76, and 0.72 for the three dimensions.

Human Inspection. We manually inspect all Korean instances and a randomly sampled set of multilingual instances to qualitatively assess the content and style of **KCaQA**. In most cases, queries capture the core topic of the learning objective,

Lang.	Setting	Lvl.	LS	CA	LU
Ko	No country	B	1.00	9.00	7.28
		I	1.00	9.41	8.68
		A	0.98	8.98	8.28
	Implicit	B	1.00	8.76	7.14
		I	1.00	9.40	8.70
		A	0.99	8.93	8.29
	Explicit	B	1.00	8.57	6.95
		I	0.99	9.31	8.64
		A	0.98	8.93	8.28
En	Explicit	B	0.99	8.00	6.64
		I	0.98	8.95	8.22
		A	0.96	8.76	8.28
Zh	Explicit	B	0.81	7.78	6.66
		I	0.83	8.78	8.27
		A	0.81	8.46	8.10
Ja	Explicit	B	0.94	7.86	6.88
		I	0.87	8.68	8.32
		A	0.73	8.41	7.89

Table 3: LLM-as-a-judge results for quality estimation of **KCaQA**. LS, CA, and LU denote the evaluation criteria: Language Selection, Cultural Appropriateness, and Language Use, respectively.

and responses align with the corresponding criteria. However, we observe minor issues in a small fraction of instances, including empty generations, code-switching, and language mismatch. We also identify cases where the target language appears to shape lexical or cultural choices in the response. For example, Japanese responses sometimes shift from Korean-standard terminology to expressions preferred in Japan. While this can be value-neutral when it reflects conventional localization (*e.g.*, using 'MEXT' instead of 'MOE'), we also observe substitutions in disputed, politically contested domains, where responses adopt Japan-aligned terminology (*e.g.*, 'Takeshima' instead of 'Dokdo' or using only 'Sea of Japan' while omitting 'East Sea'). In addition, some multilingual responses omit details of historically significant events (*e.g.*, 'March First Movement', a major rebellion against Japanese colonialism). These observations suggest that the target language may influence response faithfulness and completeness, particularly for culture-specific queries.

D.5 Experimental Setting for LLM Training using KCaQA

In Section 4.3, we train `Qwen 3 14B` using the following five datasets:

1. **CultureBank:** CultureBank (Shi et al., 2024) is a multilingual, multicultural dataset sourced from TikTok and Reddit. We extract the samples related to Korean culture.

2. **Korean LIMA:** Korean LIMA is a machine-translated version of the LIMA (Zhou et al., 2023) dataset, a high-quality, general-purpose open-domain QA.

3. **National Curriculum:** This is a source text of **KCaQA**. Here, we aim to examine the effectiveness of the **CuCu** framework by comparing its results with those of a model trained on **KCaQA**, a processed output of **CuCu** using the same national curriculum. [11]

4. **CuCu without source data:** Instead of using learning outcomes extracted from the national curriculum (Dataset 1), we ask `GPT-5.2` to generate learning outcomes by itself. Specifically, given non-cultural learning outcomes from other subjects as few-shot examples, we ask `GPT-5.2` to generate 15 learning outcomes for each of 10 culture topics (Korean history, geography, politics, law, economy, international relations, climate, traditions, social studies, and anthropology). Using these learning outcomes as source data for **CuCu**, we generate queries following steps 3–5 and their responses following steps 1–3.

5. **KCaQA (Ours):** We use Korean samples at an advanced level.

Datasets 1, 2, 4, and 5 are open-ended QA tailored for SFT, while Dataset 3 (national curriculum) is a plain corpus. To match the number of samples used in training, we randomly sample 1,030 instances of Datasets 2–4 with a seed of 42, following the number of the train set of Korean LIMA. We use accuracy for cultural knowledge evaluation, and follow the A-SVA setting of KorNAT, which asks the models to choose among three options: agree, neutral, and disagree.

[11] This national curriculum can be found at `https://ncic.re.kr/`.

E Detailed Evaluation Rubrics

The evaluation criteria are collaboratively designed by three authors, initially proposed by a single Korean L1 speaker with intermediate English proficiency and elementary proficiency in Chinese and Japanese, and subsequently reviewed and refined by two additional Korean L1 speakers.

Evaluation criteria for both human validation and LLM-as-a-judge

For the provided query and the generated response, check if they adhere to the following criteria and reject if there is any violence:

1. Language selection
   - Is the query or the response empty?
   - Is the query or the response code-mixed with multiple languages, or not written in the target language that is the same with the query?
2. Cultural appropriateness
   - Does the query or the response (especially those are not in Korean) adhere to the desired stance of MOFA (Ministry of Foreign Affairs, Republic of Korea) regarding some sensitive territory issues such as Dokdo (the official terminology that should be used instead of Takeshima) and East Sea (the official terminology that should be written in parallel with Japan Sea)?
   - Does the query or the response (especially those are not in Korean) include all important details regarding some historical events, *e.g.*, the governance and economics during Japanese colonial period?
3. Language use
   - Is the query or the response natural enough to be displayed a colloquial answer for the language speaker?
   - For Korean cases, is the synthesized query with/without country-specific expressions (*e.g.*, "our country") natural enough as a colloquial question?
   - Does the query or the response contain any terminology that corresponds to a similar concept in Korean, but that should not be transferred to other languages? (*e.g.*, the terminology for location, historical events, government departments, etc.)?

Throughout the manual inspection, we carefully review the language selection of non-English responses, as LLMs tend to respond in Korean regardless of the input language when the query includes culture-specific content or Korean-specific keywords. For cultural appropriateness, we refer to official announcements of MOFA (the Ministry of Foreign Affairs, Republic of Korea) [12], for some sensitive issues regarding territory and history (*e.g.*, the official terminology for Dokdo and East Sea/Japan Sea). We further examine whether non-Korean responses include dis/misinformation or blunt any details compared to Korean responses, especially regarding some sensitive, international historical events (*e.g.*, the governance and economics during the Japanese colonial period).

[12] https://www.mofa.go.kr/eng/wpge/m_5441/contents.do