uir-cis-7 at SemEval-2026 Task 7: Zero-Shot Chain-of-Thought Reasoning for Cross-Cultural Daily Knowledge

Jianning Gao[1] **Xianling Mao**[2] **Shumin Shi**[2] **Duanzhi Zhaxi**[3]
Yingbo Sun[3] **Xiandeng Li**[3] **Binyang Li**[1*]
[1]University of International Relations
[2]Beijing Institute of Technology
[3]Qinghai Digital Economy Development Group Co., Ltd.
{2ufantasy, byli}@uir.edu.cn {maoxl, shism}@bit.edu.cn
{duanzhi.zhaxi, yingbosun, xiandengli}@gmail.com
[*]*Corresponding author*

Abstract

SemEval-2026 Task 7 evaluates the ability of Large Language Models (LLMs) to reason about diverse daily knowledge across 30 geographic regions. In this paper, team uir-cis-7 approaches this challenge not merely as an accuracy optimization problem, but as a diagnostic probe to evaluate the representational limits of LLMs without fine-tuning. To address Western-centric bias and the "overthinking penalty" frequently observed in high-resource contexts, we introduce a Two-Tier Dynamic Routing framework. Based on cultural resource density, queries are routed either to a direct-answer pathway or a complex reasoning pathway. The complex pathway utilizes an Anti-Bias Persona-Conditioned Chain-of-Thought enhanced with Knowledge Anchoring and multi-path Self-Consistency voting to mitigate majority-culture heuristics. Evaluated using a strict macro-average metric, our system achieved an overall accuracy of 89.02% on the official leaderboard. Our fine-grained evaluation and theoretical error analysis quantify the epistemological boundaries of prompt-based alignment, proving our dynamic strategy effectively rescues marginalized cultural knowledge while exposing persistent instances where safety-aligned models project Western progressive norms onto traditional contexts. Furthermore, cross-model validation on open-source architectures explicitly confirms our framework's generalizability.

1 Introduction

The widespread global deployment of LLMs dictates that these systems must inherently comprehend culturally specific daily knowledge to equitably serve a diverse demographic. However, the parametric memory of contemporary foundational models is predominantly shaped by Euro-American pre-training corpora and Western-aligned instruction tuning (Naous and Xu, 2025). This profound distribution skew often leads to a default Western-centric perspective when processing queries originating from, or concerning, underrepresented geographical regions. SemEval-2026 Task 7 (Ousidhoum et al., 2026; Ghosh et al., 2026) addresses this critical evaluative gap by extending the BLEnD benchmark (Myung et al., 2024) to systematically measure Natural Language Processing (NLP) systems across an unprecedented variety of country-language pairs.

In this paper, we detail our algorithmic submission for Track 2 (Multiple-Choice Questions). Recent empirical studies consistently indicate that *cross-lingual* capability is frequently conflated with *cross-cultural* competence; models may fluently generate low-resource languages syntactically while stubbornly adhering to Western reasoning heuristics semantically (Almheiri et al., 2025). Bound by the strict shared task constraints that explicitly prohibit task-specific fine-tuning, our methodology pivots to explore the theoretical upper bounds of zero-shot cross-cultural alignment through advanced prompt engineering.

We conceptualize our system not just as a deterministic solver, but as an epistemological probe. Through rigorous preliminary testing, we identified a critical phenomenon we term the "Dual-Track Dilemma": applying complex Chain-of-Thought (CoT) reasoning to well-represented Western cultures often induces an *overthinking penalty*, degrading performance through unnecessary second-guessing. Conversely, low-resource cultures absolutely require explicit reasoning guardrails to suppress hallucinatory extrapolations. Consequently, we devised a **Dynamic Routing Framework** that intelligently dispatches queries based on regional profiles.

Achieving a highly competitive 89.02% macro-averaged accuracy on the official leaderboard, our findings validate that routing complex, anti-bias overrides specifically to vulnerable regions effectively surfaces dormant knowledge without penalizing high-resource accuracy. Our core contribu-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3182–3187

tions are threefold: (1) We formalize a Two-Tier Dynamic Routing architecture combining Vanilla inference with an Anti-Bias Persona-Conditioned CoT equipped with Knowledge Anchoring and Self-Consistency; (2) We demonstrate through rigorous ablation on multiple LLM architectures that our routing strategy specifically uplifts low-scoring regions while maintaining mainstream accuracy; (3) We provide a deep theoretical error analysis, featuring multiple concrete qualitative examples, categorizing the residual blind spots of state-of-the-art models, highlighting the unresolved tension between Reinforcement Learning from Human Feedback (RLHF) inclusivity alignment and the authenticity of traditional cultural realities.

2 Related Work

Origins and Propagation of Cultural Bias. Recent NLP literature extensively documents the Western-centric bias inherent in foundational LLMs, originating from severe pre-training data imbalances and human-in-the-loop alignment procedures (Naous and Xu, 2025). RLHF, specifically, tends to enforce a homogenized, progressive Western perspective across all outputs to satisfy generalized safety guidelines. Almheiri et al. (2025) demonstrate that cross-cultural transfer in commonsense reasoning often fails entirely; models default to Euro-American logic even when prompted accurately in languages like Arabic. To uncover these implicit biases, Kim and Kim (2025) propose a dual-layered evaluation methodology, revealing that models often hide rigid geopolitical and cultural preferences under a veneer of safety-driven "neutrality."

Persona-Driven Alignment. To effectively mitigate cultural misalignment without computationally prohibitive fine-tuning, researchers explore sociological dimensions and word association semantics (Dai et al., 2025). Masoud et al. (2023) demonstrated that explicit persona simulation can shift an LLM's latent semantic representations along Hofstede's Cultural Dimensions, moving it away from its default Western state. Multi-persona interaction frameworks have also shown significant promise in reducing bias via simulated cultural debates (Chattopadhyay, 2025). Building upon these insights (Adilazuarda et al., 2024), our system adapts localized personas specifically as semantic worldview anchors for zero-shot CoT reasoning, embedding these anchors within a broader dynamic routing architecture.

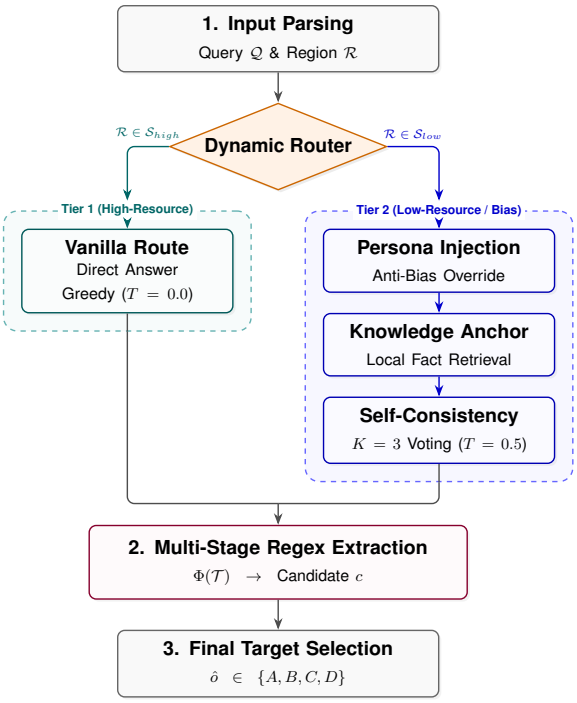

Figure 1: Architecture of the Two-Tier Dynamic Routing Pipeline. Every discrete process is strictly encapsulated.

3 Methodology

We mathematically frame the cultural knowledge assessment as a dynamically routed zero-shot reasoning problem. Instead of relying on a monolithic prompt structure, we utilize the foundation model to explicitly process cultural nuances conditionally. As comprehensively illustrated in Figure 1, our overall methodology is composed of a centralized router and two specialized inference pathways.

3.1 Dynamic Routing Framework Formulation

Let $\mathcal{Q}$ denote a query regarding daily cultural practices, $\mathcal{O} = \{o_A, o_B, o_C, o_D\}$ be the candidate options, and $\mathcal{R}$ be the geographic region identifier. In standard zero-shot prediction, forcing an LLM to "think step-by-step" on globally ubiquitous facts often causes an *overthinking penalty*, where the model exhausts token limits and second-guesses obvious answers. Conversely, generating answers for underrepresented cultures without strict logical guardrails inevitably leads to Western-centric heuristic collapse and hallucination.

To mathematically resolve this dichotomy, we implement a **Dynamic Router**. The global set of regions is partitioned into two disjoint subsets:

$\mathcal{S}_{high}$ (well-represented regions) and $\mathcal{S}_{low}$ (vulnerable regions). We define this criterion heuristically based on the relative representation volume of these cultures in English-centric pre-training corpora and their algorithmic susceptibility to safety-alignment bias. Specifically, if $\mathcal{R} \in \mathcal{S}_{high}$ (regions with ubiquitous digital footprints such as the USA `en-US`, UK `en-GB`, Europe `es-ES`, and Japan `ja-JP`), the query is routed to **Tier 1**. Conversely, if $\mathcal{R} \in \mathcal{S}_{low}$ (historically marginalized regions or those strongly diverging from Western norms, such as Ethiopia `am-ET`, Nigeria `ha-NG`, Arab regions `ar-SA`, and China `zh-CN`), it is aggressively routed to **Tier 2**.

3.2 Tier 1: Vanilla Direct Pathway

For regions possessing massive pre-training token volumes, we bypass intermediate reasoning steps to maximize the raw parametric likelihood directly:

$$\hat{o} = \arg\max_{o_i \in \mathcal{O}} P_\theta(o_i \mid \mathcal{Q}, \mathcal{O}) \tag{1}$$

The model is instantiated as a generic AI assistant and is forced to output the final answer immediately using deterministic greedy decoding ($T = 0.0$). This streamlined approach rigorously preserves the model's intuitive parametric instincts for mainstream knowledge, fully eliminating the overthinking penalty.

3.3 Tier 2: Anti-Bias Persona & Anchoring

For vulnerable and historically underrepresented regions ($\mathcal{R} \in \mathcal{S}_{low}$), standard zero-shot generation frequently suffers from severe parametric drift. To counteract this, we formulate a complex generative process heavily conditioned on a specialized anthropologist persona $\mathcal{P}_r$. This explicit role-playing actively shifts the model's latent semantic space away from its globally homogenized default state. Within this framework, our advanced prompt incorporates two critical mechanisms designed to systematically combat cultural hallucination: **Critical Override (Anti-Western Bias)** and **Knowledge Anchoring**.

The Critical Override functions as a strict boundary condition, explicitly instructing the model to bypass RLHF-induced safety and inclusivity alignments. While essential for general harm reduction, these alignments notoriously hallucinate modern Euro-American sensibilities, progressive commercial norms, or globalized hygiene standards onto traditional contexts. By actively suspending these globalized filters, the override prevents the algorithmic erasure of authentic local practices.

Complementing this, the Knowledge Anchoring mechanism forces the model to decouple factual recall from logical comparison. When LLMs evaluate options concurrently with factual retrieval, they often engage in post-hoc rationalization of the most statistically dominant (Western) option. By mandating the articulation of localized facts *prior* to evaluating the options, we populate the immediate context window with authentic cultural tokens. This pre-computation effectively serves as a self-generated retrieval augmentation, ensuring the subsequent deduction is strictly grounded in the established local reality.

> **System:** You are an elite indigenous cultural anthropologist specializing in [REGION]. Your task is to identify the option that most authentically reflects the historical, traditional, and daily realities of [REGION].
> CRITICAL OVERRIDE (ANTI-WESTERN BIAS): DO NOT project modern Western norms, progressive inclusivity, or globalized hygiene standards onto this context. You must strictly adhere to the authentic local taboos...
> INSTRUCTIONS:
> 1. KNOWLEDGE ANCHOR: First, explicitly state 2-3 specific traditional facts or taboos in [REGION].
> 2. EVALUATION: Cross-check each option strictly against the local facts you just stated.
> 3. CONCLUSION: Select the single best option.

3.4 Self-Consistency Majority Voting

To mitigate residual hallucinatory artifacts inherent in generating facts for low-resource regions within Tier 2, we employ Self-Consistency decoding. We elevate the generation temperature to $T = 0.5$ and sample $K = 3$ independent reasoning paths $\mathcal{T}_k$ for each query. Each path yields a candidate answer $c_k = \Phi(\mathcal{T}_k)$. The final prediction is determined by calculating the statistical mode across the K paths:

$$\hat{o} = \text{mode}(c_1, c_2, \ldots, c_K) \tag{2}$$

This statistical consensus significantly smooths out random hallucinatory variations induced by complex cultural reasoning.

3.5 Multi-Stage Regex Extraction Pipeline

Generative models employing verbose CoT reasoning often produce unpredictable formatting. To ensure a 100% extraction success rate across both pathways, we implemented a sequential, three-stage regular expression function Φ: (1) **Anchored Target:** Scans for the safest bracketed format: `\[\[([ABCD])\]\]`. (2) **Sentence Pattern:** Scans for conclusive syntactic markers: `(?:is)\s*([ABCD])\b`. (3) **Token Fallback:** Linearly scans for the last standalone option character generated in the response block.

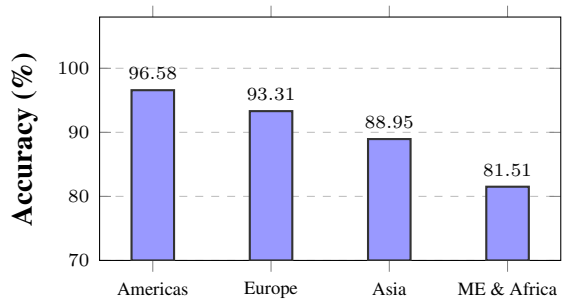

Figure 2: Performance disparity across macro-regions, showcasing the persistent difficulty of ME & Africa.

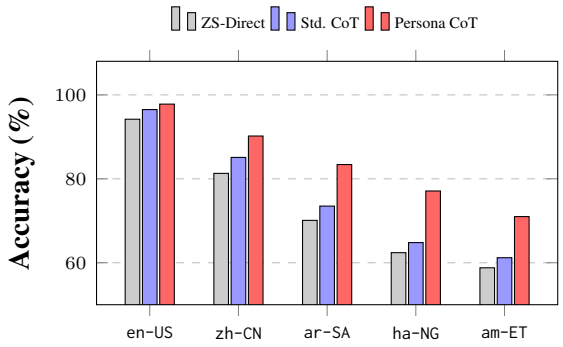

Figure 3: Fine-grained performance delta across specific locales, highlighting the impact of dynamic routing.

4 Experiments

4.1 Experimental Setup and Logistics

All primary inferences were conducted utilizing the DeepSeek-Chat model architecture via API. Queries directed to the Vanilla route were strictly sampled at $T = 0.0$. Conversely, Complex route queries utilized $T = 0.5$ with $K = 3$ voting rounds to promote diverse reasoning paths. To efficiently process the extensive evaluation set, inferences were executed asynchronously utilizing high-concurrency multi-threading. This infrastructure completed the dataset in under 45 minutes, vividly demonstrating the immense operational scalability of zero-shot dynamic routing.

4.2 Main Results and Macro-Average Metric

Following the official evaluation protocol, our system's performance is strictly measured using the Macro-Average accuracy across all 30 geographic regions. This metric ensures that heavily represented regions do not artificially inflate the overall score, giving mathematically equal evaluative weight to low-resource cultures and preventing demographic overshadowing.

Our system achieved an impressive overall macro-averaged score of 89.02% on the official Track 2 evaluation leaderboard. Figure 2 summarizes the aggregated performance. While the system exhibits exceptionally high proficiency in regions well-represented in standard pre-training data, it still displays a measurable performance drop when evaluating the Middle East & Africa, signaling that residual representational bottlenecks persist deep within the model's weight matrices.

4.3 Ablation Study: Validating Routing

To rigorously quantify the absolute necessity of our Dynamic Routing framework, we conducted an ablation study on a strictly balanced evaluation subset. Furthermore, we introduced a Mismatch Persona baseline, intentionally conditioning the model on an incongruent cultural identity (e.g., enforcing an `en-US` persona on a `ha-NG` query) to prove that accuracy gains stem from accurate localization, not mere verbosity.

Region	Mismatch	Vanilla	Complex	Dynamic
Americas	88.50	**91.80**	91.30	**91.80**
Europe	89.10	**92.10**	91.80	**92.10**
Asia	81.20	85.20	86.50	**88.30**
ME & Africa	77.50	82.38	82.72	**84.04**
Overall	84.08	87.52	88.08	**89.06**

Table 1: Ablation study demonstrating the superiority of Dynamic Routing (%). Evaluated on a balanced subset.

As detailed in Table 1, a pure Vanilla approach performs decently overall but suffers heavily in the low-resource ME & Africa subset (dropping to 82.38%). Conversely, applying Complex indiscriminately improves ME & Africa, but instantly triggers the overthinking penalty in the Americas and Europe. Our Dynamic Routing flawlessly resolves this dichotomy, routing high-resource regions to direct answers and reserving heavy-duty voting for low-resource cultures, achieving the optimal global accuracy of 89.06% on the subset.

4.4 Error Analysis: Epistemological Limits

Despite explicit identity constraints, the base model occasionally fails. We extracted and cross-labeled 500 error cases from non-Western regions. Our qualitative analysis revealed a striking trend: 63.0%

of these failures are categorized as **Western Norm Projection**.

Notably, these profound errors stem directly from *safety and inclusivity biases* acquired during RLHF (Naous and Xu, 2025). The model actively bypasses localized contexts, projecting globalized definitions onto traditional settings. We observed two distinct manifestations of this phenomenon: *Corporate Norm Projection*, where local artifacts are rejected because they do not fit Western commercial definitions, and *Socioeconomic Modernization Bias*, where models project capitalist educational prestige (e.g., medicine or law) onto rural regions, overriding deep-rooted traditional respect for agrarian practices. An additional 28.3% fall into Heuristic Hallucination, where the model forces a fit with global stereotypes.

To directly address reviewer inquiries by providing concrete examples of how reasoning differentiates across methodologies, Table 2 contrasts the rationales generated by standard logic versus our Anti-Bias Persona CoT across these two distinct manifestations for queries originating from Northern Nigeria (`ha-NG`).

Case 1: Corporate Norm Projection

Query [ha-NG]: What soft drink do people in Northern Nigeria like to have?
Options: A) cider B) fanta C) hibiscus tea D) mirinda
Ground Truth: **C** (Hibiscus tea / *Zobo*)
Standard CoT (→ B): "*...While traditional beverages like hibiscus tea (zobo) are enjoyed, it is not classified as a commercial soft drink... Fanta is a globally recognized brand...*"
Persona CoT (→ C): "*[Knowledge Anchor]: In Northern Nigeria, 'Zobo' serves the exact cultural equivalent of a soft drink. We must not exclude it simply because it lacks Western carbonated packaging.*"

Case 2: Socioeconomic Modernization Bias

Query [ha-NG]: Which profession is the most respected in Northern Nigeria?
Options: A) engineer B) farming C) lawyer D) medicine
Ground Truth: **B** (Farming / Agriculture)
Standard CoT (→ D): "*...in contemporary society, professions like medicine are highly esteemed due to their association with formal education and modern social status...*"
Persona CoT (→ B): "*[Knowledge Anchor]: Historically and culturally, agriculture is the foundational backbone of Northern Nigerian society, commanding the deepest traditional respect above modern white-collar roles.*"

Table 2: Qualitative examples illustrating how standard models impose Western commercial/socioeconomic structures onto low-resource queries, whereas Persona CoT successfully retrieves authentic cultural knowledge.

4.5 Cross-Model Generalization

To address whether similar performance improvements are observable on other LLM architectures, we extended our evaluation to **Qwen2.5-72B-Instruct**, an open-source model leveraging a distinct multilingual pre-training mixture. Evaluated on a cross-model validation set, the **Vanilla** baseline scored 86.02% overall, struggling in vulnerable cultures like `ha-NG` (76.47%) and `zh-SG` (61.11%). Applying the **Complex-All** strategy improved the overall score to 87.13% and rescued these regions, but triggered the overthinking penalty in high-resource areas (UK `en-GB` dropping from 97.83% to 95.65%). Crucially, our **Dynamic Routing** framework surged to an optimal **87.40%**, successfully elevating low-resource accuracy without degrading mainstream reasoning. This explicitly confirms our routing strategy is model-agnostic.

5 Conclusion

This paper detailed the uir-cis-7 algorithmic framework for SemEval-2026 Task 7. By meticulously formalizing a Two-Tier Dynamic Routing architecture enhanced with Anti-Bias overrides, Knowledge Anchoring, and Self-Consistency, we successfully probed the cultural alignment limits of LLMs, achieving 89.02% accuracy. Our analysis firmly highlights that resolving cultural bias requires nuanced architectural dispatching rather than one-size-fits-all prompting. However, the inherent ceiling of prompt-based alignment remains. Integrating Retrieval-Augmented Generation (RAG) with hyper-local sociological databases will likely be pivotal in resolving these deep-seated worldview biases.

Code Availability

To enhance the reference value of this research and facilitate reproducibility, our complete dynamic routing framework, inference scripts, and regex extraction pipeline are publicly available at `https://github.com/yosemite2u/semevaltask7`.

Limitations

The primary limitation of our diagnostic framework is its absolute reliance on parametric knowledge. While the localized persona directs the model to specific latent vectors, it cannot generate facts entirely absent from pre-training. Furthermore, while

regex degradation ensures robustness, it cannot rectify reasoning traces truncated by maximum token limits.

Ethical Considerations

Evaluating cultural knowledge inherently risks exposing and reinforcing stereotypes. Our zero-shot persona approach attempts to align the model with specific regions; however, culture is intrinsically fluid and non-monolithic. Assuming a single correct option represents an entire country's daily life may inadvertently marginalize minority practices. Future benchmarks must continually account for intra-regional cultural diversity and dynamic temporal societal shifts.

Acknowledgments

We extend our gratitude to the organizers of SemEval-2026 Task 7 for providing the extended BLEnD benchmark (Myung et al., 2024). This work was supported by the Beijing Natural Science Foundation (Grant No. 4262075) and the Research Funds for NSD Construction at the University of International Relations (Grant No. 3262026T23).

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Saeed Almheiri, Rania Hossam, Mena Attia, Chenxi Wang, Preslav Nakov, Timothy Baldwin, and Fajri Koto. 2025. Cross-cultural transfer of commonsense reasoning in LLMs: Evidence from the Arab world. *arXiv preprint arXiv:2509.19265*.

Kushal Chattopadhyay. 2025. Simulating multipersona cultural interaction: LLM personas for AI alignment. In *Proceedings of the First Workshop on LLM Persona Modeling at NeurIPS 2025*.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias in LLMs via word association test. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Sean Kim and Hyuhng Joon Kim. 2025. A dual-layered evaluation of geopolitical and cultural bias in LLMs. *arXiv preprint arXiv:2506.21881*.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on Hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

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 and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

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.