GUIR at SemEval-2026 Task 7: Probing Cultural Knowledge in LLMs via Multi-Agent Debate

Reihaneh Iranmanesh, Ophir Frieder, Nazli Goharian
Information Retrieval Lab
Department of Computer Science
Georgetown University
{rei, ophir, nazli}@ir.cs.georgetown.edu

Abstract

We present the GUIR system for SemEval-2026 Task 7, Everyday Knowledge Across Diverse Languages and Cultures, which probes the extent to which general-purpose LLMs encode cultural knowledge without any culture-specific supervision or fine-tuning. Our system addresses two tracks built on the BLEnD benchmark. For the short-answer question (SAQ) track, we employ zero-shot prompting with `gpt-4.1`, achieving 55.5% accuracy across 61 language locales. For the multiple-choice question (MCQ) track, we propose a three-stage pipeline: (1) zero-shot chain-of-thought inference with `gpt-5-mini`, (2) cross-locale majority voting to correct inconsistent predictions, and (3) a multi-agent debate protocol in which three LLM instances argue and adjudicate over residual errors. This pipeline achieves 97.47% overall accuracy across 30 locales, ranking **first** among all submitted systems on the MCQ track. We further conduct a targeted human evaluation on the Persian locale, revealing that BLEnD's lemma-matching scorer systematically underestimates model performance, with human annotators scoring the system 18 percentage points higher than the lemma-matching evaluation. This reveals the need for better evaluation of morphologically rich languages like Persian.

1 Introduction

SemEval-2026 (Ghosh et al., 2026) Task 7 *Everyday Knowledge Across Diverse Languages and Cultures* (Ousidhoum et al., 2026) is built on the BLEnD benchmark (Myung et al., 2024), a dataset of 52,557 question-answer pairs spanning 16 countries/regions and 13 languages, including low-resource languages such as Amharic, Assamese, Azerbaijani, Hausa, and Sundanese. Questions cover six socio-cultural domains: food, sports, family, education, holidays/celebrations/leisure, and work-life. The benchmark offers two evaluation formats. In the short-answer question (SAQ) format, models generate a free-text response that is matched against native-speaker annotations. In the multiple-choice question (MCQ) format, models select the culturally correct answer for a target country from four options, where distractors are drawn from other countries' answers to the same question, making them culturally plausible rather than arbitrary.

Prior evaluation on BLEnD reveals a performance gap of up to 57 percentage points between high-resource cultures (US English, $\sim$79% on SAQ) and low-resource ones (Ethiopian Amharic, $\sim$12%) for GPT-4, the strongest model tested at the time of the benchmark's release (Myung et al., 2024). The Pearson correlation between SAQ and MCQ performance across models is 0.93, suggesting the two formats probe the same underlying cultural knowledge.

Since this is an evaluation-only task, we rely entirely on zero-shot CoT rather than fine-tuning or few-shot prompting on BLEnD examples, directly testing how well a strong general-purpose LLM resolves culturally grounded SAQs and MCQs without any culture-specific supervision. [1]

Furthermore, we deliberately refrained from retrieval-augmented generation (RAG) pipelines that draw on external corpora: since BLEnD is publicly available, any retrieval index built from open web data could inadvertently include benchmark items, constituting a form of data contamination.

2 Background and Related Work

The cultural limitations of LLMs have attracted growing attention. Hershcovich et al. (2022) provide a systematic account of the challenges in cross-cultural NLP, noting that cultural norms are implicit, context-dependent, and highly vari-

[1] All code and analysis are publicly available in our GitHub repository: GUIR-at-SemEval-2026-Task-7.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3549–3561

able across regions. LLMs encode cultural biases stemming from imbalanced training corpora (Navigli et al., 2023): models trained predominantly on English data default to Western perspectives when queried about other regions (Durmus et al., 2024). Several benchmarks probe this gap. GeoMLAMA (Yin et al., 2022) offers geo-diverse commonsense probing across five countries; CultureBank (Shi et al., 2024) and CultureAtlas (Fung et al., 2024) collect norms from social media and Wikipedia, though both are restricted to English and formally documented information. Language-specific efforts include CLIcK (Kim et al., 2024) for Korean, COPAL-ID and IndoCulture (Wibowo et al., 2024; Koto et al., 2024) for Indonesian, and CAMEL (Naous et al., 2023) for Arabic. BLEnD (Myung et al., 2024) distinguishes itself by being fully hand-crafted by native-speaker annotators, covering mundane everyday knowledge that rarely surfaces in formal sources, and spanning a typologically diverse set of 13 languages including several low-resource ones.

Recent work has pushed toward richer and more challenging cross-cultural evaluation. Chiu et al. (2025) introduce CulturalBench, assembled via human-AI red-teaming across 45 regions, showing that even frontier models reach only 28.7-61.5% on the hard variant compared to 92.4% human performance. Romanou et al. (2025) and Singh et al. (2025b) expose systematic regional knowledge gaps in standard multilingual benchmarks such as MMLU, with consistent under-performance on non-Western locales. Liu et al. (2025) specifically target implicit cultural values embedded in natural conversation, finding that even models at human-level performance on value selection still fall substantially short on nuanced attitude detection. Ying et al. (2025) reveal a "Cultural-Linguistic Synergy" effect: models perform better when the question language matches the cultural context, suggesting that multilingual evaluation must account for both dimensions jointly rather than treating language and culture as independent axes.

While recent benchmarks achieve broad coverage, they predominantly use MCQ formats and web-sourced or formally documented knowledge. BLEnD's SAQ format is still distinctive: rather than selecting among distractors, models must generate the correct cultural answer from scratch, a far harder test of genuine cultural knowledge (Myung et al., 2024). On the modeling side, Nyandwi et al. (2025) show that grounding multimodal LLMs with culturally curated knowledge graphs yields substantial gains over general-purpose models, while Chang et al. (2024) demonstrate that dedicated monolingual models for low-resource languages outperform multilingual models on basic NLP tasks in those languages-both highlighting the persistent gap between general-purpose systems and culture- or language-specific ones that our zero-shot approach is benchmarked against.

3 System Overview

3.1 Track 1: Short Answer Questions (SAQ)

For the SAQ track, the dataset comprises 30,500 questions across 61 language codes, with 500 questions per language code. For inference, we used `gpt-4.1` via the OpenAI Chat Completions API (OpenAI, 2025). Each question was presented with a minimal zero-shot prompt as shown in Figure 1.

```
Read the following question and provide
a single answer without any explanations.
The answer should be 1-3 words.

Question: {question}
Answer:
```

Figure 1: Zero-shot prompt template used for SAQ inference similar to (Myung et al., 2024).

Evaluation Metric. The official SAQ scorer evaluates predictions via language-aware lemma matching. For each item, the predicted string is first checked for inclusion against the native-language reference answers using language-specific lemmatizers and stemmers. (See Appendix 6.1 for the full list of tools used per language). A prediction is marked correct only if *every* lemmatized token of the reference answer is present in the lemmatized token set of the prediction. A single missing token is sufficient to fail the match. If the native-language check fails, the scorer additionally attempts matching against the corresponding English reference answer using spaCy (Honnibal et al., 2020). Accuracy is then computed as the percentage of items answered correctly out of all items in the locale.

Formally, let $\mathcal{L}(\cdot, \ell)$ denote the lemmatization function for language ℓ, and let A_q be the set of reference answers for question q. A prediction y is marked correct if

$$s_q(y) = \begin{cases} 1, & \text{if } \exists\, a \in A_q \cup A_q^{en} : \\ & \quad \mathcal{L}(a, \ell) \subseteq \mathcal{L}(y, \ell) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where A_q^{en} denotes the English reference answers and $\mathcal{L}(a, \ell) \subseteq \mathcal{L}(y, \ell)$ requires every lemmatized token of a to appear in the lemmatized tokens of y. Accuracy for a locale is

$$S = \frac{1}{|Q|} \sum_{q \in Q} s_q(y_q) \times 100. \quad (2)$$

3.2 Track 2: Multiple-Choice Questions (MCQ)

Our MCQ pipeline proceeds in three sequential stages: an initial zero-shot chain-of-thought pass, a majority-voting correction step, and a debate-based re-prediction step for residual errors.

Stage 1: Zero-Shot Chain-of-Thought Baseline. The MCQ evaluation set contains 47,014 items spanning 30 language locales but unlike the SAQ track, all questions are in English. Table 1 shows an example for Persian language.

Table 1: Example entries comparing SAQ and MCQ formats for the Persian (Farsi) locale.

Format	Content
SAQ	`fa-IR_108` مردم در ایران چه غذاهای خیابانی دوست دارند بخورند؟ `en-IR_108` What street food do people from Iran like to eat?
MCQ	`fa-IR_0514` What street food do people from Iran like to eat? *(A)* couscous *(B)* falafel *(C)* stir-fried rice cakes *(D)* tacos

For the initial pass we submitted predictions generated by `gpt-5-mini` (Singh et al., 2025a) via the OpenAI Chat Completions API. We selected this model for its balance of reasoning capability and inference cost, the latter being a practical consideration given the scale of the evaluation set.

MCQs were presented to the model in a zero-shot chain-of-thought prompt (Wei et al., 2023). The prompt instructed the model to reason step by step through all four answer candidates before committing to a final choice. (See Appendix 6.2 for the full prompt).

The final answer was extracted via a deterministic parsing rule: the first occurrence of a letter in {`A, B, C, D`} following the `Answer:` prefix was taken as the prediction. If no such prefix was found, a fallback rule selected the first {`A, B, C, D`} token anywhere in the response. In the rare event of a complete API failure, the system defaulted to `A` to ensure a complete submission.

Question-Level Deduplication Cache. Inspection of the data revealed substantial lexical overlap across locales: the same underlying question stem appears in several locale-specific variants, often with identical or near-identical wording. To avoid redundant API calls, we maintained an in-memory answer cache keyed on the normalized question string. When a question had been seen previously, the cached *answer text* (i.e., the full option string of the predicted choice) was looked up among the four options of the current instance; if an exact string match was found, the corresponding letter was assigned without an API call (a cache hit). A cache miss - whether because the question was genuinely new or because the option texts differed across locales - triggered a fresh API call and updated the cache entry. This approach eliminated 92.0% of API calls. (See Appendix 6.3 for the full breakdown).

Stage 2: Cross-Locale Majority Voting. Because the same question appears across multiple locales with different distractor sets, the Stage 1 predictions for a single question stem constitute an implicit ensemble: if the model consistently selects the same answer text across locales, that answer is more likely to be correct.

We grouped all rows by their normalised question string. Within each group we count how many times each answer *text* was selected across locales. The top answer - the one with the highest selection count, with ties broken by coverage (number of locales in which it appears) and then alphabetically - is treated as the consensus prediction. For any row in the group where the top answer is present in the choices but was *not* selected by Stage 1, we flip the prediction to the top answer.

To avoid over-correcting on noisy or small groups we apply two safeguards: (i) groups with fewer than $K_{\min} = 5$ rows are left unchanged, and (ii) a flip is applied only when the top answer commands at least $\rho_{\min} = 50\%$ of the votes within the group.

These thresholds reflect natural majority semantics: $\rho_{\min} = 0.50$ requires the consensus answer to command an outright majority of votes (i.e. more

than half), and $K_{\text{min}} = 5$ ensures the group is large enough for that majority to be meaningful.

Majority-voting algorithm (Stage 2)

Input: question groups $\mathcal{G}$, Stage-1 predictions $\hat{y}$
for each group $g \in \mathcal{G}$ **do**
if $|g| < K_{\text{min}}$ **then** skip
compute votes$(a) = |\{q \in g : \hat{y}_q = a\}|$
$a^* \leftarrow \arg\max_a$ votes(a)
if votes$(a^*)/|g| < \rho_{\text{min}}$ **then** skip
for each $q \in g$ where $a^* \in$ choices(q) and $\hat{y}_q \neq a^*$ **do**
$\hat{y}_q \leftarrow a^*$

Figure 2: Pseudocode for the cross-locale majority-voting correction (Stage 2). $K_{\text{min}} = 5$; $\rho_{\text{min}} = 0.50$.

Stage 3: Debate-Based Re-Prediction. After majority voting, a set of questions remain unchanged either because their question group was too small ($|g| < K_{\text{min}} = 5$) or because no single answer commanded an outright majority ($< \rho_{\text{min}} = 50\%$). These are the predictions we treat as *low-confidence*: the cross-locale signal was insufficient to validate them, making them the most likely candidates for error. For these residual items we apply a debate-style re-prediction pass inspired by Du et al. (2023), who show that having multiple LLM instances propose and debate answers over several rounds substantially improves factuality and reasoning.

Unlike prior debate setups that treat the correct answer as unknown, our setting is asymmetric: we *assume* the current prediction is wrong. We therefore design a three-agent protocol with three API calls per question:

1. **Agent A (Critic).** Given the question, the four options, and the pre-assumed wrong answer, Agent A argues why the flagged option is incorrect and proposes an alternative answer (2-3 sentences).

2. **Agent B (Reviewer).** Seeing Agent A's argument, Agent B either endorses the proposed alternative or advocates for a different option, providing its own justification (2-3 sentences).

3. **Judge.** A neutral judge reads both arguments, is reminded that the originally flagged answer is (assumed to be) wrong, and selects the single best remaining option. The judge is instructed to respond with exactly one letter (`A`, `B`, `C`, or `D`); any other output is ignored and the question is left unchanged.

The full prompt templates for each role are provided in Appendix 6.4.

Model choice for debate. We use `GPT-OSS-20B` (OpenAI et al., 2025) for all three debate roles. This choice is motivated by our results on the development set (see Table 7 in Appendix 6.5), where we evaluated 16 open-source models across 23 language locales to identify the strongest candidate for Stage 3. `GPT-OSS-20B` achieves the highest overall MCQ accuracy (89.19%) among the open-source models we evaluate, outperforming the next best model (Qwen3-8B, 82.43%) by nearly seven percentage points. Using the strongest available model for the debate stage maximises the probability of recovering the correct answer from an initially incorrect prediction.

Evaluation Metric. MCQ predictions are evaluated with the official scorer, which reads the reference file alongside a one-hot prediction file with columns `id`, `A`, `B`, `C`, `D`. For each question, a prediction is counted correct if and only if the column corresponding to the gold answer is set to 1 and *exactly* one column is marked; items with no prediction, a malformed one-hot encoding, or a duplicate ID (resolved by keeping the last occurrence) are skipped rather than penalised. Accuracy is computed independently for each `language_region` by grouping reference items on the locale prefix of the question ID:

$$\text{Acc}_\ell = \frac{|\{q \in Q_\ell : \hat{y}_q = y_q\}|}{|Q_\ell|} \times 100, \quad (3)$$

where Q_ℓ is the set of questions for locale ℓ, y_q is the gold answer, and $\hat{y}_q$ is the predicted answer. The **overall score** is the *macro-average* across all locales,

$$\text{Acc}_{\text{overall}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{Acc}_\ell, \quad (4)$$

where $\mathcal{L}$ is the set of language -region locales. This is a *uniform* macro-average: every locale contributes equally regardless of its size, so a locale with 300 items is weighted identically to one with 3,000. All scores reported in Table 4 follow this metric.

4 Results

4.1 SAQ Results

Table 2 reports accuracy scores across all locales for both native-language and English-language

Table 2: SAQ accuracy (%) per country/region, language, and locale for `gpt-4.1` (overall: 55.5%). Each non-English locale is shown alongside its English-language counterpart (`en-*`) where available. The four English-native locales (`en-AU`, `en-GB`, `en-SG`, `en-US`) appear without an `en-*` counterpart, as do the three Singapore locales (`ms-SG`, `ta-SG`, `zh-SG`) which share `en-SG`.

Country	Lang.	Locale	Native	en-*	Country	Lang.	Locale	Native	en-*
Ethiopia	Amharic	am-ET	25.2	37.4	Indonesia	Indonesian	id-ID	73.2	57.8
Algeria	Arabic	ar-DZ	47.8	53.8	Japan	Japanese	ja-JP	63.6	48.8
Egypt	Arabic	ar-EG	53.0	49.6	North Korea	Korean	ko-KP	46.6	46.2
Morocco	Arabic	ar-MA	24.0	38.8	South Korea	Korean	ko-KR	76.4	65.8
Saudi Arabia	Arabic	ar-SA	53.4	50.6	Singapore	Malay	ms-SG	72.0	—
India	Assamese	as-AS	40.4	49.8	West Java	Sundanese	su-JB	49.2	45.4
Azerbaijan	Azerbaijani	az-AZ	67.0	60.2	Sweden	Swedish	sv-SE	63.4	53.4
Bulgaria	Bulgarian	bg-BG	58.0	49.8	Sri Lanka	Tamil	ta-LK	29.4	47.4
Greece	Greek	el-GR	64.8	55.6	Singapore	Tamil	ta-SG	42.8	—
Australia	English	en-AU	66.8	—	Philippines	Tagalog	tl-PH	60.0	53.2
United Kingdom	English	en-GB	73.8	—	China	Chinese	zh-CN	71.4	62.2
Singapore	English	en-SG	79.8	—	Singapore	Chinese	zh-SG	72.2	—
United States	English	en-US	75.8	—	Taiwan	Chinese	zh-TW	61.2	51.4
Ecuador	Spanish	es-EC	63.0	57.4	Iran	Persian	fa-IR	64.0	58.6
Spain	Spanish	es-ES	74.8	59.4	France	French	fr-FR	75.2	61.0
Mexico	Spanish	es-MX	69.6	58.0	Ireland	Irish	ga-IE	39.2	53.6
Basque (Spain)	Basque	eu-PV	48.4	46.0	Nigeria	Hausa	ha-NG	36.0	32.6
								Overall Accuracy: 55.495%	

question variants. The system achieves an overall score of **55.5%**. Performance is strongest on high-resource locales: `en-SG` (79.8), `ko-KR` (76.4), `en-US` (75.8), `fr-FR` (75.2), and `es-ES` (74.8). Scores are lowest for `ar-MA` (24.0), `am-ET` (25.2), and `ta-LK` (29.4), consistent with the performance gap between high- and low-resource cultures reported in prior BLEnD evaluations (Myung et al., 2024).

For these three lowest-scoring low-resource languages, English-language question variants score higher than their native-language counterparts (e.g., `en-ET` 37.4 vs. `am-ET` 25.2; `en-MA` 38.8 vs. `ar-MA` 24.0; `en-LK` 47.4 vs. `ta-LK` 29.4), suggesting that the model's cultural knowledge might be more reliably surfaced when queried in English. However, the reverse is true for `az-AZ` (67.0 vs. `en-AZ` 60.2), `ko-KR` (76.4 vs. `en-KR` 65.8), and `id-ID` (73.2 vs. `en-ID` 57.8), where native-language variants outperform English-locales for which `gpt-4.1` might have stronger native-language cultural grounding.

Human Evaluation for Persian. Persian's morphological richness causes lemma-based scoring to miss semantically correct answers (Iranmanesh et al., 2026): surface variants such as *nan o panir* (نان و پنیر) and *noon o panir* (نون و پنیر) ("bread and cheese") are treated as mismatches despite being semantically equivalent. To quantify this gap, we recruited three native Persian speakers who had lived in Iran for more than half their lifetime, similar to how annotators were selected in (Myung et al., 2024). We asked each annotator to independently score all 500 `fa-IR` SAQ items on a binary correct/incorrect scale, judging semantic equivalence regardless of surface form. (See annotation prompt in Appendix 6.6).

As shown in Table 3, pairwise agreement is moderate (r=0.33-0.39, κ=0.32-0.39; Fleiss' κ=0.362, all $p < 10^{-13}$), consistent with the subjectivity of open-ended cultural questions. The scorer assigns 64.0%, far below the mean human score of 80.2% and the majority-vote score of 82.2%: a gap of +18.2 pp. The asymmetry is substantial: in 73/500 items, all annotators agreed the answer was correct while the scorer marked it wrong; the reverse occurred in only 6/500 items. This confirms that lemma matching systematically under-estimates model performance on Persian, motivating soft-match scoring (Iranmanesh et al., 2026) as a complement to the BLEnD pipeline.

4.2 MCQ Results

Table 4 reports MCQ accuracy across all 30 locales. Our full pipeline ranked first, achieving **97.47%** overall, improving over the Stage 1 zero-shot CoT baseline by +3.26 pp and ranking **first** among all submitted systems on the MCQ track.

Baseline. The zero-shot CoT baseline alone exceeds 90% on most locales, suggesting that a

Table 3: Human evaluation on `fa-IR` (500 items). *Top:* inter-annotator agreement (r = Pearson, κ = Cohen's). *Bottom:* per-annotator score and agreement with the BLEnD scorer; Maj. = majority vote.

Inter-annotator agreement		
Pair	r	κ
A1 vs. A2	0.388	0.385
A1 vs. A3	0.383	0.382
A2 vs. A3	0.328	0.322
Fleiss' κ (3-way)		0.362

Human evaluation score vs. BLEnD's evaluation score				
	A1	**A2**	**A3**	**Maj.**
Score (%)	79.6	83.4	77.6	**82.2**
Pearson r	0.272	0.293	0.347	—
Agreement (%)	68.8	69.8	71.6	—

False neg. (BLEnD = 0, all human = 1): 73/500 (14.6%)
False pos. (BLEnD = 1, all human = 0): 6/500 (1.2%)
BLEnD's automatic evaluation scorer: 64.0%

strong general-purpose LLM can resolve culturally grounded MCQ without culture-specific supervision. Unsurprisingly, performance is strongest on high-resource Western locales (`en-GB`: 99.17%, `fr-FR`: 98.70%) and weakest on low-resource ones (`am-ET`: 84.18%, `ha-NG`: 85.66%, `ar-SA`: 87.16%), consistent with (Myung et al., 2024).

Pipeline Gains. Majority voting (Stage 2) corrects inconsistent predictions at negligible cost by exploiting cross-locale redundancy in BLEnD. The debate step (Stage 3) then delivers the largest gains where they matter most: `ar-SA` (+11.49 pp), `am-ET` (+7.37 pp), `el-GR` (+5.26 pp), and `ar-EG` (+5.17 pp) - all low-resource or culturally distant locales. This pattern suggests that cultural errors in these locales reflect genuine knowledge gaps that single-pass inference cannot recover, but multi-agent argumentation can. `zh-CN` is the only locale to reach a perfect 100% after Stage 3.

MCQ vs. SAQ. The strong MCQ performance should not be over-interpreted. While MCQ accuracy ranges from 89-100%, SAQ accuracy on the same locales spans 24-80%. LLMs are adept at identifying the correct cultural answer when it is present among four options, but fall considerably shorter when required to generate it from scratch (see Table 9 in Appendix 6.7). This gap is further widened by the SAQ automatic scorer, which misses semantically correct answers due to surface-level differences, an issue we confirmed through human evaluation on `fa-IR` (Section 4.1).

Table 4: MCQ accuracy (%) per locale and country/region. S1: zero-shot CoT (`gpt-5-mini`); S3: after majority voting and multi-agent debate. Δ = S3 − S1.

Locale	**Country**	**N**	**S1**	**S3**	**Δ**
`am-ET`	Ethiopia	2863	84.18	91.55	+7.37
`ar-DZ`	Algeria	2600	96.58	98.54	+1.96
`ar-EG`	Egypt	368	91.03	96.20	+5.17
`ar-MA`	Morocco	543	91.53	95.40	+3.87
`ar-SA`	Saudi Arabia	444	87.16	98.65	+11.49
`as-AS`	Assam (India)	2451	88.09	92.78	+4.69
`az-AZ`	Azerbaijan	2297	95.65	99.09	+3.44
`bg-BG`	Bulgaria	648	99.54	99.85	+0.31
`el-GR`	Greece	2734	92.76	98.02	+5.26
`en-AU`	Australia	513	94.54	98.64	+4.10
`en-GB`	United Kingdom	2167	99.17	99.95	+0.78
`en-US`	United States	1942	98.56	99.79	+1.23
`es-EC`	Ecuador	977	98.67	99.59	+0.92
`es-ES`	Spain	1931	96.69	98.91	+2.22
`es-MX`	Mexico	1899	95.42	99.26	+3.84
`eu-PV`	Basque Country	1075	94.23	96.93	+2.70
`fa-IR`	Iran	3699	91.02	93.40	+2.38
`fr-FR`	France	307	98.70	99.35	+0.65
`ga-IE`	Ireland	856	98.48	99.77	+1.29
`ha-NG`	Nigeria	2008	85.66	89.29	+3.63
`id-ID`	Indonesia	1995	94.84	98.35	+3.51
`ja-JP`	Japan	410	91.71	96.34	+4.63
`ko-KP`	North Korea	2185	90.89	94.78	+3.89
`ko-KR`	South Korea	2512	96.14	99.40	+3.26
`su-JB`	Indonesia	2345	94.71	97.95	+3.24
`sv-SE`	Sweden	447	94.41	97.09	+2.68
`ta-LK`	Sri Lanka	1114	97.13	99.55	+2.42
`tl-PH`	Philippines	1327	96.53	98.49	+1.96
`zh-CN`	China	1929	98.50	100.00	+1.50
`zh-SG`	Singapore	428	93.69	97.20	+3.51
Overall		47,014	**94.21**	**97.47**	**+3.26**

5 Conclusion

We presented the GUIR system for SemEval-2026 Task 7, achieving 97.47% on the MCQ track (first place) and 55.5% on the SAQ track. Our results confirm that a strong general-purpose LLM with zero-shot CoT already encodes substantial cultural knowledge, with majority voting and multi-agent debate providing targeted gains on low-resource locales. The MCQ-SAQ gap highlights that recognising cultural knowledge is significantly easier than generating it. While culturally grounded and language-specific LLMs are a growing research direction, our experiments suggest that open-source multilingual models cannot yet compete with closed-source state-of-the-art models on this task; closing this gap remains an important direction for future work.

Limitations

The majority voting step exploits a structural property of BLEnD, the same question stem appearing across multiple locales, that may not be present in other cultural benchmarks. Adapting this stage to datasets without cross-locale redundancy is an interesting direction, for instance through cross-model ensemble voting rather than cross-locale voting. Our human evaluation is limited to Persian, motivated by its morphological complexity, but similar scorer underestimation likely affects other morphologically rich languages in the benchmark such as Amharic, Assamese, and Azerbaijani; a broader multilingual human evaluation study would provide a more complete picture of lemma-matching scorer reliability. Finally, our pipeline depends on proprietary model APIs, which evolve over time; we release all prompts and pipeline code to ensure our methodology remains relatively reproducible even as underlying model versions change.

Acknowledgments

Thank you to the three adjudicators for our human evaluation: Homa Kashfi, Hannaneh Shojaei and Amirali Iranmanesh.

References

Adrien Barbaresi. 2021. simplemma: A simple multilingual lemmatizer for Python.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *Preprint*, arXiv:2408.10441.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & LM benchmarking. *Preprint*, arXiv:2402.09369.

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.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Reihaneh Iranmanesh, Saeedeh Davoudi, Pasha Abrishamchian, Ophir Frieder, and Nazli Goharian. 2026. Taraz: Persian short-answer question benchmark for cultural evaluation of language models. *Preprint*, arXiv:2602.22827.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in korean. In *Proceedings of the 2024 Joint International Conference*

on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3335–3346, Torino, Italia. ELRA and ICCL.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically-influenced cultural commonsense reasoning across eleven indonesian provinces. *Preprint*, arXiv:2404.01854.

Taku Kudo. 2006. MeCab: Yet another part-of-speech and morphological analyzer.

Ziyi Liu, Priyanka Dey, Jen tse Huang, Zhenyu Zhao, Bowen Jiang, Rahul Gupta, Yang Liu, Yao Du, and Jieyu Zhao. 2025. Can llms grasp implicit cultural values? benchmarking llms' cultural intelligence with cq-bench. *Preprint*, arXiv:2504.01127.

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.

Tarek Naous, Michael J Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal LLMs with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24187–24231, Suzhou, China. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI. 2025. Gpt-4.1. `https://platform.openai.com/docs/models/gpt-4.1`.

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.

Eunjeong L. Park and Sungzoon Cho. 2014. KoNLPy: Korean natural language processing in Python. In *Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology*.

Peng Qi, Yuhao Zhang, Yuhao Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

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

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, and 1 others. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. *Preprint*, arXiv:2404.15238.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025a. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025b. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. *Preprint*, arXiv:2311.01012.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251, Vienna, Austria. Association for Computational Linguistics.

6 Appendix

6.1 SAQ Scoring Tools by Language

Language(s)	Tool
Arabic, Greek, Tamil	Stanza (Qi et al., 2020)
Japanese	MeCab (Kudo, 2006)
Korean	KoNLPy (Park and Cho, 2014)
Indonesian, Persian, Spanish, French, Bulgarian, Tagalog, Malay	`simplemma` (Barbaresi, 2021)
Hausa	`hausastemmer`
Amharic	`amharicNLP`
Azerbaijani	Custom stemmer
Irish, Basque	`PyStemmer` (Basque model)

Table 5: Language-specific lemmatization and stemming tools used by the SAQ scorer.

6.2 Zero-shot Prompt Template Used for MCQ

```
You are answering a multiple-choice
question. Think through it carefully,
then provide your answer.

Question: {question}

A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}

Please reason through this question step
by step, considering each option carefully.
Then, provide your final answer as a
single letter (A, B, C, or D) on the last
line, prefixed with ``Answer: ''

Format your response like this:
[Your reasoning here]
Answer: [A, B, C, or D]
```

Figure 3: Zero-shot chain-of-thought prompt template used for all MCQ inference. Curly-brace placeholders are filled with the normalized question and answer option strings at runtime.

6.3 API Usage Statistics for Stage 1 MCQ Inference

Table 6: API usage and caching statistics for Stage 1 MCQ inference across 47,014 questions. The 92% cache hit rate reflects substantial overlap in prompts across locales, reducing cost to $3.23 total.

Statistic	Value
Total questions	47,014
API calls made	3,781
Cache hits	43,233
Cache hit rate	92.0%
Total cost (USD)	$3.23

6.4 Debate Prompt Templates

Agent A - Critic

System:
You are a critical reasoning agent. You are told that a specific answer to a multiple-choice question is WRONG. Your job is to explain why it is wrong and argue for what the correct answer should be. Be concise (2-3 sentences).

User:
Question: `{question}`
Choices: `{choices}`
The answer (`{wrong}`): `{wrong_text}` has been marked as WRONG.
Explain why it is wrong and argue for the correct answer.

Agent B - Reviewer

System:
You are a critical reasoning agent reviewing another agent's argument. You agree the flagged answer is wrong. Either support the other agent's proposed answer or argue for a better alternative. Be concise (2-3 sentences).

User:
Question: `{question}`
Choices: `{choices}`
(`{wrong}`): `{wrong_text}` is WRONG.
Agent A argues: "`{arg_a}`"
Do you agree with Agent A's proposed answer, or is there a better option? Justify your position.

Judge

System:
You are a neutral judge. Two agents have debated the correct answer to a multiple-choice question. A specific answer has already been assumed to be WRONG. Read both arguments and select the single best answer from the remaining options. Respond with ONLY a single letter: A, B, C, or D. Nothing else.

User:
Question: `{question}`
Choices: `{choices}`
(`{wrong}`) WRONG.
Valid options: `{valid}`
Agent A: "`{arg_a}`"
Agent B: "`{arg_b}`"
Which answer is correct? Reply with one letter only.

Figure 4: Prompt templates for the three debate roles used in Stage 3. Curly-brace placeholders are filled at runtime.

6.5 MCQ Results: Development Phase

Language	GPT-OSS 20B	Qwen3-8B	Gemma-2 9B	Llama-3 8B	Qwen2.5-7B	Command-R 7B	Llama-3.1 8B	Qwen3-4B Think.	SeaLLMs 7B	Qwen3-4B Inst.	Llama-3.2 3B	Gemma-2 2B	Qwen3-1.7B	Qwen2.5-1.5B	Gemma-3 12B	Llama-3.2 1B
Overall	**89.19**	**82.43**	**81.08**	**77.70**	**77.70**	**77.03**	**76.35**	**76.35**	**73.65**	**72.97**	**68.92**	**66.22**	**64.19**	**61.49**	**53.38**	**32.43**
ar-EG	**85.71**	85.71	57.14	71.43	85.71	71.43	71.43	85.71	57.14	57.14	71.43	42.86	42.86	57.14	57.14	57.14
ar-MA	**100.00**	100.00	100.00	71.43	100.00	100.00	71.43	57.14	85.71	42.86	71.43	42.86	57.14	71.43	57.14	14.29
ar-SA	**71.43**	42.86	42.86	28.57	57.14	71.43	42.86	57.14	42.86	14.29	28.57	28.57	42.86	28.57	42.86	28.57
bg-BG	85.71	71.43	**100.00**	85.71	71.43	57.14	85.71	71.43	57.14	71.43	28.57	57.14	85.71	57.14	71.43	0.00
el-GR	**100.00**	80.00	80.00	100.00	60.00	80.00	100.00	80.00	60.00	80.00	60.00	80.00	60.00	40.00	80.00	40.00
en-AU	**100.00**	100.00	85.71	85.71	85.71	85.71	85.71	71.43	71.43	85.71	85.71	71.43	85.71	85.71	57.14	57.14
en-GB	**100.00**	100.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	100.00	100.00	80.00	80.00	80.00	40.00	60.00
es-EC	**100.00**	100.00	87.50	75.00	75.00	75.00	75.00	62.50	87.50	62.50	75.00	75.00	25.00	50.00	37.50	12.50
es-ES	**100.00**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	20.00
es-MX	**100.00**	100.00	100.00	100.00	100.00	80.00	100.00	100.00	80.00	100.00	100.00	80.00	100.00	80.00	40.00	20.00
eu-ES	**85.71**	57.14	85.71	85.71	71.43	42.86	57.14	71.43	42.86	57.14	71.43	42.86	28.57	71.43	42.86	28.57
fa-IR	**100.00**	100.00	100.00	80.00	80.00	80.00	80.00	100.00	80.00	80.00	80.00	80.00	100.00	100.00	80.00	60.00
fr-FR	62.50	37.50	50.00	37.50	37.50	50.00	37.50	37.50	50.00	62.50	**75.00**	50.00	37.50	37.50	25.00	12.50
ga-IE	57.14	85.71	57.14	**100.00**	42.86	42.86	71.43	57.14	57.14	57.14	71.43	57.14	57.14	71.43	71.43	42.86
id-ID	**100.00**	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	80.00	40.00
ja-JP	**85.71**	85.71	85.71	71.43	85.71	85.71	71.43	85.71	85.71	85.71	71.43	71.43	57.14	71.43	57.14	57.14
ko-KR	60.00	40.00	60.00	40.00	60.00	**100.00**	60.00	40.00	40.00	60.00	20.00	40.00	40.00	20.00	60.00	20.00
ms-SG	**100.00**	100.00	100.00	100.00	100.00	85.71	100.00	100.00	100.00	100.00	71.43	100.00	71.43	71.43	28.57	28.57
ta-LK	85.71	71.43	85.71	71.43	**100.00**	85.71	71.43	71.43	71.43	57.14	28.57	57.14	28.57	42.86	42.86	71.43
ta-SG	**100.00**	85.71	85.71	85.71	57.14	85.71	85.71	100.00	85.71	71.43	71.43	71.43	100.00	28.57	42.86	14.29
tl-PH	**87.50**	87.50	75.00	75.00	75.00	75.00	75.00	75.00	87.50	75.00	75.00	75.00	62.50	25.00	50.00	12.50
zh-CN	**100.00**	100.00	80.00	80.00	100.00	80.00	80.00	100.00	100.00	100.00	60.00	80.00	80.00	100.00	60.00	40.00
zh-SG	**100.00**	85.71	85.71	85.71	85.71	85.71	85.71	85.71	85.71	100.00	85.71	71.43	85.71	71.43	42.86	28.57

Table 7: MCQ accuracy (%) per model and language locale. Models are sorted by overall score in descending order. There are 23 language codes and 148 total questions in the development phase. **Bold gold** cells indicate the best-performing model for each language row.

Language Codes	# Questions
tl-PH, fr-FR, es-EC	8
zh-SG, ta-SG, ta-LK, ms-SG, ja-JP, ga-IE, eu-ES, en-AU, bg-BG, ar-SA, ar-MA, ar-EG	7
zh-CN, ko-KR, id-ID, fa-IR, es-MX, es-ES, en-GB, el-GR	5

Table 8: Number of MCQ questions per language code in Track 2 development phase.

6.6 Human Evaluation Annotation Prompt

Dear Participant,

Thank you sincerely for volunteering to take part in this study. Your participation plays an important role in improving the quality of this research.

This task consists of 500 questions in Persian. The goal of this evaluation is to assess everyday knowledge related to the culture, customs, lifestyle, language, and general knowledge commonly found in Iranian society. Please read each question carefully and evaluate the provided answer solely based on your personal knowledge and cultural experience.

Scoring:

- If the provided answer is fully correct and culturally valid, enter **1** in the "Score" column.
- If the answer is incorrect, misleading, culturally invalid, inappropriate, or irrelevant, enter **0**.

Important Notes:

- Please base your evaluation only on your own knowledge and personal judgment.
- Strictly avoid using any external resources such as Google, books, social media, or AI tools.
- If you are uncertain about an answer, decide based on the general and common perception within Iranian culture.
- Please respond carefully and attentively so that the evaluation results are of the highest quality.

شرکت‌کننده‌ی گرامی،

از شما بابت داوطلب شدن برای انجام این مطالعه صمیمانه سپاسگزاریم. مشارکت شما نقش مهمی در بهبود کیفیت این پژوهش دارد.

این کار شامل ۵۰۰ پرسش به زبان فارسی است. هدف این ارزیابی، سنجش دانش روزمره مرتبط با فرهنگ، آداب‌ورسوم، سبک زندگی، زبان و دانش عمومی رایج در جامعه ایران است. لطفاً هر پرسش را با دقت مطالعه کرده و پاسخ ارائه‌شده را صرفاً بر اساس دانش شخصی و تجربه فرهنگی خود ارزیابی کنید.

نحوه امتیازدهی:

- اگر پاسخ ارائه‌شده به پرسش کاملاً صحیح و از نظر فرهنگی معتبر است، عدد 1 را در ستون «امتیاز» وارد کنید.
- اگر پاسخ نادرست، گمراه‌کننده، از نظر فرهنگی نامعتبر، نامناسب، یا غیرمرتبط است، عدد 0 را وارد کنید.

نکات مهم:

- لطفاً ارزیابی خود را فقط بر اساس دانش و قضاوت شخصی انجام دهید.
- از استفاده از هرگونه منبع خارجی مانند گوگل، کتاب‌ها، شبکه‌های اجتماعی، یا ابزارهای هوش مصنوعی جداً خودداری کنید.
- اگر در مورد پاسخ تردید دارید، بر اساس برداشت کلی و رایج در فرهنگ عمومی ایران تصمیم‌گیری کنید.
- لطفاً با دقت و تمرکز پاسخ دهید تا نتایج ارزیابی از بالاترین کیفیت برخوردار باشد.

Figure 5: Annotation instructions provided to Persian-speaking human evaluators: English (left) and original Persian (right).

6.7 SAQ vs. MCQ Failure Examples for Persian

Table 9 shows two Persian examples with identical questions where MCQ is correct but SAQ is not, illustrating distinct failure modes of open-ended generation. In the first, the model hedges by naming both rival teams (استقلال و پرسپولیس, "Esteghlal and Persepolis"), semantically correct but failing lemma matching, while MCQ forces a single correct choice. In the second, the model generates the wrong answer in SAQ format, yet identifies the correct choice in MCQ format.

Table 9: Persian examples where MCQ is correct but SAQ is not.

Track	Question	Prediction	Gold	✓/×
MCQ	What is the most popular sport team in Iran? (A) america (B) mc algiers (C) **persepolis** (D) real madrid	*persepolis*	persepolis	✓
SAQ	درایران محبوب‌ترین تیم ورزشی کدام است؟	استقلال و پرسپولیس	persepolis	×
MCQ	Which one of the daily meals is commonly shared with family members in Iran? (A) breakfast (B) **dinner** (C) morning (D) snacks	*dinner*	dinner	✓
SAQ	درایران کدام یک از وعده‌های غذایی روزانه معمولاً با اعضای خانواده به اشتراک گذاشته می‌شود؟	ناهار (lunch)	dinner	×