K-NLPers at SemEval-2026 Task 7: Multiple LLM Agent Debate System for Everyday Knowledge Across Diverse Languages and Cultures

Jiwoo Song*, **Sihyeong Yeom*** **and Harksoo Kim**

Department of Artificial Intelligence, Konkuk University

Correspondence: nlpdrkim@konkuk.ac.kr

Abstract

This paper presents the K-NLPers system for SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures. The task extends the BLEnD benchmark to evaluate cultural understanding of language models across more than 30 language-country pairs. Although Large Language Models (LLMs) achieve strong overall performance, they exhibit performance disparities across cultural contexts and tend to produce regionally biased responses. To address this limitation, we propose a continent-based multi-agent debate framework that leverages culture-specific performance differences instead of relying on a single model. For the Short Answer Question (SAQ) track, we employ three agents: a general-purpose model, a continent-specific model, and a country-level or culturally adjacent model. These agents engage in independent generation, mutual refinement, and final adjudication. For the Multiple-Choice Question (MCQ) track, we adopt a debate structure centered on high-performing general-purpose models due to the track's simpler structure. Our system participated in all language-region pairs and achieved overall scores of 55.75 on SAQ and 88.32 on MCQ. Further analysis reveals that grouping the performance of various individual models by continent explains performance patterns more consistently than language-based grouping, highlighting the importance of cultural and historical context in model generalization.

1 Introduction

Large Language Models (LLMs) achieve strong performance across natural language understanding and generation tasks (Zhao et al., 2026; Baek et al., 2025), yet they remain limited in capturing cultural context (Lee et al., 2025). Due to pretraining data concentrated on specific regions and high-resource languages, LLMs often overrepresent Western-centric perspectives and underrepresent low-resource languages and regions. This limitation highlights the need for structural approaches that better accommodate cultural diversity.

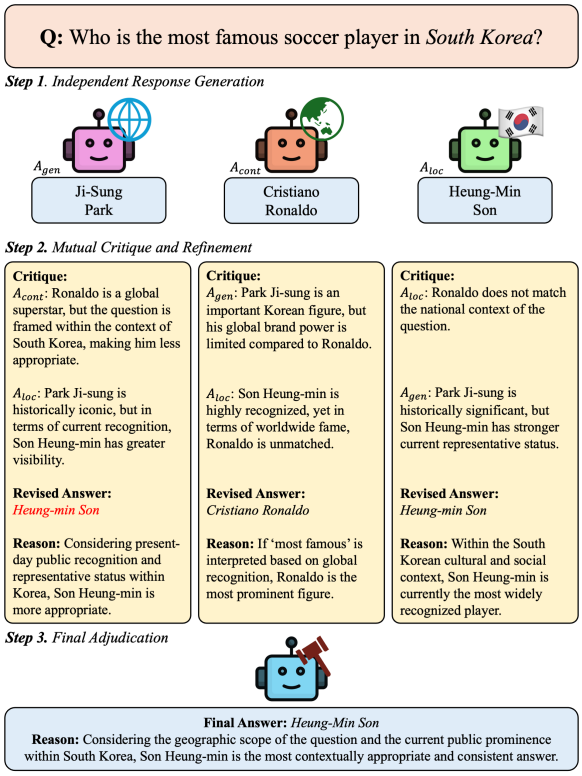


Figure 1: Overview of the multi-agent debate framework for the SAQ track. The framework consists of three stages–independent response generation, mutual critique and refinement, and final adjudication. Multiple agents first produce initial answers, revise them after reviewing peer responses, and a judge model then selects the final output.

SemEval-2026 Task 7: Everyday Knowledge Across Diverse Languages and Cultures (Ghosh et al., 2026; Ousidhoum et al., 2026) extends the BLEnD (Myung et al., 2024) benchmark to evaluate culturally grounded everyday knowledge across more than 30 language-country pairs. The SAQ

* These authors contributed equally to this work.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3280–3296

track requires the open-ended generation of culturally appropriate responses, while the MCQ track evaluates the selection of culturally appropriate answers from multiple options.

We begin from the observation that a single model does not perform uniformly across cultural contexts. For SAQ, we employ three agents: a general-purpose model, a continent-specific model, and a country-level or culturally adjacent model. These agents derive the final answer through independent generation, mutual refinement, and final decision-making. For MCQ, we maintain the same debate structure but use high-performing general-purpose models to ensure stable predictions.

The contributions of this paper are as follows:

- We propose a multi-LLM debate-based inference framework for modeling cultural context.
- We systematically analyze country- and region-level performance differences, demonstrating culture-specific model strengths.
- We design differentiated strategies for SAQ and MCQ and experimentally validate their effectiveness.

2 Related Work

2.1 Cultural Bias and Multilingual Cultural Evaluation Benchmarks

As LLMs become globally deployed, understanding cultural context has emerged as a critical challenge. BLEnD (Myung et al., 2024) evaluated everyday knowledge across 16 countries and demonstrated that country-specific knowledge differences persist even within the same language group. KoBBQ (Jin et al., 2024) further showed that cultural context cannot be preserved through simple translation.

While these benchmarks diagnose cultural bias, they do not propose mechanisms to mitigate performance imbalances. In contrast, our work incorporates empirically observed culture-specific model strengths into a multi-agent debate framework to reduce cross-cultural performance gaps.

2.2 Multi-Agent Debate and Expert-Integrated LLM Collaboration

Multi-agent debate frameworks have been proposed to improve LLM reasoning (Ki et al., 2025). Prior work shows that independent generation followed by iterative cross-review can outperform single-model baselines (Liang et al., 2024). However, debate may also reinforce shared misconceptions or converge to majority bias when models exhibit similar inductive biases (Estornell and Liu, 2024). To address this issue, intervention strategies such as *Misconception Refutation* have been introduced.

We extend the multi-agent debate paradigm to culturally grounded evaluation. By constructing an agent pool that reflects task characteristics and culture-specific model strengths, we design a three-stage architecture consisting of independent response generation, mutual critique & refinement, and final adjudication.

3 System Overview

Our system is a multi-LLM debate-based inference framework for culturally grounded response generation. It consists of three stages: (1) independent response generation, (2) mutual critique and refinement, and (3) final adjudication. Figure 1 illustrates the overall debate process for the SAQ track.

3.1 Multi-Agent Debate Framework

Given a question q, we construct a set of N agents A. Each agent A_i generates responses through the following procedure:

$$A = \{A_1, A_2, \cdots, A_N\}. \tag{1}$$

Independent Response Generation. Each agent independently produces an initial response $r_i^{(0)}$ to the question q:

$$r_i^{(0)} = A_i(q). \tag{2}$$

At this stage, agents do not access each other's outputs.

Mutual Critique and Refinement. Each agent receives the set of initial responses. $R^{(0)}$ from the other agents and generates a revised response $r_i^{(1)}$ by reviewing and refining them:

$$R^{(0)} = \{r_1^{(0)}, r_2^{(0)}, \cdots, r_N^{(0)}\}. \tag{3}$$

$$r_i^{(1)} = A_i\big(q, R^{(0)} \setminus r_i^{(0)}\big). \tag{4}$$

This step compensates for missing cultural context and mitigates individual model biases.

Final Adjudication. A judge model J determines the final answer $\hat{y}$ based on the question q and the revised responses:

Model	Africa		Middle East		Europe			Americas		East Asia			SE Asia		overall
	am-ET	ha-NG	ar-SA	fa-IR	fr-FR	en-GB	es-ES	en-US	es-MX	zh-CN	ja-JP	ko-KR	id-ID	tl-PH	
gpt-4o-mini	29.6	29.8	**51.8**	**63.8**	70.4	72.4	62.0	75.8	58.6	66.8	60.0	64.8	67.4	**62.6**	52.97
gemma-3-27b-it	25.4	27.0	46.4	61.6	64.6	71.8	56.8	73.4	49.0	69.0	58.4	71.2	70.6	56.0	49.14
Qwen2.5-14B-Instruct	14.0	10.4	40.0	39.2	66.2	**73.6**	62.8	**77.4**	60.2	73.6	59.0	56.8	63.8	38.6	45.26
EuroLLM-22B-Instruct	2.2	5.6	41.6	29.0	70.0	72.4	66.4	76.0	**61.0**	68.2	53.0	63.6	49.4	29.4	46.10
Ours	**50.6**	**40.8**	46.8	54.0	**71.6**	72.6	**69.6**	75.6	59.6	**78.0**	**63.6**	**78.0**	**73.2**	60.0	**55.75**

Table 1: SAQ performance of major single-model baselines and the proposed framework across representative locales from each continent. Performance differences are observed across locales, with general-purpose and region-specific models exhibiting varying strengths depending on cultural context. The proposed multi-agent debate framework integrates these complementary tendencies, leading to more consistent performance across culturally diverse settings.

$$R^{(1)} = \{r_1^{(1)}, r_2^{(1)}, \cdots, r_N^{(1)}\}. \tag{5}$$

$$\hat{y} = J(q, R^{(1)}). \tag{6}$$

The judge selects the final output by evaluating consistency, validity, and cultural appropriateness. In the case of MCQ, unlike SAQ which involves open-ended outputs, the output structure is constrained to a fixed set of answer options. Therefore, when the agents produce the same final choice (i.e., select the same option number), we omit the use of a Judge LLM to improve computational efficiency.

3.2 Agent Configuration for SAQ

The SAQ track is an open-ended generation problem that requires culturally grounded responses. To reflect culture-specific strengths, we construct a three-agent configuration.

The agent set is define as follows:

$$A_{\text{SAQ}} = \{A_{\text{gen}}, A_{\text{cont}}, A_{\text{loc}}\}. \tag{7}$$

- A_{gen}**:** a high-performing general-purpose model with stable overall performance.
- A_{cont}**:** a model developed within the continent of the target country.
- A_{loc}**:** a country-specific model or one developed in a culturally adjacent region.

This configuration reflects model- and region-level performance differences identified in our preliminary analysis (Section 5.3), where continent- or country-developed models showed relatively higher scores on related regional questions.

The three agents derive the final response through the three-stage debate process described in Section 3.1.

3.3 Agent Configuration for MCQ

The MCQ track is a multiple-choice classification problem. Unlike SAQ, it evaluates the selection of culturally appropriate answers from predefined candidates. Accordingly, we do not employ culture-specific agent specialization.

While retaining the same multi-agent debate framework, we construct the agent set using high-performing general-purpose LLMs with stable overall performance. Given the structural simplicity and constrained response space of MCQ, the focus is on improving decision consistency rather than generative diversity.

4 Experimental Setup

Model. We perform no additional training or fine-tuning. All experiments are conducted in an inference-only setting using pretrained LLMs as-is.

For the SAQ track, we use ten LLMs to construct the multi-agent debate framework. For each question, three debater agents are instantiated, with `gpt-4o-mini` fixed as the A_{gen} due to its stable performance. The remaining two agents are selected from models developed within the continent or country corresponding to the target locale. We use `gemma3-27b-it` as the judge model for SAQ. The country-level routing table is provided in Appendix A. The models used in our experiments are as follows:

- gpt-4o-mini (OpenAI, 2024)
- gemma-3-27b-it (Team et al., 2025)
- Llama-3.1-8B-Instruct (Grattafiori et al., 2024)
- Midm-2.0-Base-Instruct (Shin et al., 2026)
- Mistral-Nemo-Japanese-Instruct-2408 (Ishigami, 2024)
- Qwen2.5-14B-Instruct (Team, 2024b)
- Qwen-SEA-LION-v4-32B-IT (Singapore, 2024)
- Falcon3-10B-Instruct (Team, 2024a)

- Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)
- EuroLLM-22B-Instruct-2512 (Ramos et al., 2026)

For MCQ, we construct the agent set using high-performing general-purpose LLMs. Specifically, `gpt-4o-mini` (OpenAI, 2024) and `claude-3.5-haiku` (Anthropic, 2024) serve as debaters, and `gpt-4o`(OpenAI et al., 2024) serves as the judge.

Implementation Details. Open-source LLM experiments were conducted on two NVIDIA A6000 GPUs. Closed-source models were accessed via their official APIs. For all models, the temperature was set to 0 to ensure deterministic, fact-based responses.

5 Results & Analysis

5.1 SAQ Result

Table 1 reports performance on representative locales from each continent. The overall column reflects the official score computed across all SAQ locales rather than the average of representative ones. Detailed results are provided in Appendix B.

Our method achieved an overall score of 55.75, improving by approximately 2.8 points over the best single model, `gpt-4o-mini` (52.97). This confirms that the multi-agent debate framework yields consistent gains over single-model baselines in open-ended culturally grounded generation tasks.

Improvements were most pronounced in low-resource languages. In locales such as Amharic (am-ET) and Hausa (ha-NG), where single-model performance was low, Ours achieved substantial gains. In contrast, improvements were limited for high-resource languages such as English (en-US, en-GB), where baseline performance was already strong. This suggests that multi-agent debate enhances robustness in settings with larger performance inconsistencies.

Including locally or regionally developed models as debate agents further contributed to performance gains. For example, `Qwen2.5` was used for Chinese, `Midm` for Korean, and `Mistral` and `EuroLLM` for European languages, leveraging region-specific strengths. By combining regional cultural signals with general-purpose stability, the proposed method maintains strong performance across diverse cultural contexts. However, in regions such as the Middle East (ar-SA, fa-IR), where low-performing models were included as agents, they appeared to introduce noise rather than complementary cultural signals, resulting in performance degradation. This highlights the importance of agent quality in the debate framework.

Locale/Region	$S(M_1)$	$S(M_2)$	D	Δ_1	Δ_2
am-ET	**68.46**	61.82	66.36	-2.10	+4.54
ar-DZ	91.27	90.73	**92.23**	+0.96	+1.50
ar-EG	90.22	89.67	**91.03**	+0.81	+1.36
ar-MA	77.53	79.19	**80.85**	+3.32	+1.66
ar-SA	76.58	79.28	**81.31**	+4.73	+2.03
as-AS	76.70	77.97	**78.95**	+2.25	+0.98
az-AZ	80.37	77.80	**81.28**	+0.91	+3.48
bg-BG	**96.60**	95.06	95.68	-0.92	+0.62
el-GR	**93.93**	92.39	**93.93**	0.00	+1.54
en-AU	92.20	88.11	**92.59**	+0.39	+4.48
en-GB	94.97	94.32	**95.80**	+0.83	+1.48
en-US	95.37	94.64	**95.62**	+0.25	+0.98
es-EC	98.46	**98.57**	**98.57**	+0.11	0.00
es-ES	94.30	91.09	**95.18**	+0.88	+4.09
es-MX	**94.21**	93.42	94.10	-0.11	+0.68
eu-PV	87.53	86.60	**90.79**	+3.26	+4.19
fa-IR	77.45	77.70	**79.10**	+1.65	+1.40
fr-FR	**96.09**	93.49	**96.09**	0.00	+2.60
ga-IE	90.07	89.02	**91.24**	+1.17	+2.22
ha-NG	79.18	73.75	**79.43**	+0.25	+5.68
id-ID	84.56	86.52	**87.77**	+3.21	+1.25
ja-JP	83.17	**85.85**	**85.85**	+2.68	0.00
ko-KP	**73.59**	71.08	73.50	-0.09	+2.42
ko-KR	88.22	**89.49**	90.13	+1.91	+0.64
su-JB	83.75	78.38	**87.55**	+3.80	+9.17
sv-SE	82.55	86.13	**87.47**	+4.92	+1.34
ta-LK	**96.95**	92.73	95.96	-0.99	+3.23
tl-PH	87.72	87.26	**88.47**	+0.75	+1.21
zh-CN	87.82	88.34	**90.88**	+3.06	+2.54
zh-SG	91.36	90.65	**91.82**	+0.46	+1.17
Overall	85.67	84.30	**88.32**	+2.65	+4.02

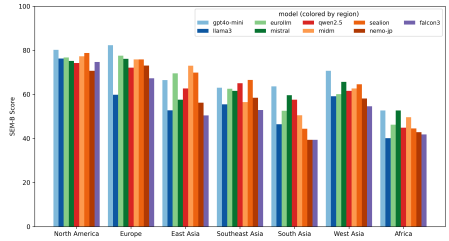
Table 2: Main results for Track 2 (MCQ). Comparison of single-model accuracies $S(M_1)$ and $S(M_2)$ with the final performance D achieved by the debate-based framework across locales/regions. Here, M_1 and M_2 denote `gpt-4o-mini` and `claude-3.5-haiku`, respectively. Δ_1 and Δ_2 represent $D-S(M_1)$ and $D-S(M_2)$, respectively.

5.2 MCQ Result

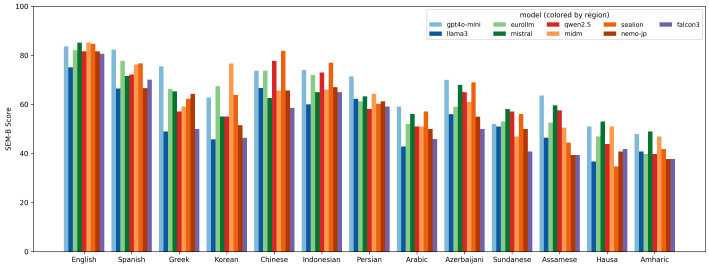
Table 2 reports locale-level performance and overall averages for the MCQ track. The proposed multi-agent debate model achieved an overall accuracy of 88.32, improving by 2.65 points over `gpt-4o-mini` (85.67) and 4.02 points over `claude-3.5-haiku` (84.30).

While slight decreases were observed in a few locales (e.g., am-ET, bg-BG), most locales showed positive gains. This indicates that multi-agent feedback mitigates individual model biases and leads to more stable decisions.

Although MCQ does not use culture-specific agent specialization, the debate structure itself functions as a mechanism for mutual error correction during option comparison.



(a) Performance analysis by Continental Grouping.



(b) Performance analysis by Language Grouping.

Figure 2: Agent selection analysis for SAQ. The results compare SEM-B scores across different continental and linguistic groupings to determine the optimal agent configuration.

5.3 Agent Selection Analysis for SAQ

To determine the SAQ agent configuration, we analyzed single-model performance across language-country pairs. From the BLEnD (Myung et al., 2024) dataset, we randomly selected 150 samples per country and evaluated them according to the official evaluation protocol. In accordance with SemEval-2026 Task 7 regulations, the data were used solely for analysis and not for training or fine-tuning.

We compared the average single-model performance across groups under two grouping criteria: (1) continental grouping, (2) language grouping.

Continental Grouping. When aggregated by continent, models developed within the corresponding continent showed higher average performance, particularly in Europe and East Asia. For example, `EuroLLM` performed strongly in European regions, and `Qwen2.5` in Chinese contexts. In Africa, European-based models also showed relatively strong performance, possibly reflecting historical and linguistic connections.

Language Grouping. Under language-based grouping, country-specific or culturally adjacent models performed well in single-country-centered languages (e.g., Korean, Chinese, Indonesian). However, for languages spoken across multiple countries (e.g., English, Spanish), no consistent country-level dominance was observed, and performance varied within the same language.

Overall, language grouping explained certain localized advantages but lacked consistent global patterns. In contrast, continental grouping yielded more stable and interpretable trends. Accordingly, we adopted continent-based agent selection for SAQ.

6 Conclusion

We proposed a multi-LLM debate-based inference framework to address the limitations of single models in culturally grounded multilingual settings. Based on observed model- and country-level performance variations, we adopted continent- and region-based agent selection for SAQ, while using high-performing general-purpose models for MCQ to ensure stable classification.

The proposed method achieved overall scores of 55.75 on SAQ and 88.32 on MCQ. The multi-agent debate framework improved culturally grounded scores over single-model baselines, particularly in open-ended generation. These results highlight the importance of structurally incorporating cultural diversity in multilingual evaluation benchmarks.

7 Ethical Considerations

Our study does not involve new data collection or human subjects, but continent-level grouping may still simplify cultural diversity and should therefore be interpreted with care.

Limitations

We employ a static agent configuration for both SAQ and MCQ. Recent routing and orchestration approaches dynamically adjust model selection or collaboration based on query characteristics (e.g., MASRouter (Yue et al., 2025); Agent-Oriented Planning (Li et al., 2025)). The absence of such dynamic mechanisms remains a limitation of our framework.

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00343989, Enhancing the

Ethics of Data Characteristics and Generation AI Models for Social and Ethical Learning).

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00553041, Enhancement of Rational and Emotional Intelligence of Large Language Models for Implementing Dependable Conversational Agents).

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Hye-Yoon Baek, Jinho Choi, Jimyeung Seo, Xiongnan Jin, Dongcheon Lee, and Byungkook Oh. 2025. Relation-faceted graph pooling with llm guidance for dynamic span-aware information extraction. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 109–118, New York, NY, USA. Association for Computing Machinery.

Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principals, and interventions. In *Advances in Neural Information Processing Systems*, volume 37, pages 28938–28964. Curran Associates, Inc.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ryosuke Ishigami. 2024. Mistral-nemo-japanese-instruct-2408.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple LLM agents debate for equitable cultural alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24841–24877, Vienna, Austria. Association for Computational Linguistics.

Woojin Lee, Yujin Sim, Hongjin Kim, and Harksoo Kim. 2025. Multilingual, not multicultural: Uncovering the cultural empathy gap in LLMs through a comparative empathetic dialogue benchmark. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 791–809, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Ao Li, Yuexiang Xie, Songze Li, Fugee Tsung, Bolin Ding, and Yaliang Li. 2025. Agent-oriented planning in multi-agent systems. In *The Thirteenth International Conference on Learning Representations*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

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

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

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

Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. 2026. Eurollm-22b: Technical report. *Preprint*, arXiv:2602.05879.

Donghoon Shin, Sejung Lee, Soonmin Bae, Hwijung Ryu, Changwon Ok, Hoyoun Jung, Hyesung Ji, Jeehyun Lim, Jehoon Lee, Ji-Eun Han, Jisoo Baik, Mihyeon Kim, Riwoo Chung, Seongmin Lee, Wonjae Park, Yoonseok Heo, Youngkyung Seo, Seyoun Won, Boeun Kim, and 47 others. 2026. Mi:dm 2.0 korea-centric bilingual language models. *Preprint*, arXiv:2601.09066.

AI Singapore. 2024. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`.

Falcon-LLM Team. 2024a. The falcon 3 family of open models.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. 2025. MasRouter: Learning to route LLMs for multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15549–15572, Vienna, Austria. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2026. A survey of large language models. *Preprint*, arXiv:2303.18223.

A SAQ Agent Routing Table by Continent and Country

For the SAQ track, we adopt a continent-based agent routing strategy to reflect cultural characteristics. Preliminary experiments analyzing single-model performance revealed significant disparities across countries and regions. We found that grouping these performance results by continent yielded more consistent patterns than language-based grouping. Accordingly, to mitigate the limitations of a single model, we construct a three-agent configuration for each locale consisting of a general-purpose model (A_{gen}), a continent-specific model (A_{cont}), and a country-level or culturally adjacent model (A_{loc}).

Locale	A_{gen}	A_{cont}	A_{loc}
en-US	gpt-4o-mini	Llama-3.1-8B-Instruct	gemma-3-27b-it
en-AU	gpt-4o-mini	Llama-3.1-8B-Instruct	gemma-3-27b-it
en-GB	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Mistral-7B-Instruct-v0.3
es-MX	gpt-4o-mini	Llama-3.1-8B-Instruct	gemma-3-27b-it
en-MX	gpt-4o-mini	Llama-3.1-8B-Instruct	gemma-3-27b-it
es-EC	gpt-4o-mini	Llama-3.1-8B-Instruct	gemma-3-27b-it
en-EC	gpt-4o-mini	Llama-3.1-8B-Instruct	gemma-3-27b-it
eu-PV	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
en-PV	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
bg-BG	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
en-BG	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
fr-FR	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Mistral-7B-Instruct-v0.3
en-FR	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Mistral-7B-Instruct-v0.3
el-GR	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
en-GR	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
ga-IE	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
en-IE	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
es-ES	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
en-ES	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
sv-SE	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
en-SE	gpt-4o-mini	Mistral-7B-Instruct-v0.3	EuroLLM-22B-Instruct-2512
zh-CN	gpt-4o-mini	Qwen-SEA-LION-v4-32B-IT	Qwen2.5-14B-Instruct
en-CN	gpt-4o-mini	Qwen-SEA-LION-v4-32B-IT	Qwen2.5-14B-Instruct
ja-JP	gpt-4o-mini	Qwen2.5-14B-Instruct	Mistral-Nemo-Japanese-Instruct-2408
en-JP	gpt-4o-mini	Qwen2.5-14B-Instruct	Mistral-Nemo-Japanese-Instruct-2408
ko-KP	gpt-4o-mini	Qwen2.5-14B-Instruct	Midm-2.0-Base-Instruct
en-KP	gpt-4o-mini	Qwen2.5-14B-Instruct	Midm-2.0-Base-Instruct
ko-KR	gpt-4o-mini	Qwen2.5-14B-Instruct	Midm-2.0-Base-Instruct
en-KR	gpt-4o-mini	Qwen2.5-14B-Instruct	Midm-2.0-Base-Instruct
zh-TW	gpt-4o-mini	Qwen-SEA-LION-v4-32B-IT	Qwen2.5-14B-Instruct
en-TW	gpt-4o-mini	Qwen-SEA-LION-v4-32B-IT	Qwen2.5-14B-Instruct
id-ID	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
en-ID	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
su-JB	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
en-JB	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
tl-PH	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
en-PH	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
ms-SG	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
ta-SG	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
zh-SG	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
en-SG	gpt-4o-mini	Qwen2.5-14B-Instruct	Qwen-SEA-LION-v4-32B-IT
az-AZ	gpt-4o-mini	Qwen2.5-14B-Instruct	gemma-3-27b-it
en-AZ	gpt-4o-mini	Qwen2.5-14B-Instruct	gemma-3-27b-it
as-AS	gpt-4o-mini	Qwen2.5-14B-Instruct	gemma-3-27b-it
en-AS	gpt-4o-mini	Qwen2.5-14B-Instruct	gemma-3-27b-it
ta-LK	gpt-4o-mini	Qwen2.5-14B-Instruct	gemma-3-27b-it

Locale	A_{gen}	A_{cont}	A_{loc}
en-LK	gpt-4o-mini	Qwen2.5-14B-Instruct	gemma-3-27b-it
fa-IR	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
en-IR	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
ar-SA	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
en-SA	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
ar-DZ	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
en-DZ	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
ar-EG	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
en-EG	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
am-ET	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Qwen2.5-14B-Instruct
en-ET	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Qwen2.5-14B-Instruct
ar-MA	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
en-MA	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Falcon3-10B-Instruct
ha-NG	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Qwen2.5-14B-Instruct
en-NG	gpt-4o-mini	EuroLLM-22B-Instruct-2512	Qwen2.5-14B-Instruct

B Detailed Performance Results for the SAQ Track across All Locales

This appendix presents detailed SAQ performance results for all language-country pairs. Across most locales, the proposed multi-agent debate framework achieved the highest performance, outperforming the single closed-source model `gpt-4o-mini`. However, in certain countries that maintain their own domestically developed LLMs, such as Korea, region-specific single models occasionally achieved the best performance. In addition, some models exhibited limited multilingual input capabilities, leading to performance degradation on questions requiring deeper cultural understanding or more complex linguistic structures. These limitations contribute to the observed performance disparities across locales.

Model	am-ET	en-ET	ar-DZ	en-DZ	ar-EG	en-EG	ar-MA	en-MA	ar-SA	en-SA	as-AS
gpt-4o-mini	29.60	**43.80**	47.20	**52.80**	52.40	**55.20**	21.60	**36.60**	**51.80**	**52.60**	27.80
gemma-3-27b-it	25.40	38.00	44.20	47.80	48.80	49.20	26.20	35.40	46.40	46.20	27.60
Llama-3.1-8B-Instruct	4.00	33.60	27.20	41.60	33.80	48.20	15.00	29.60	26.80	43.60	12.00
Midm-2.0-Base-Instruct	3.60	**43.80**	9.00	49.60	13.60	53.00	7.00	35.20	12.00	48.20	1.40
Mistral-Nemo-Japanese-Instruct-2408	2.00	35.80	23.20	42.80	25.40	46.20	12.40	32.20	23.20	44.00	7.00
Qwen2.5-14B-Instruct	14.00	38.20	40.20	48.80	48.20	48.00	20.00	31.60	40.00	43.40	19.60
Qwen-SEA-LION-v4-32B-IT	16.20	38.80	37.80	51.80	45.80	54.00	23.00	33.80	39.00	48.20	24.00
Falcon3-10B-Instruct	2.40	36.40	8.80	41.80	11.40	45.00	8.00	32.40	8.60	43.40	2.00
Mistral-7B-Instruct-v0.3	0.40	41.20	26.60	47.00	31.60	50.60	21.00	36.20	29.60	45.80	1.20
EuroLLM-22B-Instruct-2512	2.20	41.20	42.80	49.00	44.20	50.60	18.40	34.80	41.60	48.20	4.60
Ours	**50.60**	38.60	**49.20**	49.20	**53.20**	52.80	**36.40**	36.40	46.80	46.60	**49.20**

Model	en-AS	az-AZ	en-AZ	eu-PV	en-PV	bg-BG	en-BG	zh-CN	en-CN	zh-SG	en-AU
gpt-4o-mini	**53.40**	55.20	**60.00**	36.40	48.20	48.20	**49.40**	66.80	64.80	67.60	68.00
gemma-3-27b-it	43.40	52.60	52.20	32.40	41.00	44.40	45.00	69.00	60.60	66.40	64.60
Llama-3.1-8B-Instruct	39.20	34.60	45.40	26.80	33.80	30.80	39.60	58.20	56.00	45.40	61.80
Midm-2.0-Base-Instruct	51.00	16.60	52.20	4.60	35.00	19.00	45.00	47.40	60.40	45.00	**69.00**
Mistral-Nemo-Japanese-Instruct-2408	41.20	20.20	50.00	18.20	37.40	21.80	42.60	56.40	58.80	55.40	60.20
Qwen2.5-14B-Instruct	48.40	24.60	54.80	15.60	**48.60**	30.20	44.40	73.60	62.80	66.80	65.40
Qwen-SEA-LION-v4-32B-IT	46.20	43.40	**60.00**	27.00	43.40	39.60	49.00	77.20	**68.40**	70.80	66.80
Falcon3-10B-Instruct	38.00	7.40	47.40	13.00	36.80	10.40	41.60	38.60	57.00	44.60	59.20
Mistral-7B-Instruct-v0.3	44.80	23.60	54.00	6.40	42.80	36.00	44.20	47.60	58.40	59.00	67.80
EuroLLM-22B-Instruct-2512	49.20	31.40	57.40	14.20	46.20	48.00	47.20	68.20	61.60	65.60	65.20
Ours	49.00	**59.20**	56.20	**44.00**	42.00	**48.40**	47.60	**78.00**	66.00	**74.40**	68.00

Model	en-GB	en-US	fr-FR	en-FR	el-GR	en-GR	ha-NG	en-NG	id-ID	en-ID	ga-IE
gpt-4o-mini	72.40	75.80	70.40	60.40	56.20	58.60	29.80	35.00	67.40	**63.00**	36.60
gemma-3-27b-it	71.80	73.40	64.60	50.00	58.60	57.40	27.00	29.20	70.60	57.40	29.60
Llama-3.1-8B-Instruct	67.40	74.60	57.80	47.00	34.20	49.60	15.80	28.20	49.00	48.80	17.60
Midm-2.0-Base-Instruct	74.20	**79.20**	57.40	53.80	13.60	46.40	3.80	**39.20**	46.00	57.60	7.40
Mistral-Nemo-Japanese-Instruct-2408	68.60	73.60	62.40	52.20	21.20	53.60	3.20	31.80	43.60	51.80	4.00
Qwen2.5-14B-Instruct	73.60	77.40	66.20	55.60	28.00	54.40	10.40	31.40	63.80	56.20	17.20
Qwen-SEA-LION-v4-32B-IT	**74.60**	78.80	65.60	54.60	36.40	57.00	10.20	30.20	71.80	58.60	14.80
Falcon3-10B-Instruct	69.00	74.00	57.20	52.40	11.20	49.40	9.00	31.60	37.40	52.80	7.20
Mistral-7B-Instruct-v0.3	73.60	74.60	59.80	56.20	20.00	56.40	5.80	36.60	57.20	55.20	6.80
EuroLLM-22B-Instruct-2512	72.40	76.00	70.00	**63.20**	55.80	**59.80**	5.60	36.00	49.40	59.40	36.60
Ours	72.60	75.60	**71.60**	58.20	**60.00**	59.40	**40.80**	35.40	**73.20**	58.40	**54.80**

Model	en-IE	ja-JP	en-JP	ko-KP	en-KP	ko-KR	en-KR	ms-SG	zh-TW	en-TW	fa-IR
gpt-4o-mini	51.40	60.00	**50.80**	40.40	44.60	64.80	**61.40**	68.60	53.60	52.00	**63.80**
gemma-3-27b-it	48.60	58.40	45.60	36.60	37.20	71.20	59.20	65.40	55.20	53.60	61.60
Llama-3.1-8B-Instruct	44.40	43.00	42.00	31.00	36.20	49.00	45.60	52.20	43.60	40.20	43.80
Midm-2.0-Base-Instruct	53.60	37.00	49.80	**58.60**	44.00	**83.80**	57.20	42.40	34.40	49.40	11.60
Mistral-Nemo-Japanese-Instruct-2408	49.00	53.60	43.40	24.60	38.80	42.80	50.00	42.60	33.60	49.80	30.20
Qwen2.5-14B-Instruct	49.40	59.00	47.80	39.00	38.00	56.80	51.00	55.00	55.00	46.80	39.20
Qwen-SEA-LION-v4-32B-IT	50.20	57.00	50.20	39.60	**46.80**	60.40	58.60	64.80	59.80	**54.80**	46.60
Falcon3-10B-Instruct	44.60	14.40	41.60	11.20	37.80	16.60	45.20	31.00	25.80	45.40	9.20
Mistral-7B-Instruct-v0.3	53.60	38.40	46.80	24.20	38.20	43.40	54.20	53.80	33.00	45.60	29.80
EuroLLM-22B-Instruct-2512	53.40	53.00	48.60	39.20	42.60	63.60	53.80	41.20	48.60	51.80	29.00
Ours	**54.40**	**63.60**	49.40	49.20	44.40	78.00	57.60	**80.40**	**60.40**	**54.80**	54.00

Model	en-IR	es-EC	en-EC	es-MX	en-MX	es-ES	en-ES	su-JB	en-JB	sv-SE	en-SE
gpt-4o-mini	**57.80**	**54.00**	**55.40**	58.60	**60.80**	62.00	58.80	38.40	**50.00**	52.40	**53.20**
gemma-3-27b-it	50.20	43.40	41.60	49.00	47.60	56.80	52.00	31.00	41.60	48.80	42.80
Llama-3.1-8B-Instruct	47.40	40.60	40.20	52.00	53.80	50.40	48.00	21.00	35.60	34.60	41.80
Midm-2.0-Base-Instruct	51.40	45.40	47.20	55.80	58.40	56.20	57.00	13.40	40.60	31.20	49.40
Mistral-Nemo-Japanese-Instruct-2408	50.20	44.00	43.20	50.40	54.20	48.40	53.00	12.20	40.00	25.00	47.40
Qwen2.5-14B-Instruct	50.60	48.20	47.40	60.20	58.80	62.80	57.00	31.80	43.40	39.60	45.60
Qwen-SEA-LION-v4-32B-IT	55.60	49.40	43.60	60.60	50.80	64.60	52.40	38.80	42.20	44.60	44.00
Falcon3-10B-Instruct	48.20	39.00	45.00	49.00	52.60	55.80	56.00	10.40	35.80	22.00	40.80
Mistral-7B-Instruct-v0.3	50.60	42.20	49.40	53.20	60.00	54.20	58.60	16.20	39.40	43.80	48.80
EuroLLM-22B-Instruct-2512	52.40	52.00	51.80	**61.00**	58.60	66.40	**63.20**	15.20	43.40	52.80	51.00
Ours	53.80	51.40	46.60	59.60	52.80	**69.60**	59.60	**50.80**	42.80	**54.40**	51.40

Model	tl-PH	en-PH	ta-SG	ta-LK	en-LK	en-SG	Overall
gpt-4o-mini	**62.60**	56.00	35.00	23.60	**44.40**	**82.00**	52.97
gemma-3-27b-it	56.00	50.80	53.20	30.60	41.20	72.20	49.14
Llama-3.1-8B-Instruct	38.80	47.00	14.00	8.60	33.80	65.80	39.70
Midm-2.0-Base-Instruct	26.40	52.20	0.60	0.00	39.40	75.40	39.70
Mistral-Nemo-Japanese-Instruct-2408	28.00	46.40	17.60	6.40	33.20	71.80	38.56
Qwen2.5-14B-Instruct	38.60	48.40	14.60	7.00	36.60	71.60	45.26
Qwen-SEA-LION-v4-32B-IT	57.00	56.00	22.20	14.40	39.60	77.20	48.50
Falcon3-10B-Instruct	24.00	43.80	4.20	2.60	34.20	70.80	33.48
Mistral-7B-Instruct-v0.3	38.00	50.60	4.60	1.60	37.40	75.20	41.03
EuroLLM-22B-Instruct-2512	29.40	46.00	8.20	3.60	41.40	75.00	46.10
Ours	60.00	**56.80**	**79.20**	**43.20**	43.20	77.80	**55.75**

Table 4: Detailed Performance Results for the SAQ Track across All Locales.

C Track 1: Short Answer Questions (SAQ) Prompts

This section presents the detailed prompt used for the SAQ track in our multi-agent debate framework. While the base prompt was developed in English, for non-English linguistic contexts, we used `gpt-4o-mini` to translate the prompt into the respective local languages. The prompt operationalizes a three-stage debate process consisting of independent response generation, mutual critique & refinement, and final adjudication by explicitly defining agent roles, interaction protocols, and evaluation criteria. It is designed to elicit culturally grounded open-ended responses, guiding each agent to generate an initial answer, review peer outputs, and revise its reasoning before a judge model determines the final response. The structure emphasizes interpretability, cultural appropriateness, and robustness in generative reasoning.

Step 1. Independent Response Generation

You are a person from {COUNTRY_REGION} who is trying to explain your country's culture to a foreigner.
Answer the following question, providing a single answer without any explanations.

Step 2. Mutual Critique and Refinement

 # SYSTEM

You are an expert reviewer in a multi-agent debate pipeline for Short Answer Questions (SAQ). Your job is to critique other agents' answers and revise your own answer to maximize correctness against human reference answers.

Goals:

- Produce the most reference-matchable short answer.
- Prefer canonical, standardized, minimal answers.

Rules:

- Do NOT re-derive from scratch. Focus on compare → verify → correct → finalize.
- If multiple answers are plausible, choose ONE canonical answer; do not list multiple candidates.
- Keep the final answer short.
- Follow the output JSON schema exactly. Output ONLY JSON, no extra text.

Output schema (strict):

```
{
  "critique": {
    "agent_1": "<one-sentence evaluation>",
    "agent_2": "<one-sentence evaluation>"
  },
  "final_answer": "<single-line short answer>",
  "why_final": "<1-2 sentences, concise>",
  "confidence": <number from 0.00 to 1.00>
}
```

Evaluation priority:

1. Factual correctness

2. Likelihood of matching the reference answer
3. Locale-appropriate wording and formatting
4. Brevity (SAQ-friendly)

 # USER

You are Agent {AGENT_ID} performing Step (2) Model Feedback.

Question:
{QUESTION}

Your Step (1) answer:
{MY_ANSWER}

Other agents' Step (1) answers:
Agent {OTHER_AGENT_1_ID}:
{OTHER_ANSWER_1}

Agent {OTHER_AGENT_2_ID}:
{OTHER_ANSWER_2}

TASK:

- Critique the other two answers (one sentence each).
- Decide whether to keep or revise your answer.
- Output ONE best final answer.

Return ONLY valid JSON.

Step 3. Final Adjudication

 # System
You are the final Judge in a multi-agent pipeline for culturally grounded short-answer questions (SAQ). You are evaluating an SAQ for {LOCALE} ({COUNTRY}).
These questions aim to capture cultural norms, customs, and everyday shared understandings in {COUNTRY} using a brief answer. Your job is to choose ONE final answer from the candidates proposed by the agents.
At this stage, you must not invent new content. You must choose one of the provided candidate answers, or make only minimal surface edits (e.g., spelling/formatting/very small phrasing cleanup) without adding new information.

———————

Reference about agents (for context only)

———————

- LLM1: a general-purpose language model trained broadly across many languages and countries.
- LLM2: a model designed to reflect language/cultural tendencies at a continental or macro-regional level.

- LLM3: a model developed in the country/region relevant to this locale.

This information is provided only to help your judgment; it does not guarantee correctness. Your final decision must be based on the suitability of the answer itself for this cultural SAQ.

Output format (STRICT)

Output ONLY valid JSON (no markdown, no extra text), in this exact schema and key order:

```
{
  "reason": "<1-2 sentences explaining why you chose this answer>",
  "final_answer": "<the final answer>"
}
```

 # User
Below is information for a culturally grounded short-answer question (SAQ).
[Question]
{QUESTION}
[Candidate answers]

- Agent LLM1:
  {A1_FINAL}
- Agent LLM2:
  {A2_FINAL}
- Agent LLM3:
  {A3_FINAL}

[Agents' selection reasons (for reference)]

- Agent LLM1:
  {A1_REASON}
- Agent LLM2:
  {A2_REASON}
- Agent LLM3:
  {A3_REASON}

Instructions:

- Base your decision only on the candidate answers above.
- Do NOT add new information.
- Follow the STRICT JSON output format from the system prompt.

D Track 2: Multiple-Choice Questions (MCQ) prompt

This section presents the detailed prompt used for the MCQ track in our multi-agent debate framework. The prompt applies the same three-stage debate structure to a multiple-choice decision setting and specifies the roles of debater agents and the judge model. Unlike the SAQ configuration, it prioritizes structure comparison of predefined options and consistency in final answer selection rather than generative diversity. The interaction protocol is designed to mitigate individual model bias and promote stable, well-justified decisions.

Step 1. Independent Response Generation

Choose only one from the given alphabet choices (e.g., A, B, C). Assume that exactly one option is correct and do NOT consider the possibility that none of the options apply.
Provide as JSON format:

```
{"answer_choice":""}
```

Explain your answer in less than three sentences.
Question:
{question}
Options:
{options}
Answer:

Step 2.1 Mutual Critique

You are currently discussing whether the given answer is culturally plausible and typical with another discussant. Respond to the discussant by providing any relevant feedback. Respond in less than three sentences.
Question:
{question}
Options:
{options}
You:
{your_response}
Discussant:
{other_response}
Response:

Step 2.2 Refinement

You are currently discussing whether the given answer is culturally plausible and typical with another discussant.
Question:
{question}
Options:
{options}
You:
{your_response}
Discussant:
{other_response}
Your feedback:
{your_feedback}
Discussant feedback:

{other_feedback}
Based on the above discussion, critically think and make your final decision. Choose only one from the given alphabet choices (e.g., A, B, C). Provide as JSON format:

```
{"answer_choice":""}
```

Answer:

Step 3. Final Adjudication

You are a judge responsible for making a final decision based on the debate history between Model1 and Model2. They have debated whether the given answer is culturally plausible and typical. Do NOT make any independent judgments; base your final decision solely on the debate.
Respond with a final decision as JSON format:

```
{"answer_choice":""}
```

Question:
{question}
Options:
{options}
*** Debate starts ***
# Model1 opinion: {model_1_response}
# Model2 opinion: {model_2_response}
# Model1 feedback: {model_1_feedback}
# Model2 feedback: {model_2_feedback}
# Model1 final decision: {model_1_final_decision}
# Model2 final decision: {model_2_final_decision}
*** Debate ends ***
Final decision:

E Impact of Judge Model Scales and Closed-source LLMs on SAQ Performance

Judge Model	Overall
gemma-3-1b-it	48.55
gemma-3-4b-it	55.11
gemma-3-12b-it	55.55
gemma-3-27b-it	55.75
gpt-4o-mini	**57.47**

Table 5: Impact of Judge Model Scale and Closed-Source Substitution on SAQ Overall Performance.

To analyze the impact of the judge model, we varied its parameter scale within the same family (`Gemma-3`) and additionally evaluated a closed-source LLM. As shown in Table 5, SAQ performance consistently improves with increasing judge capacity, rising from 48.55 with `gemma-3-1b-it` to 55.75 with `gemma-3-27b-it`. This indicates that a stronger judge more effectively evaluates consistency, cultural appropriateness, and response validity.

Replacing the open-source judge with the closed-source model `gpt-4o-mini` further increased the overall score to 57.47, surpassing the officially submitted score of 55.75. This suggests that judge quality directly constrains the upper bound of performance in our framework.