Pinetree at SemEval-2026 Task 7: A Large-Scale Failure Analysis of Cultural Grounding in Language Models

Yen Yee Yam
Pinetree Research
National University of Singapore
e1354172@u.nus.edu

Hong Meng Yam
Pinetree Research
Stanford University
hongmeng@stanford.edu

Abstract

Using a simple prompting strategy without fine-tuning or retrieval augmentation, our system achieved 88.85% micro-average and 90.55% macro-average accuracy, ranking #4 overall on SemEval-2026 Task 7. Our primary contribution is a failure analysis of 5,241 incorrect predictions (11.15% of the dataset), categorized using the six-topic BLEnD taxonomy. Errors concentrate in Food (39.42%) and Holidays/Celebration/Leisure (15.76%), but within-topic error rates are highest on Family (21.04%) and Work life (20.45%), which topics with limited representational density. Global-brand attractor errors account for only 2.50% of failures and are tightly localized: 98.5% fall on a single template (most popular sport team) in four low-resource cultures. Outside these templates, brand-default effects are statistically negligible. These findings support representational sparsity and knowledge-density asymmetry, not ideological skew, as the dominant cause of cultural misalignment in everyday behavioral tasks.

1 Introduction

Cultural alignment in large language models has primarily been studied through survey-based value instruments such as the World Values Survey and Hofstede dimensions (Tao et al., 2024; Cao et al., 2023), which document skew toward Anglosphere and Western European value clusters. Yet *everyday behavioral knowledge*, the routine, majority-default practices of daily life, has received far less scrutiny, despite being where most user-facing cultural failures actually occur.

SemEval-2026 Task 7 Track 2 (Ghosh et al., 2026; Ousidhoum et al., 2026) targets this dimension directly through multiple-choice questions about contemporary everyday behavior across 30 locales. We show that a frozen LLM with carefully designed prompting alone, with no fine-tuning, retrieval, or ensembling, achieves 88.85% micro- and 90.55% macro-average accuracy, ranking #4 overall.

Our central contribution, however, is what comes after the leaderboard: to our knowledge, the first large-scale failure analysis of cultural grounding on a behavioral benchmark, covering all 5,241 errors organized along the six-topic BLEnD taxonomy (Myung et al., 2024). Three findings stand out. First, errors concentrate not where intuition predicts (Food, in absolute terms) but where representational density is lowest (Family, Work life, by within-topic rate). Second, contrary to common assumptions about systemic Western-default collapse, global-brand attractor effects are template-bounded: 98.5% fall on a single sport-team question in four low-resource cultures, and are statistically negligible elsewhere. Third, numeric institutional errors are overwhelmingly rank-adjacent (71.1%), indicating fuzzy-but-close knowledge rather than absent priors. Together, these reframe cultural misalignment in everyday tasks as representational sparsity, not ideological skew.

2 Related Work

2.1 Cultural Bias and Value Alignment in LLMs

A growing literature evaluates cultural alignment in LLMs using structured survey instruments. Tao et al. (2024) show that GPT-family models cluster near self-expression–oriented societies on the Inglehart–Welzel cultural map when answering World Values Survey (WVS) items without cultural prompting. Similarly, Cao et al. (2023) find systematic alignment with American value patterns across Hofstede dimensions. These studies frame cultural bias primarily as skew in expressed moral or political values.

Beyond value surveys, broader work on social bias in language models has documented demographic and cultural skew in text generation

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3399–3407

(Nadeem et al., 2021). In multimodal systems, Ross et al. (2021) show that representational biases can be structurally embedded in model architectures, suggesting that bias is not purely a surface prompting artifact. However, these analyses largely focus on normative attitudes or protected attributes rather than modeling everyday behavioral typicality.

2.2 Cultural Prompting and Steering

Prompt-based steering strategies have been explored to reduce cultural misalignment. Explicitly instructing models to answer from the perspective of an "average person" in a given country can reduce measured cultural distance on WVS-style evaluations, though improvements are uneven and sometimes unstable (Tao et al., 2024). More generally, work on controllable generation and alignment demonstrates that prompt framing can meaningfully influence model outputs without parameter updates (Ouyang et al., 2022; Bai et al., 2022).

While these approaches show that cultural perspective can be partially modulated at inference time, they predominantly target value alignment and normative positioning. They do not systematically analyze how models fail on concrete, majority-default everyday behaviors across diverse national contexts.

2.3 Everyday Cultural Knowledge and Benchmarking

Benchmarks evaluating factual and commonsense reasoning across languages are well established (Hendrycks et al., 2020; Fujinuma et al., 2023), yet they typically emphasize encyclopedic knowledge or general commonsense rather than culturally specific daily-life practices. SemEval-2026 Task 7 Track 2 builds on the BLEnD benchmark (Myung et al., 2024), which explicitly evaluates culturally grounded everyday behavior through forced-choice questions about contemporary life across 16 countries and six topics. This design isolates majority-default practices that are socially unremarkable yet culturally distinctive.

To our knowledge, prior work has not conducted large-scale failure analyses of modern LLMs on everyday cultural behavioral benchmarks. Existing cultural alignment studies quantify aggregate value distance but rarely dissect error types or analyze structural causes of misprediction at scale. Our work addresses this gap by providing a systematic empirical characterization of 5,241 incorrect predictions, identifying dominant failure modes within each of the six BLEnD topics and quantifying cross-cutting failure mechanisms (brand-attractor capture, answer-space mismatch, numeric rank-adjacency) orthogonally to topic.

3 System Description

3.1 Model and Inference Configuration

Our system uses GPT-5.2 as a frozen, general-purpose language model. No fine-tuning, supervised training, retrieval augmentation, web search, or external knowledge sources are employed. All predictions are generated purely via prompting at inference time.

Decoding is performed with fixed hyperparameters (temperature = 0.0, maximum output length = 40 tokens). No chain-of-thought or intermediate reasoning traces are requested. Each instance is processed independently in a single-pass pipeline.

3.2 Prompting Strategy

Each multiple-choice question is formatted into a structured prompt that instructs the model to select the option that best reflects mainstream, contemporary everyday behavior in the specified country.

The prompt incorporates lightweight behavioral constraints designed to reduce common failure modes observed during development. Specifically, the model is encouraged to:

- Prefer majority-default behaviors that are widely practiced in ordinary daily life,
- Prioritize nationally salient norms when regional variation exists,
- Avoid selecting options based purely on lexical overlap with the question,
- Disregard exaggerated stereotype cues unless they reflect genuine mainstream practice,
- Anchor judgments in contemporary behavior rather than historical or ceremonial customs.

These constraints are implemented as natural-language instructions within the prompt template. They do not modify model parameters and introduce no additional supervision. The full prompt template and complete constraint wording are provided in Appendix A.2 for reproducibility.

3.3 Output Formatting and Validation

The model is required to output exactly one capital letter (A/B/C/D). Outputs are parsed automatically. If a response contains additional text or cannot be parsed into a valid label, a deterministic fallback rule selects the first valid label present in the output to guarantee exactly one prediction per instance.

Predictions are written to the required one-hot TSV format, with exactly one positive label per row, as specified by the SemEval-2026 Task 7 Track 2 guidelines.

4 Experimental Setup

We evaluate on SemEval-2026 Task 7 Track 2 (MCQ), which consists of 47,014 multiple-choice questions about contemporary everyday behavior across 30 countries and language settings, extending the BLEnD benchmark (Myung et al., 2024) with additional locales.

No parameter updates are performed at any stage. The official development split is used solely for prompt wording stabilization and output formatting verification. The final submitted system applies a fixed prompt template and fixed decoding configuration to the full test set without modification. Each instance is processed independently, without batching-based cross-instance information sharing. No external corpora, lexicons, or handcrafted knowledge bases are used.

5 Results

Our system achieves:

- **Micro-average accuracy: 88.85%**
- **Macro-average accuracy: 90.55%**

This performance ranked #4 overall on the official SemEval-2026 Track 2 leaderboard.

The macro-average exceeding the micro-average indicates relatively balanced performance across countries rather than dominance in high-frequency or high-resource locales. This suggests that the simple prompting strategy generalizes reasonably well across diverse cultural contexts.

Country-level accuracy ranges from 69.96% to 99.28%. Lower performance is observed in locales characterized by greater internal heterogeneity or limited globally accessible documentation of everyday practices, while higher-performing locales tend to exhibit more stable and widely represented daily-life norms. A full per-country breakdown is provided in Appendix A.1.

Despite the absence of fine-tuning or retrieval augmentation, the system achieves competitive leaderboard performance. In the following section, we move beyond aggregate accuracy and conduct a detailed failure analysis of 5,241 incorrect predictions (11.15% of the dataset) to identify systematic limitations in culturally grounded reasoning.

6 Failure Analysis

We analyze all 5,241 incorrect predictions (11.15% of the dataset) along three axes: **(i) topic distribution**, using the six-topic taxonomy defined by the BLEnD authors (Myung et al., 2024); **(ii) cross-cutting failure mechanisms** orthogonal to topic; and **(iii) country-level distributional variation**.

Topic labels are not produced by our heuristic; they are recovered by joining each MCQ row to its base BLEnD question ID and looking up the authoritative topic from the BLEnD release. The join uses the BLEnD MCQ release as the primary source (94.8% of error rows) with the BLEnD SAQ release as fallback for SemEval-added items not present in BLEnD's MCQ (5.2%). All 5,241 error rows receive a topic label, and BLEnD's ID-to-Topic mapping is verified consistent across all 16 country files.

6.1 Topic-Level Distribution and Error Rates

Errors are unevenly distributed across topics in absolute count (Table 1), but the more revealing statistic is the within-topic error rate, i.e., what fraction of questions in a topic the model gets wrong, controlling for topic frequency.

The most informative pattern is the divergence between absolute error count and within-topic error rate. Food dominates absolute count (39.42%) but has only a moderate within-topic rate (13.63%), reflecting its 32% share of the benchmark. Family inverts this: it contributes only 4.56% of errors yet has the highest within-topic rate (21.04%), consistent with the lower BLEnD inter-annotator agreement on this topic (Myung et al., 2024, Table 7) and with genuine within-country pluralism on questions such as *"What is a popular family activity with a child to do on weekends in X?"*. Work life is the second-hardest topic (20.45%), driven primarily by numeric institutional questions (e.g., *"At what age do most people start working in X?"*) rather than behavioral knowledge, as shown in Section 6.3. Sport

Topic (Myung et al., 2024)	Errors	% of Errors	Total Qs	Error rate within topic (%)
Food	2,066	39.42	15,157	13.63
Work life	988	18.85	4,831	**20.45**
Holidays/Celebration/Leisure	826	15.76	11,974	6.90
Education	637	12.15	3,947	16.14
Sport	485	9.25	9,969	**4.87**
Family	239	4.56	1,136	**21.04**
Total	**5,241**	**100.00**	**47,014**	**11.15**

Table 1: Distribution of errors across the six BLEnD topics (Myung et al., 2024). Topic labels are recovered from BLEnD's authoritative ID-to-Topic mapping. The within-topic error rate (column 5) is the more meaningful measure of model difficulty; absolute error count (column 2) reflects topic frequency in the benchmark.

has the lowest error rate (4.87%), with one concentrated brand-attractor exception analyzed in Section 6.3. Holidays/Celebration/Leisure at 6.90% is the second-easiest, contradicting the SAQ-format finding in Myung et al. (2024) that holiday questions are among the hardest, perhaps because the MCQ format substantially constrains the answer space and disproportionately benefits this topic.

6.2 Within-Topic Sub-Types

Within each topic, errors cluster around recurring question templates. The dominant sub-types per topic (each accounting for at least 15% of errors in that topic) are:

- **Food**: staple food / meals (27.1%), cooking utensils (23.1%), beverages (22.0%).
- **Work life**: age-related numeric questions (57.3% combined); profession choice (16.3%).
- **Holidays/Celebration/Leisure**: festival drink offerings (72.9%); congratulatory gestures and gifts (12.7%).
- **Education**: school-life specifics (42.4%); numeric age questions (16.5%).
- **Sport**: sport-preference questions (45.4%); brand-attractor capture on team identity (26.6%, see Section 6.3); stadium / venue (18.4%).
- **Family**: recreational activities (49.0%); food-with-family overlap (15.5%).

6.3 Cross-Cutting Failure Mechanisms

Three failure mechanisms cut across topics and warrant separate quantification (Table 2). They are orthogonal to topic and reported as flags rather than as a competing taxonomy.

Brand-attractor capture (2.50%, n=131). The model selects a globally-recognized Western brand when the gold answer is a local alternative. The distribution is sharply localized: 129 of 131 (98.5%) fall on the single question template *"What is the most popular sport team in {country}?"*, with predictions concentrated as *Real Madrid* (n=82) and *Manchester United* (n=47). Country distribution: Ethiopia (`am-ET`, n=100), Indonesia (`id-ID`, n=18), Algeria (`ar-DZ`, n=9), and South Korea (`ko-KR`, n=2). The remaining 2 cases occur on Australian payment-method questions (`en-AU`), where the model predicted *Apple Pay* or *Google Pay* over the gold answer of *cash*. Outside these two question templates, brand-default effects are negligible (zero errors across 5,108 other cases). This pattern supports a *template-bounded, localized* rather than systemic interpretation of Western-default collapse: the model exhibits clear brand priors only in domains where Western entities dominate global media coverage (international club football, US-origin tech-payment platforms).

Answer-space mismatch (0.38%, n=20). The predicted option is in a clearly different semantic space than the question requires. Notably, 13 of the 20 ASM errors come from a single question template — *"What do people do to celebrate New Year's Day in {Philippines, Sri Lanka}?"* — where the model consistently substituted *"gather with friends at a restaurant or at home, preparing a festive meal"* for activity-centric gold answers (*firecracker*, *cleaning the house*, *midnight*). The remaining 7 are scattered across Food (n=3, Morocco) and Education (n=4) on questions where

Mechanism	Count	% of Errors	Concentration
Brand-attractor capture	131	2.50	98.5% Sport (am-ET, id-ID, ar-DZ, ko-KR); 2 cases on Australian payment-method
Answer-space mismatch	20	0.38	65% are one repeated template (New Year's Day in PH/LK)
Numeric rank-adjacency	839	16.01	71.1% of all numeric errors; varies 66.1–100% by topic

Table 2: Cross-cutting failure mechanisms reported as flags orthogonal to topic. Each row's concentration column summarizes where the mechanism occurs.

Topic	Numeric err.	Adjacent	%
Family	29	29	100.0
Food	68	59	86.8
Education	241	194	80.5
Holidays	25	17	68.0
Work life	817	540	66.1
Sport	0	0	—
Overall	**1,180**	**839**	**71.1**

Table 3: Numeric rank-adjacency by topic. Adjacency means the predicted option is one rank away from the gold answer when the four numeric options are sorted by value.

the predicted entity was domain-adjacent but cross-domain. ASM is a small failure mode dominated by one recurrent pattern.

Numeric rank-adjacency (16.01%, n=839). Of 1,180 numeric errors, 71.1% are *adjacent in option-rank* to the gold answer rather than wide-margin gaps. The within-topic adjacency rate varies (Table 3). The model is rarely order-of-magnitude wrong on numeric institutional questions; it is typically one option-rank from the correct value, indicating fuzzy-but-close knowledge. This effect is most pronounced on Education and Family questions and weakest on Work life, consistent with retirement-age and labor-entry-age questions exhibiting higher genuine variation across countries than school-related numerics.

6.4 Country-Level Failure Patterns

Country-level accuracy ranges from 69.96% (`am-ET`) to 99.28% (`es-EC`). The bottom-six countries by accuracy (`am-ET`, `ha-NG`, `ko-KP`, `as-AS`, `fa-IR`, `ar-MA`) all combine three structural disadvantages: (i) lower documentation density in globally-aggregated training corpora, (ii) higher within-country cultural heterogeneity, and (iii) lower BLEnD answer-agreement scores even among native annotators (Myung et al., 2024, Table 7). The brand-attractor concentration described in Section 6.3 reinforces this pattern: all 131 brand-capture errors fall in low- or mid-resource countries; none appear in high-performing locales.

6.5 Structural Interpretation

The error distribution reveals three structural limitations in modern LLMs:

- **Granularity ceiling.** Models encode coarse cultural patterns but struggle with fine-grained majority-default calibration in the Food, Education, and Sport-preference sub-domains. Combined, these account for 60.8% of errors.
- **Discrete precision limits in numeric reasoning.** Approximate reasoning yields rank-adjacent answers on numeric institutional questions (71.1% of numeric errors), suggesting the model approximates the correct range but lacks precise calibration. This effect concentrates in the Work life and Education topics.
- **Localized brand-attractor capture.** Western-brand defaults appear, but in tightly-bounded contexts: international football team identity in low-resource cultures (n=129) and tech-payment defaults in Australia (n=2). They do not generalize to a systemic Western-default tendency across topics.

A fourth observation worth distinguishing from these structural modes is the elevated within-topic error rate on **Family** questions (21.04%) which — uniquely among topics — appears driven by genuine within-country pluralism in family-life practice rather than by representational gaps in the

model. This is supported by correspondingly low BLEnD inter-annotator agreement on Family items.

6.6 Limitations of the Categorization

The topic axis is provided directly by the BLEnD authors and inherits their topic-design decisions. Topic recovery uses the BLEnD MCQ release for 94.8% of error rows; the remaining 5.1% (n=269) used the BLEnD SAQ release as fallback because the corresponding question templates were not promoted to BLEnD's MCQ. SAQ-fallback rows are slightly over-represented in Education (35.3% of SAQ-fallback errors vs. 12.2% overall) and Food (24.9% vs. 39.4%), reflecting the topics where SemEval expanded BLEnD's MCQ inventory. Topic assignments remain authoritative because BLEnD's ID-to-Topic mapping is consistent across all 16 country files.

The cross-cutting flag definitions (brand-attractor, answer-space mismatch, numeric adjacency) are operationalized via deterministic rules over option text and a fixed brand list (full definitions in Appendix A.3). We validated the heuristic flags against an independent GPT-5.2 blind labeler over the 4,061 non-numeric errors and observed 90.91% agreement; disagreements concentrate at three taxonomy boundaries (gift vs. wedding-custom; brand-attractor vs. sport-team-topic; eating-utensil vs. dataset-quality issues with non-utensil gold answers). The brand-attractor count is robust to this validation: the labeler concurs on 117 of the 131 cases.

7 Discussion

7.1 Reframing Cultural Misalignment

Much prior literature conceptualizes cultural bias primarily in ideological or normative terms. Our findings suggest a different framing for everyday cultural grounding tasks: performance degradation is driven less by value skew and more by representational granularity and knowledge density asymmetries.

Brand-attractor effects exist but are template-bounded: 129 of 131 brand-capture errors occur on a single question template (*most popular sport team*) in four low-resource cultures, and the remaining 2 occur on Australian payment-method questions. Outside these two templates, the model does not systematically default to Western templates, failing to converge on nationally dominant defaults, with the failure mode varying by topic (granularity for Food, precision for Work life, internal heterogeneity for Family). The localized rather than systemic distribution of brand-attractor effects suggests that Western-default collapse arises from domain-specific media coverage asymmetries rather than a general ideological prior.

7.2 Implications for Cultural Evaluation

Our analysis suggests three implications:

1. Cultural grounding benchmarks should distinguish ideological bias from material-culture granularity.

2. Majority-default tasks expose representational sparsity rather than moral skew.

3. Improvements may require denser geographically localized corpora or targeted institutional knowledge augmentation.

By isolating structural error modes , this work contributes an empirical foundation for understanding where LLMs fail in culturally grounded everyday reasoning.

8 Conclusion

Herein, using a simple prompting strategy without fine-tuning or retrieval augmentation, our system achieved 88.85% micro-average and 90.55% macro-average accuracy, ranking #4 overall on the official leaderboard.

By systematically analyzing 5,241 incorrect predictions, we identify three dominant structural limitations: (i) a granularity ceiling in localized material culture knowledge; (ii) discrete precision limits in numeric institutional reasoning; and (iii) localized brand-attractor capture. We additionally observe a high within-topic error rate on Family questions (21.04%) reflecting genuine within-country pluralism rather than representational gaps in the model.

Future work should explore denser geographically localized corpora, structured augmentation of institutional knowledge, and evaluation frameworks that distinguish between ideological bias and material-culture granularity. By grounding analysis in the BLEnD topic taxonomy and large-scale empirical error patterns, this study contributes a clearer understanding of structural limitations in multilingual cultural grounding.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *Preprint*, arXiv:2303.17466.

Yoshinari Fujinuma, Siddharth Varia, Nishant Sankaran, Srikar Appalaraju, Bonan Min, and Yogarshi Vyas. 2023. A multi-modal multilingual benchmark for document image classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14361–14376.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina 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. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5356–5371.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, Online. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

A Appendix

A.1 Full Country-Level Accuracy

Table A.1 reports per-country accuracy for Track 2 (MCQ), sorted in descending order. Overall micro- and macro-averages are shown at the bottom for reference.

Country	Accuracy (%)
es-EC	99.28
bg-BG	98.77
fr-FR	98.05
en-GB	96.59
ga-IE	94.98
ar-EG	94.57
el-GR	94.37
es-ES	94.30
en-US	94.23
es-MX	94.10
su-JB	93.77
sv-SE	93.74
ta-LK	93.54
zh-CN	93.36
ko-KR	93.27
en-AU	93.18
ar-SA	93.02
ar-DZ	92.58
zh-SG	92.29
tl-PH	91.79
eu-PV	91.63
id-ID	88.42
az-AZ	87.85
ja-JP	86.83
ar-MA	81.95
fa-IR	81.32
as-AS	81.15
ko-KP	80.64
ha-NG	76.94
am-ET	69.96
TOTAL_MACRO	**90.55**
TOTAL_MICRO	**88.85**

Table: Per-country accuracy for Track 2 (MCQ).

A.2 Prompt Template for GPT-5.2 Baseline

The following prompt template is used for all GPT-5.2 runs reported in Section 3.2. The placeholders `{country}`, `{question}`, and `{A}`–`{D}` are filled

per example from the BLEnD MCQ TSV. Temperature is fixed at 0.0 and no in-context examples are provided; the seven operational constraints are the only task-specific supervision.

```
### Role
You are a Senior Cultural Anthropologist and
    Ethnographer specializing in
contemporary global social norms. You
    distinguish carefully between
official rules, stereotypes, and lived
    everyday behavior.

### Objective
Identify the most culturally authentic answer
    to a multiple-choice question
about daily life in {country}.

### Operational Constraints
1. **The 80% Rule**
   Select the option that approximately 80%
    of locals would recognize as the
   standard or default behavior in an
    ordinary, everyday situation.

2. **Current Era Bias (2026)**
   Base your judgment on how people actually
    live today. Disregard traditions,
   etiquette, or customs that exist mainly in
    textbooks, tourism narratives,
   or historical accounts.

3. **National Salience Rule**
   If behavior varies by region, class, or
    subculture, prioritize the behavior
   that carries the greatest national
    cultural weight and is most visible in
   mainstream daily life.

4. **Anti-Stereotype Filter**
   Do not choose answers that reflect foreign
    cliches or exaggerated cultural
   tropes unless they are genuinely practiced
    by locals in everyday life.

5. **Linguistic & Surface Bias Check**
   Do not select an option simply because it
    repeats words or phrasing from the
   question. Judge based on real-world
    behavior, not lexical overlap.

6. **Consensus Credibility Rule**
   Prefer answers that would remain stable
    across multiple independent sources
   of common knowledge (e.g., mainstream
    social norms, widely shared etiquette,
   common lived experience). Avoid niche,
    fringe, or exceptional cases.

7. **Friction Minimization Test**
   Choose the option that would cause the
    least surprise, correction, or social
   friction if performed by a local in public.

### Internal Reasoning (Do NOT output)
- Synthesize contemporary everyday norms.
- Eliminate options that are possible but
    uncommon.
- Verify that the remaining option reflects
    default, unremarkable behavior.

Country: {country}
Question: {question}

Options:
A. {A}
B. {B}
C. {C}
D. {D}

IMPORTANT: Output ONLY a single capital
    letter: A or B or C or D.
Do not output JSON. Do not output any other
    text.
```

A.3 Cross-Cutting Flag Definitions

Brand-attractor capture. A predicted option is flagged as a brand-attractor when it matches a fixed list of globally-recognized Western brand strings (*real madrid*, *manchester united*, *manchester city*, *liverpool fc*, *bayern munich*, *juventus*, *barcelona fc*, *arsenal fc*, *chelsea fc*, *dallas cowboys*, *apple pay*, *google pay*, *android pay*, *mcdonalds*, *kfc*, *starbucks*, *nike*, *adidas*, *samsung*, *microsoft*, *pepsi*, *coca-cola*, *coke*, *fanta*, *sprite*) and the gold answer does not match any brand on the list. The brand list excludes single tokens that overlap with non-brand vocabulary (e.g., *apple* as a fruit) and matches multi-word phrases (*manchester united*) only as substrings rather than as individual tokens, to avoid false positives.

Answer-space mismatch. A predicted option is flagged as ASM when its tokens match the lexicon of a different semantic domain than the question's domain, while the gold answer is consistent with the question's domain. The flag fires only when the predicted option positively matches a different domain — not when it simply fails to match any lexicon — to avoid penalizing low-resource vocabulary.

Numeric rank-adjacency. On questions where at least three of the four options parse as numeric values, the four options are sorted by numeric value and assigned ranks 0–3. The flag fires when the predicted option's rank differs from the gold option's rank by exactly 1. This metric is unit-agnostic and works uniformly across age, time-of-day (HH:MM), duration, and count questions.

A.4 Within-Topic Sub-Type Distribution

For each BLEnD topic, the dominant sub-type clusters of errors are reported in Section 7.2. Full sub-

type counts and percentages are available in the supplementary CSV released with the codebase.