

# 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. Applying **CuCu** to the Korean national social studies curriculum, we construct **KCaQA**, comprising 34.1k open-ended QA pairs. Our quantitative and qualitative analyses suggest that **KCaQA** covers culture-specific topics and produces responses grounded in local sociocultural contexts.<sup>1</sup>

## 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 (QA) supervision.

We introduce **CuCu** (from national **Curricula** to **Cultural 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 **Q**A, 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.

<sup>1</sup>**CuCu** and **KCaQA** are available at <https://anonymous.4open.science/r/cucu/>.

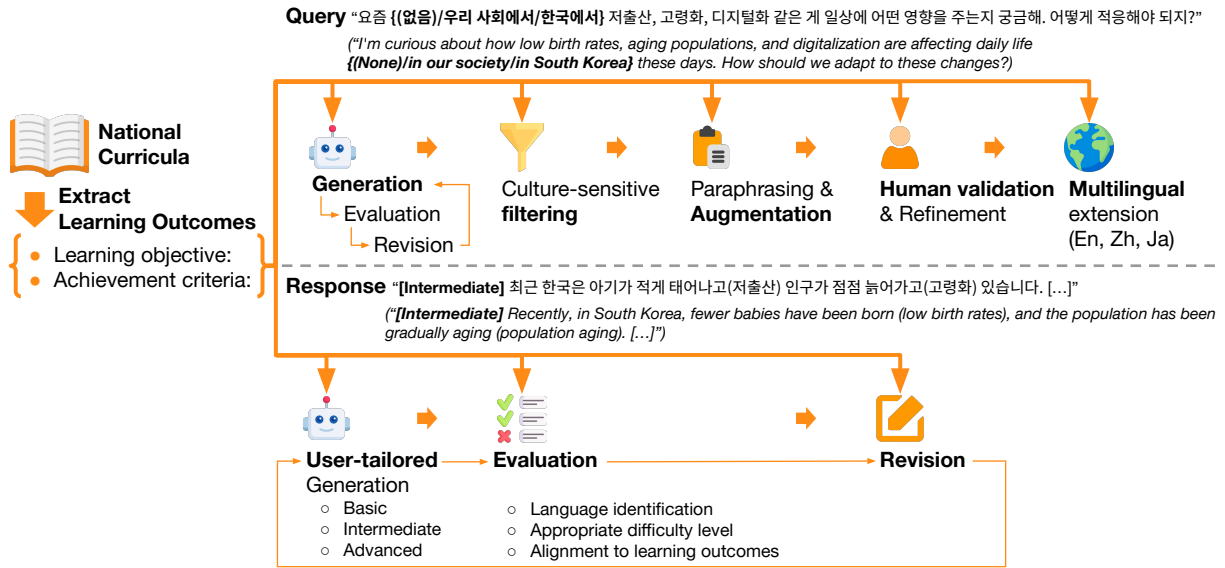


Figure 1: Overview of CuCu. CuCu constructs an open-ended culture-specific QA pairs using national curricula.

## 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

Regardless of the data creation being manual or AI-assisted, culture-specific NLP data construction has long been a topic of interest among diversity and inclusivity interest groups. Though culture incorporates a wide range of dimensions, including region and nation, ethnic, social groups, custom, etc. (Spencer-Oatey and Franklin, 2012), we mainly focus on social consensus that lies in a specific culture and language. This differs from approaches that focus on professional knowledge, and is rather close to approaches that aim at ensuring or benchmarking language models’ capability of understanding

customs and teachings (Li et al., 2024), norm (Rao et al., 2025), and common grounds (Tanwar et al., 2025) shared among cultural groups.

Especially for Korean, prior studies have tackled the evaluation of language model alignment considering cultural heritage (Kim et al., 2022), historical knowledge (Son et al., 2024), or social bias (Jin et al., 2024; Lee et al., 2023). Lee et al. (2024) has investigated social values and common knowledge regarding Korean among LLMs, but the dataset is close-ended and does not necessarily account for LLM-targeted questions, while Ko et al. (2025) only handles factuality issues instead of sociocultural norms and values. In this light, we aim to construct a framework and the following dataset that handles open-ended culture-specific questions in view of standard and refined source of value.

## 3 CuCu and KCaQA

We introduce CuCu (from national Curricula to Cultural 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.

**Query Generation.** We use Solar Pro 2 (Upstage, 2025) and GPT-4o<sup>2</sup> (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

<sup>2</sup>Model version: gpt-4o-2024-08-06

141	achievement criterion). We (1) generate a query,	query, (2) appropriateness of difficulty in terms	188
142	(2) evaluate whether it fully reflects the learning	of style/readability and content depth, and (3)	189
143	outcome, and (3) revise based on the feedback,	alignment with the corresponding learning out-	190
144	repeating this loop up to five times.	comes.	191
145	<b>2. Culture-sensitive filtering:</b> We filter out	<b>3. Response revision:</b> We revise the response and	192
146	queries that involve general, culture-agnostic	repeat this process up to five times.	193
147	knowledge, further improving their naturalness		
148	and language use.	To showcase <b>CuCu</b> , we use the 2022 Revised El-	194
149	<b>3. Paraphrasing &amp; Augmentation:</b> We para-	mentary, Middle, and High School National Cur-	195
150	phrase the initial query into two additional	riculum of Social Studies <sup>3</sup> , published by the Min-	196
151	queries that target the same learning outcomes	istry of Education, Republic of Korea, as a source	197
152	but differ in style and tone. Each resulting query	data. We extract 357 atomic pairs of learning ob-	198
153	is then expanded into three variants: (1) without	jectives and achievement criteria and filter out 158	199
154	mentioning any culture or country, (2) implicitly	learning outcomes, generating 2,844 queries ( <i>i.e.</i> , 3	200
155	referring to “our/my” culture or country, and (3)	paraphrasing × (3 Korean variants + 3 translations)	201
156	explicitly stating the name of a specific culture	× 158 learning outcomes). In total, we construct	202
157	or country.	<b>KCaQA (Korean Culturally-aware QA)</b> , comprising	203
158	<b>4. Human validation &amp; Refinement:</b> One of the	34,128 QA pairs ( <i>i.e.</i> , three different responses per	204
159	authors, who is a Korean L1 speaker born and	2,844 queries) that reflect Korean cultural knowl-	205
160	raised in Korea, who has been educated under	edge and perspective. We provide further details on	206
161	the Korean national curricula and thus is familiar	our design choice and discuss wider adaptability in	207
162	with its norms and culture, manually validates	Appendices B.1 and B.2.	208
163	and refines queries generated by multi-LLM	<b>4 Discussions</b>	209
164	agents. This process mainly includes (1) revis-	<b>4.1 KCaQA Quantitative Analysis</b>	210
165	ing queries to be natural colloquial questions,	<b>Culture-specific Query.</b> We analyze whether	211
166	(2) inspecting if the variants imply the same	<b>KCaQA</b> queries cover culture-specific topics us-	212
167	meaning regardless of the presence of culture	ing topic modeling. In particular, we compare	213
168	or country-related expressions, and (3) assuring	the Korean explicit queries in <b>KCaQA</b> with Korean	214
169	that queries adhere to the learning objective and	<b>LIMA (Zhou et al., 2023)</b> dataset <sup>4</sup> , a high-quality,	215
170	the achievement criteria, not handling topics that	general-domain QA dataset for LLM alignment	216
171	are overly generic or too professional.	training. The resulting distributions diverge sharply,	217
172	<b>5. Multilingual extension:</b> We finally extend the	with a JSD of 0.807 between <b>KCaQA</b> and Korean	218
173	queries into three additional languages ( <i>i.e.</i> , En-	<b>LIMA</b> ; a 1k permutation test confirms the differ-	219
174	glish, Chinese, and Japanese) by translating the	ence is significant ( $p < 0.01$ ). A qualitative inspec-	220
175	ones explicitly stating the culture/country name.	tion of the top-10 most skewed topics suggests that	221
176	<b>Response Generation.</b> We use four open-source	<b>KCaQA</b> mainly covers Korean sociocultural content	222
177	LLMs ( <i>i.e.</i> , GPT-OSS (OpenAI et al., 2025), Qwen3,	( <i>e.g.</i> , Korean geography, economy, history, climate,	223
178	Qwen3-Next (Yang et al., 2025), and DeepSeek	and politics), whereas Korean <b>LIMA</b> is dominated	224
179	R1 (DeepSeek-AI et al., 2025)) to generate re-	by general-purpose daily-life topics ( <i>e.g.</i> , STEM,	225
180	sponse texts.	recipes, and advice).	226
181	<b>1. User-tailored response generation:</b> For each	<b>User-tailored Response.</b> We examine whether	227
182	query, we produce three variants targeting differ-	<b>KCaQA</b> responses match their intended difficulty	228
183	ent audience bands— <i>Basic</i> , <i>Intermediate</i> , and	level ( <i>i.e.</i> , <i>Basic</i> , <i>Intermediate</i> , and <i>Advanced</i> ). Ta-	229
184	<i>Advanced</i> —to cover a range of readability and	ble 1 reports readability statistics across languages	230
185	assumed background knowledge.	in terms of response length and lexical diversity,	231
186	<b>2. Response evaluation:</b> We evaluate each can-	measured by the number of tokens, the number	232
187	didate for (1) language consistency with the		

<sup>3</sup><https://ncic.re.kr/>

<sup>4</sup>Translated version of LIMA dataset. Accessible at <https://huggingface.co/datasets/taeshahn/ko-lima>

of sentences, tokens per sentence, and the rare token ratio. In general, **KCaQA** responses become longer and syntactically denser with increasing target level, while lexical diversity also increases.

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 1: 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.

## 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 include the following dimensions, with detailed rubrics described in Appendix D.

- **Language selection:** If both query and response are monolingual in the target language;
- **Cultural appropriateness:** If both query and response adhere to the desired stance regarding culture-sensitive concepts;
- **Language use:** If the style and expressions of both query and response are appropriate and natural.

**LLM-as-a-Judge.** We employ GPT-5.2<sup>5</sup> (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

<sup>5</sup>Model version: gpt-5.2-2025-12-11

language use, respectively. These results suggest that **KCaQA** produced by **CuCu** largely satisfies the intended language constraints and stance requirements for culture-specific concepts. Full results are reported in Table 2 in Appendix B.3.

**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, 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.

## 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. This study provides a practical and reproducible path to constructing culturally grounded supervision for sovereign or culture-adapted LLMs.

## 308 Limitation

309 This study incorporates several conceptual and  
310 methodological limitations that require acknowl-  
311 edgment.

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

322 Second, our analysis focuses exclusively on  
323 Seoul Korean as used in the Republic of Korea  
324 (ROK) and references the standard curriculum of  
325 Republic of Korea. This scope excludes Pyongyang  
326 Korean as employed in the Democratic People’s  
327 Republic of Korea (DPRK) and does not consider  
328 North Korea’s educational standards. While we at-  
329 tempted to select queries and responses that would  
330 minimize controversial elements, certain aspects  
331 may not align with North Korean linguistic or edu-  
332 cational norms.

333 Third, our use of national curricula as primary  
334 sources for query and response generation, while  
335 ensuring verified and socially accepted content,  
336 does not fully capture the range of geocultural char-  
337 acteristics expressed through language. Curricula  
338 represent institutionalized knowledge rather than  
339 the complete spectrum of cultural expression. Nev-  
340 ertheless, our framework maintains sufficient flexi-  
341 bility to accommodate alternative source materials  
342 that may more comprehensively represent linguis-  
343 tic and cultural phenomena when such resources  
344 become available.

## 345 Broader Impact Statement

346 Our dataset construction process mainly consists of  
347 automated synthesis and validation, which is trans-  
348 parently shared to ensure reproducibility. Addition-  
349 ally, our study does not involve the recruitment  
350 or engagement of human subjects, thereby min-  
351 imizing subjective influence. While we conduct  
352 some manual inspection for qualitative analysis,  
353 we adopt the same predefined guidelines as those  
354 provided to LLMs.

355 We regard the national curricula of the Repub-  
356 lic of Korea as a reliable data source, which may  
357 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 ad-  
dressed. 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<sup>6</sup>, 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

- 374
- 375 Badr Alkhamissi, Muhammad ElNokrashy, Mai  
376 Alkhamissi, and Mona Diab. 2024. *Investigating  
377 cultural alignment of large language models*. In *Pro-  
378 ceedings of the 62nd Annual Meeting of the Associa-  
379 tion for Computational Linguistics (Volume 1: Long  
380 Papers)*, pages 12404–12422, Bangkok, Thailand. As-  
381 sociation for Computational Linguistics.
- 382 Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo,  
383 Thomas Wolf, and Leandro von Werra. 2024. *Cos-  
384 mopedia*.
- 385 Damian Blasi, Antonios Anastasopoulos, and Gra-  
386 ham Neubig. 2022. *Systematic inequalities in lan-  
387 guage technology performance across the world’s  
388 languages*. In *Proceedings of the 60th Annual Meet-  
389 ing of the Association for Computational Linguistics  
390 (Volume 1: Long Papers)*, pages 5486–5505, Dublin,  
391 Ireland. Association for Computational Linguistics.
- 392 Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and  
393 Iryna Gurevych. 2024. *Are multilingual LLMs  
394 culturally-diverse reasoners? an investigation into  
395 multicultural proverbs and sayings*. In *Proceedings  
396 of the 2024 Conference of the North American Chap-  
397 ter of the Association for Computational Linguistics:  
398 Human Language Technologies (Volume 1: Long Pa-  
399 pers)*, pages 2016–2039, Mexico City, Mexico. Asso-  
400 ciation for Computational Linguistics.
- 401 Chul-Ki Cho, HyeSook Kim, and Soyoung Lee. 2023.  
402 *Analysis of the issues that emerged in the revision of  
403 the national social studies curriculum in South Korea:  
404 Text mining and semantic network analysis of the  
405 comments at the public hearing on YouTube*. *Journal  
406 of Education and E-Learning Research*, 10(3):463–  
407 473.

<sup>6</sup>[https://en.wikipedia.org/wiki/Bias\\_in\\_curricula](https://en.wikipedia.org/wiki/Bias_in_curricula)

408	Eunjung Cho, Won Ik Cho, and Soomin Seo. 2025. <a href="#">Hermit kingdom through the lens of multiple perspectives: A case study of LLM hallucination on North Korea</a> . In <i>Proceedings of the 31st International Conference on Computational Linguistics</i> , pages 3353–3371, Abu Dhabi, UAE. Association for Computational Linguistics.	463
409		464
410		465
411		466
412		467
413		468
414		
415	Jacob Cohen. 1960. <a href="#">A coefficient of agreement for nominal scales</a> . <i>Educational and Psychological Measurement</i> , 20:37–46.	469
416		470
417		471
418	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. <a href="#">DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning</a> . <i>arXiv preprint arXiv:2501.12948</i> .	472
419		473
420		474
421		475
422		
423		476
424		477
425		478
426	Joseph L. Fleiss. 1971. <a href="#">Measuring nominal scale agreement among many raters</a> . <i>Psychological Bulletin</i> , 76(5):378–382.	479
427		480
428		481
429	Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. <a href="#">Discovering language-neutral sub-networks in multilingual language models</a> . In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 7560–7575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	482
430		483
431		484
432		485
433		
434		486
435		487
436		488
437	Maarten Grootendorst. 2022. <a href="#">BERTopic: Neural topic modeling with a class-based TF-IDF procedure</a> . <i>arXiv preprint arXiv:2203.05794</i> .	489
438		490
439		491
440	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. <a href="#">Textbooks are all you need</a> . <i>arXiv preprint arXiv:2306.11644</i> .	492
441		493
442		494
443		495
444		496
445		497
446		498
447		499
448	Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. <a href="#">CARE: Multilingual human preference learning for cultural awareness</a> . In <i>Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing</i> , pages 32854–32883, Suzhou, China. Association for Computational Linguistics.	500
449		501
450		502
451		
452		503
453		504
454		505
455	William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. <a href="#">A material lens on coloniality in NLP</a> . <i>arXiv preprint arXiv:2311.08391</i> .	506
456		507
457		508
458	Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. <a href="#">KoBBQ: Korean bias benchmark for question answering</a> . <i>Transactions of the Association for Computational Linguistics</i> , 12:507–524.	509
459		510
460		511
461		512
462		513
	Erik Jones, Hamid Palangi, Clarisse Simões Ribeiro, Varun Chandrasekaran, Subhabrata Mukherjee, Arindam Mitra, Ahmed Hassan Awadallah, and Ece Kamar. 2024. <a href="#">Teaching language models to hallucinate less with synthetic tasks</a> . In <i>The Twelfth International Conference on Learning Representations</i> .	514
		515
		516
		517
		518
		519
	Gyeongmin Kim, Jinsung Kim, Junyoung Son, and Heuiseok Lim. 2022. <a href="#">KoCHET: A Korean cultural heritage corpus for entity-related tasks</a> . In <i>Proceedings of the 29th International Conference on Computational Linguistics</i> , pages 3496–3505, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.	519
		520
		521
		522
		523
		524
		525
		526
		527
		528
		529
		530
		531
		532
		533
		534
		535
		536
		537
		538
		539
		540
		541
		542
		543
		544
		545
		546
		547
		548
		549
		550
		551
		552
		553
		554
		555
		556
		557
		558
		559
		560
		561
		562
		563
		564
		565
		566
		567
		568
		569
		570
		571
		572
		573
		574
		575
		576
		577
		578
		579
		580
		581
		582
		583
		584
		585
		586
		587
		588
		589
		590
		591
		592
		593
		594
		595
		596
		597
		598
		599
		600
		601
		602
		603
		604
		605
		606
		607
		608
		609
		610
		611
		612
		613
		614
		615
		616
		617
		618
		619
		620
		621
		622
		623
		624
		625
		626
		627
		628
		629
		630
		631
		632
		633
		634
		635
		636
		637
		638
		639
		640
		641
		642
		643
		644
		645
		646
		647
		648
		649
		650
		651
		652
		653
		654
		655
		656
		657
		658
		659
		660
		661
		662
		663
		664
		665
		666
		667
		668
		669
		670
		671
		672
		673
		674
		675
		676
		677
		678
		679
		680
		681
		682
		683
		684
		685
		686
		687
		688
		689
		690
		691
		692
		693
		694
		695
		696
		697
		698
		699
		700
		701
		702
		703
		704
		705
		706
		707
		708
		709
		710
		711
		712
		713
		714
		715
		716
		717
		718
		719
		720
		721
		722
		723
		724
		725
		726
		727
		728
		729
		730
		731
		732
		733
		734
		735
		736
		737
		738
		739
		740
		741
		742
		743
		744
		745
		746
		747
		748
		749
		750
		751
		752
		753
		754
		755
		756
		757
		758
		759
		760
		761
		762
		763
		764
		765
		766
		767
		768
		769
		770
		771
		772
		773
		774
		775
		776
		777
		778
		779
		780
		781
		782
		783
		784
		785
		786
		787
		788
		789
		790
		791
		792
		793
		794
		795
		796
		797
		798
		799
		800
		801
		802
		803
		804
		805
		806
		807
		808
		809
		810
		811
		812
		813
		814
		815
		816
		817
		818
		819
		820
		821
		822
		823
		824
		825
		826
		827
		828
		829
		830
		831
		832
		833
		834
		835
		836
		837
		838
		839
		840
		841
		842
		843
		844
		845
		846
		847
		848
		849
		850
		851
		852
		853
		854
		855
		856
		857
		858
		859
		860
		861
		862
		863
		864
		865
		866
		867
		868
		869
		870
		871
		872
		873
		874
		875
		876
		877
		878
		879
		880
		881

520	Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. <a href="#">Bactrian-X: Multilingual replicable instruction-following models with low-rank adaptation</a> . <i>arXiv preprint arXiv:2305.15011</i> .	576
521		577
522		578
523		579
524		
525	Pratyush Maini, Vineeth Dorna, Parth Doshi, Aldo Carranza, Fan Pan, Jack Urbanek, Paul Burstein, Alex Fang, Alvin Deng, Amro Abbas, and 1 others. 2025. <a href="#">BeyondWeb: Lessons from scaling synthetic data for trillion-scale pretraining</a> . <i>arXiv preprint arXiv:2508.10975</i> .	580
526		581
527		582
528		583
529		584
530		585
531	Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. <a href="#">Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions</a> . In <i>Proceedings of the 31st International Conference on Computational Linguistics</i> , pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.	586
532		587
533		588
534		589
535		590
536		
537		
538		
539	OpenAI. 2025. <a href="#">GPT-5 System Card</a> .	
540	OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. <a href="#">GPT-4 technical report</a> . <i>arXiv preprint arXiv:2303.08774</i> .	
541		
542		
543		
544		
545		
546		
547		
548	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, Sebastian Bubeck, Che Chang, and 107 others. 2025. <a href="#">gpt-oss-120b &amp; gpt-oss-20b model card</a> . <i>arXiv preprint arXiv:2508.10925</i> .	
549		
550		
551		
552		
553		
554		
555		
556	Suzi Park and Hyopil Shin. 2021. Kr-sbert: A pre-trained korean-specific sentence-bert model. <a href="https://github.com/snunlp/KR-SBERT">https://github.com/snunlp/KR-SBERT</a> .	
557		
558		
559	Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. <a href="#">A survey of multilingual large language models</a> . <i>Patterns</i> , 6(1):101118.	
560		
561		
562		
563	Neel Prabhajan Rachamalla, Aravind Konakalla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. 2025. <a href="#">Pragyaan: Designing and curating high-quality cultural post-training datasets for Indian languages</a> . In <i>Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)</i> , pages 285–321, Suzhuo, China. Association for Computational Linguistics.	
564		
565		
566		
567		
568		
569		
570		
571	Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. <a href="#">NormAd: A framework for measuring the cultural adaptability of large language models</a> . In <i>Proceedings of the 2025 Conference of the Nations of the Americas</i>	
572		
573		
574		
575		
	<i>Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.	576
		577
		578
		579
	Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Sneha 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. <a href="#">INCLUDE: Evaluating multilingual language understanding with regional knowledge</a> . In <i>The Thirteenth International Conference on Learning Representations</i> .	580
		581
		582
		583
		584
		585
		586
		587
		588
		589
		590
	Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024. <a href="#">HAE-RAE bench: Evaluation of Korean knowledge in language models</a> . In <i>Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)</i> , pages 7993–8007, Torino, Italia. ELRA and ICCL.	591
		592
		593
		594
		595
		596
		597
		598
	Robyn Speer. 2022. <a href="#">rspeer/wordfreq: v3.0</a> .	599
	Helen Spencer-Oatey and Peter Franklin. 2012. What is culture. <i>A compilation of quotations. GlobalPAD Core Concepts</i> , 1(22):1–21.	600
		601
		602
	Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang Wang, and Tanmoy Chakraborty. 2025. <a href="#">Do you know about my nation? investigating multilingual language models’ cultural literacy through factual knowledge</a> . In <i>Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing</i> , pages 14967–14990, Suzhou, China. Association for Computational Linguistics.	603
		604
		605
		606
		607
		608
		609
		610
		611
	Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. <a href="#">Cultural bias and cultural alignment of large language models</a> . <i>PNAS Nexus</i> , 3(9):pgae346.	612
		613
		614
	Upstage. 2025. <a href="#">Solar Pro 2</a> .	615
	Zifeng Wang, Chun-Liang Li, Vincent Perot, Long Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. 2024. <a href="#">CodeLM: Aligning language models with tailored synthetic data</a> . In <i>Findings of the Association for Computational Linguistics: NAACL 2024</i> , pages 3712–3729, Mexico City, Mexico. Association for Computational Linguistics.	616
		617
		618
		619
		620
		621
		622
	Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. <a href="#">Self-pluralising culture alignment for large language models</a> . In <i>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)</i> , pages 6859–6877, Albuquerque, New Mexico. Association for Computational Linguistics.	623
		624
		625
		626
		627
		628
		629
		630

631 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang,  
632 Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,  
633 Chengen Huang, Chenxu Lv, Chujie Zheng, Dayi-  
634 heng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge,  
635 Haoran Wei, Huan Lin, Jialong Tang, and 41 oth-  
636 ers. 2025. [Qwen3 technical report](#). *arXiv preprint*  
637 *arXiv:2505.09388*.

638 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan  
639 Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,  
640 Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang,  
641 Joseph E Gonzalez, and Ion Stoica. 2023. [Judging](#)  
642 [LLM-as-a-judge with MT-bench and chatbot arena](#).  
643 In *Advances in Neural Information Processing Sys-*  
644 *tems*, volume 36, pages 46595–46623. Curran Asso-  
645 ciates, Inc.

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

653 Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun,  
654 Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin,  
655 António Branco, and Deyi Xiong. 2024. [Multilingual](#)  
656 [large language models: A systematic survey](#). *arXiv*  
657 *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 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.

### B Additional Discussions

#### B.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

**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.

#### B.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 gov-

ernments or educational authorities and are publicly available—e.g., United Kingdom<sup>7</sup>, India<sup>8</sup>, Bangladesh<sup>9</sup>, and Australia<sup>10</sup>. 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.

**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.

### B.3 Quality Estimation of KCaQA

Table 2 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

<sup>7</sup><https://www.gov.uk/government/collections/national-curriculum>

<sup>8</sup>[http://www.ncert.nic.in/rightside/links/nc\\_framework.html](http://www.ncert.nic.in/rightside/links/nc_framework.html)

<sup>9</sup><https://nctb.gov.bd/>

<sup>10</sup><https://www.australiancurriculum.edu.au/>

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 2: 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.

in English, annotates the samples using the same rubric as the LLM-as-a-Judge (Appendix D). 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.

### 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 변화되

는지 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 장

려금·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 child-birth 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 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)<sup>11</sup>, 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

<sup>11</sup>[https://www.mofa.go.kr/eng/wpge/m\\_5441/contents.do](https://www.mofa.go.kr/eng/wpge/m_5441/contents.do)

793 details compared to Korean responses, especially  
794 regarding some sensitive, international historical  
795 events (*e.g.*, the governance and economics during  
796 the Japanese colonial period).