

Toward a Culture Map of Chatbots: Evaluating LLM Understanding of Cross-Cultural Workplace Norms

Thomas Dubedout, Carlo Alberto Bono, Francesco Pierri, Anna Bernasconi
Dipartimento di Elettronica, Informazione e Bioingegneria – Politecnico di Milano
thomas.dubedout@ensta.fr, carlo.bono@polimi.it,
francesco.pierri@polimi.it, anna.bernasconi@polimi.it

Abstract

As chatbots powered by Large Language Models (LLMs) are increasingly deployed across cultural contexts, evaluating whether they can reason about culturally conditioned norms becomes essential. We explore whether the eight workplace-culture dimensions of Erin Meyer’s *The Culture Map* can provide a practical foundation for such an evaluation. We operationalize these dimensions into 104 behavioral scenarios that test two complementary capabilities: classifying behaviors along a cultural dimension and judging the plausibility of nationality–behavior pairings. We evaluate nine open-weight LLMs spanning 7B to approximately 671B parameters. Nationality information helps some models classify behavior, but most still struggle with group-conditioned plausibility. Explicit cultural rules benefit the reasoning model most, possibly suggesting that this capability depends more on reasoning than on scale. These findings demonstrate both the diagnostic potential of the proposed framework and the limitations that a broader evaluation should capture. Our study is therefore a first step toward a large-scale, open, and multidimensional “Culture Map of Chatbots” benchmark for LLM cross-cultural understanding. Code is available at <https://github.com/TH-77-oss/CMC-Benchmark>.

1 Introduction

Chatbots powered by large language models (LLMs) are increasingly deployed in international settings, where they interact with users from diverse cultural backgrounds (Liu et al., 2025a). This makes it essential to evaluate whether their outputs reflect the plurality of norms and perspectives that such interactions involve. Yet when annotators disagree, a common practice is to aggregate their judgments into a single gold label, typically through majority voting (Aroyo and Welty, 2015; Uma et al., 2021), while treating minority judgments as noise. This approach overlooks settings

in which disagreement does not arise from annotation error, but from systematic differences across social or cultural groups. In such cases, multiple answers may be legitimate, each reflecting the norms or expectations of a particular group (Plank, 2022; Basile et al., 2021; Mostafazadeh Davani et al., 2022).

This issue is especially consequential when users rely on chatbots for guidance in intercultural interactions. Cross-cultural friction at work is a well documented issue, with multicultural teams reporting, among the causes of conflict, direct versus indirect feedback, and differing attitudes towards hierarchy or decision-making norms (Brett et al., 2006), with a meta-analysis of 108 studies linking cultural diversity and task conflict (Stahl et al., 2010). As chatbots enter everyday work tasks, with writing and practical guidance among the most common work-related uses (Chatterji et al., 2025), they are increasingly asked for guidance on these interactions. For instance, a user preparing critical feedback for a colleague from another country may ask whether to communicate the criticism directly or soften it, whether to raise the issue in a group meeting or in private, and how explicitly to state the desired changes. The appropriateness of such advice may depend on culturally conditioned expectations rather than on a single universal standard. An AI assistant that presents the norms of one cultural context as universally appropriate may therefore provide ineffective or alienating guidance.

Existing work has established the value of open, large-scale evaluation (Chiang et al., 2024; Pierri et al., 2025), modeled how situational context shapes social norms (Ziems et al., 2023), and developed resources and tools for culturally aware language technologies (Nguyen et al., 2024; Pandey et al., 2025; Saha et al., 2025; Liu et al., 2025b, 2026; Kumar et al., 2026). However, recent surveys note that cultural competence remains an underspecified construct in NLP (Liu et al., 2025a;

Zhou et al., 2025). We argue this gap extends to evaluation itself: existing efforts rarely test how models reason about concrete behaviors under explicitly defined, group-conditioned cultural norms, and instead tend to reduce culturally structured variation to a single aggregated target.

To investigate this capability in a concrete setting, we draw on *The Culture Map* (Meyer, 2016), which characterizes cross-national differences in workplace interactions along eight dimensions: communicating, evaluating, persuading, leading, deciding, trusting, disagreeing, and scheduling. These dimensions provide a practical basis for benchmark construction by organizing the broad notion of culture into interpretable and behaviorally distinct axes that can be translated into testable scenarios and group-conditioned expectations. We operationalize the dimensions as binary behavioral scenarios and use nationality as group-level context to evaluate behavior classification and nationality-behavior plausibility. The nationality mappings represent stylized country-level tendencies from the underlying framework, rather than deterministic claims about individual behavior. We explore this approach as a first step toward building a large-scale, open, and multidimensional benchmark for LLM cross-cultural understanding, which we refer to as a “Culture Map of Chatbots.”

We organize the investigation around two research questions:

- **RQ1: Behavior Classification.** Can LLMs correctly classify a described behavior according to a given cultural dimension?
- **RQ2: Behavior Plausibility.** Can LLMs judge whether a described behavior is culturally plausible for individuals of a stated nationality?

To address these questions, we construct 104 culturally grounded scenarios spanning the eight dimensions of *The Culture Map*, each describing the behavior of two individuals, Person A and Person B. We derive two paired experimental conditions for each task.

For **RQ1**, models first **(a)** classify each person’s behavior using the scenario and the relevant dimension definition, and then **(b)** repeat the task with nationality provided as additional context. This comparison tests whether group information improves behavior classification beyond the behavioral description itself.

For **RQ2**, models judge the plausibility of each nationality-behavior pairing, first **(a)** without and

then **(b)** with a formal definition of the relevant cultural dimension. This comparison tests whether an explicit description of the cultural rule helps models make group-conditioned plausibility judgments.

Overall, our work contributes a preliminary benchmark dataset covering Meyer’s eight cultural dimensions, an evaluation framework separating behavior classification from nationality-conditioned plausibility judgment, and an empirical assessment of nine open-weight LLMs ranging from 7B to approximately 671B parameters, verified through a single human check.

2 Related Work

Large-scale evaluation platforms such as Chatbot Arena (Chiang et al., 2024) have demonstrated the value of open evaluation for comparing conversational AI systems in realistic settings. Other resources focus more directly on social and cultural knowledge. NormBank (Ziems et al., 2023) represents how situational context shapes social norms, while Nguyen et al. (2024) construct a cultural commonsense knowledge base to support intercultural dialogue. These resources provide broad coverage of culturally relevant situations, but do not offer a dimension-specific evaluation of concrete interpersonal behaviors such as giving feedback, expressing disagreement, or building trust. A related line of work examines whether language models can recognize culturally specific communication gaps. Saha et al. (2025) show that LLMs struggle to identify such gaps in user-generated text, while Pandey et al. (2025) develop a reading assistant that helps users interpret culturally specific expressions and references. These studies primarily address culture-specific *content*, whereas our work focuses on culture-conditioned *behavior*. More broadly, Guo et al. (2025) investigate whether LLMs capture cross-cultural differences, and Guan (2025) and Shen et al. (2026) study culturally aware language understanding and dialogue disambiguation, respectively. Recent surveys (Liu et al., 2025a; Zhou et al., 2025) emphasize that culture remains an underspecified construct in NLP evaluation and call for more explicit and systematically grounded operationalizations. We respond to this challenge by adopting the eight-dimension framework of *The Culture Map* (Meyer, 2016). Its dimensions translate broad differences in workplace interaction into interpretable behavioral con-

trasts, such as *direct* versus *indirect feedback*, providing a practical basis for constructing testable evaluation tasks. Rao et al. (2025) introduce NormAd, which evaluates the ability of LLMs to judge the social acceptability of situational descriptions, similarly to our RQ2. Unlike NormAd, we pair this judgment with a behavior classification (RQ1) and organize norms as axes of workplace interaction. Our approach also relates to perspectivist NLP, which treats disagreement as a meaningful signal rather than automatically collapsing it into a single aggregated label (Plank, 2022; Basile et al., 2021; Mostafazadeh Davani et al., 2022). We extend this perspective to group-conditioned cultural expectations by using nationality as a group-level variable and separating two capabilities: classifying behavior along a cultural dimension and judging the plausibility of a nationality–behavior pairing.

3 Data and Methods

Figure 1 gives an overview of our approach. Starting from eight cultural dimensions (Meyer, 2016) (Section 3.1), we construct a dataset of behavioral scenarios (Section 3.2), and validate this dataset with a human check. We then map nationalities onto the scale of each dimension (Section 3.3). In the experiments (Section 4), we will describe the four task variants, one for each research question, used for LLM evaluation.

3.1 Cultural Dimensions

We consider the eight cultural dimensions described in *The Culture Map* (Meyer, 2016). Each defines a qualitative scale along which the author positions national cultures. From the *low* to the *high* pole, the dimensions are:

Evaluating. Indirect vs. direct negative feedback.
Persuading. Applications-first vs. principles-first.
Communicating. Low-context vs. high-context.
Leading. Egalitarian vs. hierarchical.
Deciding. Consensual vs. top-down.
Trusting. Relationship-based vs. task-based.
Disagreeing. Conflict-avoidant vs. confrontational.
Scheduling. Flexible vs. linear time.

3.2 Dataset Construction

Scenarios. We build the dataset from two components of *The Culture Map* (Meyer, 2016). First, we use the book’s narrative examples to construct behavioral scenarios spanning its eight cultural dimensions. Second, we use its country-level scales

to define the nationality subsets introduced in Section 3.3. Each scenario describes a short interpersonal situation involving two fictional individuals, Person A and Person B, whose behaviors instantiate poles of the dimension being tested. We select examples that cover varied contexts and behavioral manifestations within each dimension.

We constructed scenarios according to three criteria intended to ensure that models rely on the relevant cultural behavior rather than on superficial cues:

1. Scenarios should be structurally distinct, rather than repeating the same interaction with only minor changes in setting or wording.
2. The main behavioral cue should be specific to the dimension being tested and should not primarily reflect another dimension.
3. The *high* and *low* poles should be balanced between Person A and Person B, preventing models from exploiting label position.

Scenarios meeting all three criteria are retained unchanged as *book-adapted*; those requiring changes are rewritten while preserving the intended dimension and labels, and are referred to as *synthetic*. Most scenarios (88%) are retained directly from the book, with the remaining 12% authored as synthetic replacements. All scenarios are subsequently reviewed for coherence and label consistency, yielding a final dataset of 104 scenarios across the eight dimensions.

Task formalization. Each scenario describes two fictional individuals, Person A and Person B, whose behaviors represent opposite poles of the cultural dimension being tested. For example, a scenario from the Evaluating dimension may read:

During feedback on a startup pitch, Person A tells directly “The market analysis in this pitch is wrong; you need to redo it,” without qualifying the statement. Reviewing another pitch, Person B says “There might be a few areas worth reconsidering, perhaps around the market analysis,” presenting the same concern as a suggestion.

For each person, the base dataset stores a *high* or *low* label indicating which pole of the dimension their behavior represents. In the example above, Person A is labeled *high*, corresponding to direct negative feedback, while Person B is labeled *low*, corresponding to indirect negative feedback. The

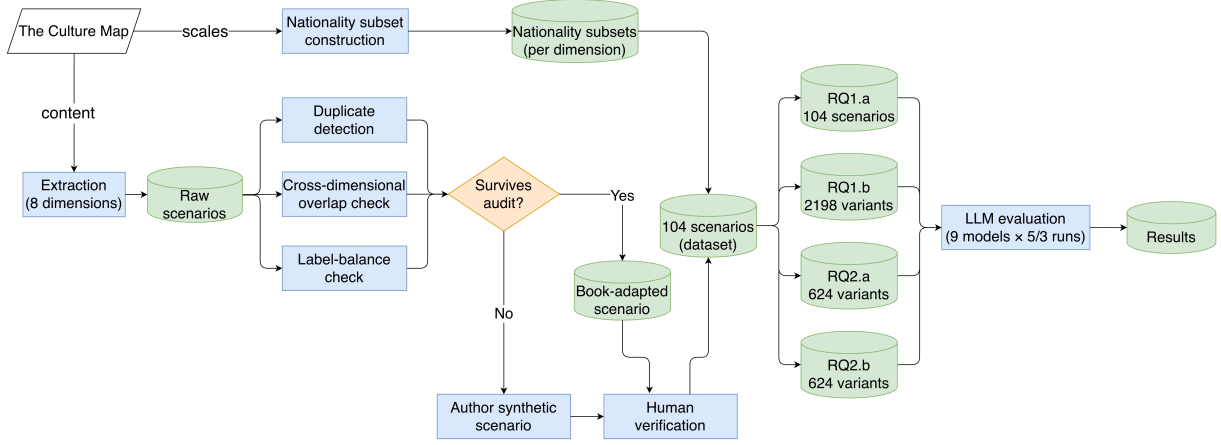


Figure 1: Methodological workflow.

complete definitions of the two poles for all dimensions are provided in the repository.

In the two **RQ1** conditions, the *high/low* label is the prediction target. We describe scenarios where it holds that one of Person A/B is *high*, and the other one is *low*. Models classify each person’s behavior first without nationality information and then with nationality provided as additional context. Because the label is determined by the behavior described in the scenario, adding nationality does not change the correct answer.

In the two **RQ2** conditions, the prediction target is instead *plausible* or *implausible*. This label is derived by comparing each person’s behavioral class with the subset containing their assigned nationality. A nationality–behavior pairing is labeled *plausible* when the nationality belongs to the subset associated with the person’s *high/low* behavior, and *implausible* when it belongs to the opposite subset. The behavior of both Person A/B can be *plausible* or *implausible*. The two **RQ2** conditions use the same scenarios, nationality assignments, and plausibility labels. They differ only in whether the prompt includes the formal definition of the cultural dimension.

3.3 Mapping Nationality to Labels

We derive nationality subsets from the visual country-level scales presented in *The Culture Map* (Meyer, 2016). For each dimension, countries visually positioned clearly toward one pole are assigned to either the *high* or *low* subset. Membership indicates only the side of the scale on which a country is placed; it does not imply an ordinal ranking among countries within the same subset. Countries with ambiguous positions are excluded, resulting in

different subset sizes across dimensions. The complete subsets are reported in the repository. These mappings are used differently across the two tasks. For behavior classification, nationalities consistent with each person’s behavioral label are provided as additional context. For plausibility judgment, a nationality–behavior pairing is labeled *plausible* when the nationality belongs to the subset associated with the described behavior and *implausible* when it belongs to the opposite subset. These labels represent the group-level expectations encoded by the source framework rather than deterministic claims about individual behavior.

4 Experimental Setup

Each research question is evaluated with its own experiment and corresponding dataset. Datasets are built starting from the 104 base scenarios (Section 3.2) by adding task-specific information: a nationality assignment for Person A and/or Person B, drawn from the pools defined in Section 3.3, and, for some tasks, the dimension’s formal definition.

RQ1.a: Behavior Classification (w/o nationality).

The 104 base scenarios themselves, with the dimension’s formal definition provided but no nationality information - isolating classification ability from group-conditioned reasoning. 88% of scenarios are adapted from Meyer (2016); all are human-verified.

RQ1.b: Behavior Classification (w/ nationality).

For each base scenario, every nationality in the pool matching Person A’s true class is paired with every nationality in the pool matching Person B’s true class, so assigned nationalities are always consistent with the described behavior. Taking all such pairwise combinations across the eight dimensions

yields 2,198 variants. This dataset requires no separate verification, as it is built directly from the already-verified RQ1.a scenarios.

RQ2.a: Plausibility Judgment (w/o dimension definitions). The target changes to *plausible/improbable*: each person’s nationality is sampled from either pool, matching (plausible) or contradicting (improbable) their true class. Since either-pool sampling makes exhaustive enumeration impractical, we instead sample 6 independent nationality draws per scenario, chosen so that, in expectation, each yields a balanced 3 plausible/3 improbable split per person.¹ This yields 624 variants (104 base scenarios times 6 draws each).

RQ2.b: Plausibility Judgment (w/ dimension definitions). The same 624 RQ2.a variants, with the dimension’s formal definition (the RQ1.b mapping) added to the prompt - focusing on whether an explicit cultural rule improves plausibility judgment.

4.1 Models

We select nine open-weight models spanning a wide range of parameter counts, from 7B to ~671B: *mistral-7b* (Jiang et al., 2023), *qwen2.5:7b* (Yang et al., 2024), *apertus-v1.1:8b* (Project Apertus, 2025), *llama3.1:8b* (Grattafiori et al., 2024), *gemma2:9b* (Gemma Team et al., 2024), *gemma3:12b* (Gemma Team et al., 2025), *gpt-oss:20b* (Agarwal et al., 2025), *qwen2.5:72b* (Yang et al., 2024), and *deepseek-V3:671b* (DeepSeek-AI, 2024).

We direct the evaluation at open-weight models for reproducibility and cost reasons; extending the benchmark to closed-weight alternatives is left to future work. We do not override sampling parameters (temperature, top-*p*) for any model, relying instead on each serving engine’s default configuration. Most models are run locally via Ollama, which enables full reproducibility without reliance on proprietary APIs; the two largest models, *qwen2.5:72b* and *deepseek-V3:671b*, exceed our local hardware capacity and are instead accessed through Hugging Face Inference Providers. Note that one model, *gpt-oss:20b*, has native chain-of-thought reasoning capability, distinguishing it architecturally from the other eight.

¹Expectation, not guarantee: each scenario has a small chance ($\approx 3\%$) of degenerating to a uniform outcome for a given person, predicting roughly 6 such cases dataset-wide.

4.2 Prompting Strategy

All prompts require the model to return a single JSON object under strict formatting constraints – no free-text explanation, values restricted to the exact label set for the task – which removes the need for free-text parsing. Prompts differ across experimental settings in two respects:

- (i) whether nationality is included as input
- (ii) whether the formal *high/low* definition of the dimension; being tested (the *definition*) is included.

RQ1.a provides the definition only; RQ1.b adds nationality to test whether models correctly apply a stated cultural rule. RQ2.a omits the definition, so models must rely on their own cultural knowledge rather than an explanation we supply; RQ2.b adds it to separate a lack of knowledge from a failure to apply it without explicit guidance. Full prompt templates for every condition are provided in the Appendix.

4.3 Evaluation Metrics

- **RQ1.** We report accuracy against the ground-truth label (*high/low*).²
- **RQ2.** We report accuracy against the ground-truth label (*plausible/improbable*), using the same aggregation as RQ1. We additionally report *flip rate*: the rate at which a model’s prediction changes when the assigned nationality pool changes, which in our experiments measures the model’s sensitivity to group-conditioned plausibility, independent of whether the model was already correct or not, without having access to group information.

5 Results

RQ1: Behavior Classification. Figure 2 and Table 1 report per-dimension classification accuracy for RQ1.a and RQ1.b.

RQ1.a (without nationality). Most models reach high global accuracy from the behavior description alone: *deepseek-V3:671b* (98.2%), *qwen2.5:72b* (97.8%), and *gpt-oss:20b* (97.0%) are near ceiling, followed by *qwen2.5:7b* (93.3%) and *gemma3:12b* (86.1%). Four models perform markedly lower: *gemma2:9b* (75.4%), *llama3.1:8b* (63.7%), *mis-*

²Mean and standard deviation across dimensions and runs jointly: the mean is computed by first averaging within each (dimension, run) pair, and the standard deviation is computed on the pooled set of these per-dimension, per-run values, rather than on run-level averages alone.

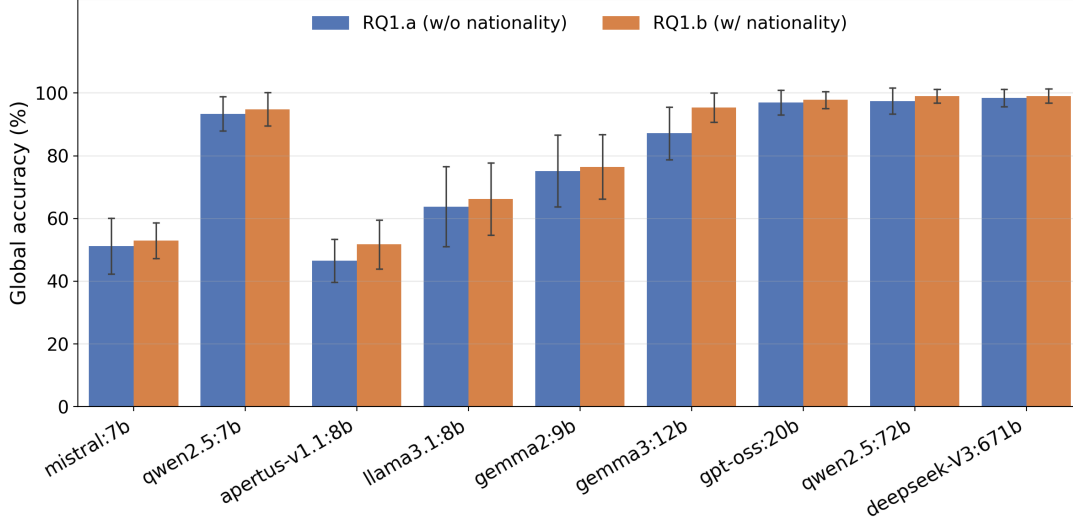


Figure 2: Global accuracy by model (ordered by parameter size), RQ1.a vs RQ1.b (mean $\pm$ std, pooled across dimensions and runs).

Model	Comm.	Dec.	Disagr.	Eval.	Lead.	Pers.	Sched.	Trust.	Global
<i>RQ1.a</i>									
mistral:7b	53.3 $\pm$ 0.0	62.7 $\pm$ 2.8	48.9 $\pm$ 6.1	46.7 $\pm$ 0.0	53.3 $\pm$ 0.0	34.7 $\pm$ 8.7	50.0 $\pm$ 0.0	60.0 $\pm$ 2.3	51.2 $\pm$ 1.1
qwen2.5:7b	84.7 $\pm$ 1.8	93.3 $\pm$ 0.0	88.9 $\pm$ 0.0	89.3 $\pm$ 1.5	100.0 $\pm$ 0.0	94.7 $\pm$ 3.0	96.2 $\pm$ 3.4	100.0 $\pm$ 0.0	93.3 $\pm$ 0.5
apertus-v1.1:8b	42.7 $\pm$ 4.3	46.7 $\pm$ 4.7	46.7 $\pm$ 7.5	42.0 $\pm$ 9.9	56.0 $\pm$ 3.7	46.7 $\pm$ 0.0	50.0 $\pm$ 6.2	41.7 $\pm$ 4.2	46.4 $\pm$ 1.8
llama3.1:8b	68.7 $\pm$ 4.5	49.3 $\pm$ 3.7	62.2 $\pm$ 7.2	64.7 $\pm$ 4.5	67.3 $\pm$ 6.8	54.7 $\pm$ 5.6	55.0 $\pm$ 6.8	88.3 $\pm$ 9.5	63.7 $\pm$ 2.8
gemma2:9b	69.3 $\pm$ 3.7	81.3 $\pm$ 3.0	77.8 $\pm$ 0.0	88.0 $\pm$ 5.6	72.0 $\pm$ 3.0	58.7 $\pm$ 3.0	62.5 $\pm$ 0.0	91.7 $\pm$ 0.0	75.4 $\pm$ 1.1
gemma3:12b	86.7 $\pm$ 2.4	93.3 $\pm$ 0.0	88.9 $\pm$ 0.0	86.7 $\pm$ 0.0	74.0 $\pm$ 4.3	76.0 $\pm$ 3.7	100.0 $\pm$ 0.0	91.7 $\pm$ 0.0	86.1 $\pm$ 0.7
gpt-oss:20b	94.0 $\pm$ 4.3	93.3 $\pm$ 0.0	92.2 $\pm$ 5.0	98.0 $\pm$ 1.8	98.7 $\pm$ 3.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	97.0 $\pm$ 1.3
qwen2.5:72b	97.8 $\pm$ 3.8	93.3 $\pm$ 0.0	88.9 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	97.8 $\pm$ 0.6
deepseek-V3:671b	94.4 $\pm$ 1.9	93.3 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	98.2 $\pm$ 0.3
<i>RQ1.b</i>									
mistral:7b	53.3 $\pm$ 0.0	50.2 $\pm$ 0.5	56.7 $\pm$ 0.6	46.7 $\pm$ 0.0	55.3 $\pm$ 0.7	46.7 $\pm$ 0.0	50.0 $\pm$ 0.0	64.7 $\pm$ 1.0	53.3 $\pm$ 0.2
qwen2.5:7b	98.2 $\pm$ 0.4	93.3 $\pm$ 0.0	83.2 $\pm$ 0.7	92.1 $\pm$ 1.0	97.5 $\pm$ 0.1	94.4 $\pm$ 0.6	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	95.1 $\pm$ 0.2
apertus-v1.1:8b	40.5 $\pm$ 1.7	47.0 $\pm$ 0.6	67.9 $\pm$ 3.2	53.3 $\pm$ 1.6	56.1 $\pm$ 0.3	46.7 $\pm$ 0.0	53.2 $\pm$ 1.7	49.1 $\pm$ 0.8	51.3 $\pm$ 0.4
llama3.1:8b	67.0 $\pm$ 0.8	49.4 $\pm$ 0.9	64.4 $\pm$ 2.2	61.6 $\pm$ 1.5	69.8 $\pm$ 1.0	59.8 $\pm$ 1.3	65.5 $\pm$ 1.9	92.0 $\pm$ 1.1	67.4 $\pm$ 0.4
gemma2:9b	76.0 $\pm$ 0.4	78.9 $\pm$ 0.4	82.0 $\pm$ 0.8	84.4 $\pm$ 0.8	66.5 $\pm$ 0.6	66.2 $\pm$ 1.0	63.1 $\pm$ 0.3	94.7 $\pm$ 0.9	77.4 $\pm$ 0.3
gemma3:12b	98.3 $\pm$ 0.3	93.3 $\pm$ 0.0	100.0 $\pm$ 0.0	94.8 $\pm$ 0.6	99.2 $\pm$ 0.2	86.1 $\pm$ 0.4	100.0 $\pm$ 0.0	91.7 $\pm$ 0.0	95.5 $\pm$ 0.2
gpt-oss:20b	95.6 $\pm$ 0.8	93.1 $\pm$ 0.3	95.2 $\pm$ 1.0	98.5 $\pm$ 0.2	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	97.9 $\pm$ 0.1
qwen2.5:72b	100.0 $\pm$ 0.0	93.3 $\pm$ 0.0	99.4 $\pm$ 0.5	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	99.6 $\pm$ 0.7	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	99.3 $\pm$ 0.1
deepseek-V3:671b	99.6 $\pm$ 0.3	93.1 $\pm$ 0.3	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	100.0 $\pm$ 0.0	99.4 $\pm$ 0.0

Table 1: Per-dimension accuracy (mean $\pm$ std) for RQ1.a and RQ1b. Dimension names abbreviated: Comm. = Communicating, Dec. = Deciding, Disagr. = Disagreeing, Eval. = Evaluating, Lead. = Leading, Pers. = Persuading, Sched. = Scheduling, Trust. = Trusting. Corresponding line plots are shown in Figure 4 (see Appendix).

tral:7b (51.2%), and *apertus-v1.1:8b* (46.4%), with the latter two close to chance.

RQ1.b (with nationality). Adding nationality context leaves near-ceiling models largely unchanged – *deepseek-V3:671b* from 98.2% to 99.4% (+1.2%), *qwen2.5:72b* from 97.8% to 99.3% (+1.5%), *gpt-oss:20b*: from 97.0% to 97.9% (+0.9%) – but produces a clear gain for *gemma3:12b* from 86.1% to 95.5% (+9.4%) and smaller gains for *gemma2:9b* from 75.4% to 77.4% (+2.0%) and *llama3.1:8b* from 63.7% to 67.4% (+3.7%), while *apertus-v1.1:8b* and *mistral:7b* remain near-chance (51.3% and 53.3%, respectively).

Nationality context helps models that are

not already at ceiling on RQ1.a, most clearly *gemma3:12b*, but does not lift the two weakest models above chance. This suggests their limitation on RQ1 is not the absence of nationality information, but a more basic difficulty with the classification task itself.

RQ2: Behavior Plausibility. Figure 3 reports global accuracy and flip rate for RQ2.a and RQ2.b. Table 2 reports per-dimension flip rate, and Table 3 (see Appendix) reports per-dimension accuracy.

RQ2.a (without dimension definition). Most models obtain accuracy near 50%, failing to capture intercultural signal when classifying a behavior

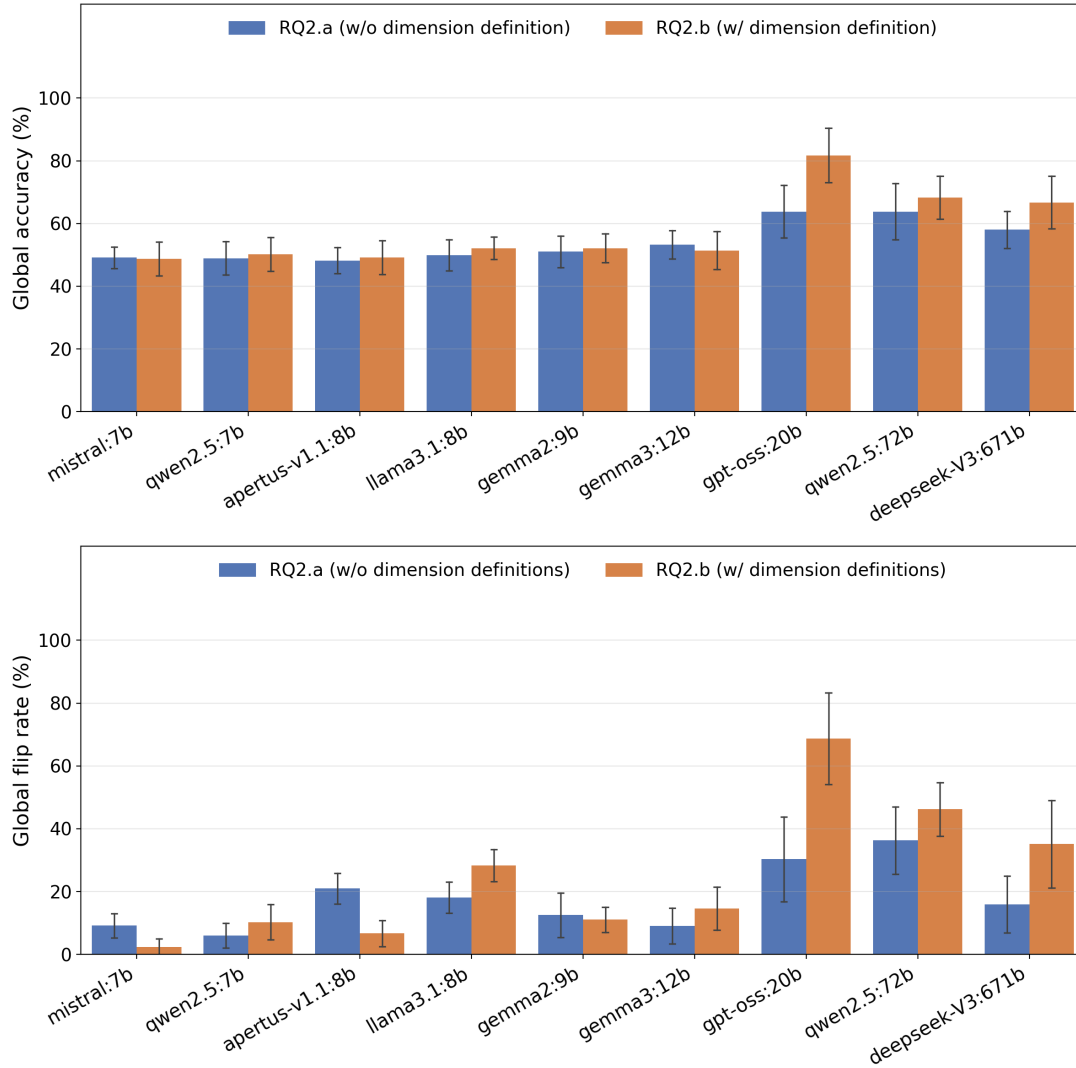


Figure 3: Global accuracy (above) and flip rate (below) by model, indicating RQ2.a vs RQ2.b (mean $\pm$ std, pooled across dimensions and runs). Per-dimension breakdown for global-accuracy are shown with line plots in Figure 5 (see Appendix); for flip rate, refer to Table 2.

as plausible or not. Two models stand out when measuring global accuracy: *qwen2.5:72b* (62.9%) and *gpt-oss:20b* (63.3%), followed by *deepseek-V3:671b* (57.4%). Flip rate also highlights this split: models with near-chance accuracy report flip rate close to 0% (*qwen2.5:7b*: 6.1%, *mistral:7b*: 8.9%, *gemma3:12b*: 8.9%), indicating a fixed response bias rather than noisy but input-sensitive guessing, while *qwen2.5:72b* (35.5%), *gpt-oss:20b* (30.0%), and *apertus-v1.1:8b* (20.7%) show some sensitivity to nationality even without explicit guidance.

RQ2.b (with dimension definition). Models with a larger number of parameters are those that show the greatest gain: *gpt-oss:20b* improves the most, reaching 81.8% global and up to 97.7% for Leading (see Table 3, in Appendix), while *qwen2.5:72b* rises to 67.8% global accuracy and

deepseek-V3:671b to 65.9%. Flip rate increases accordingly, most sharply for *gpt-oss:20b* (30.0% to 69.0%), followed by *qwen2.5:72b* (35.5% to 46.0%) and *deepseek-V3:671b* (15.1% to 35.0%), while remaining low or even decreasing for smaller models (e.g., *apertus-v1.1:8b*: 20.7% to 6.8%).

Providing the explicit cultural rule increases not only correctness but sensitivity to nationality, often with positive impact, and does so unevenly across models: the gain is concentrated in *gpt-oss:20b*, while larger models improve only modestly, and smaller models barely move. This contrast between RQ1.b, where nationality context reliably helps, and RQ2, where the same information helps only some models, is discussed next.

Model	Comm.	Dec.	Disagr.	Eval.	Lead.	Pers.	Sched.	Trust.	Global
<i>RQ2.a</i>									
mistral:7b	7.2±1.2	8.4±1.0	11.0±3.2	9.4±0.5	12.9±2.3	1.6±0.5	9.1±1.4	13.4±1.5	8.9±0.9
qwen2.5:7b	10.9±2.1	4.4±1.2	1.6±2.4	6.6±2.3	0.0±0.0	10.6±1.1	7.1±1.8	6.2±0.6	6.1±0.8
apertus-v1.1:8b	18.3±3.1	20.1±3.0	19.3±5.1	25.7±2.5	14.6±4.8	21.7±2.7	26.3±1.4	21.9±4.5	20.7±0.6
llama3.1:8b	15.8±1.6	10.5±2.9	20.1±5.4	20.4±3.9	15.3±2.4	19.7±2.7	20.2±4.0	22.6±4.5	17.8±1.2
gemma2:9b	16.8±1.4	12.3±2.8	6.6±2.2	4.2±1.4	14.0±3.5	26.0±2.3	5.9±2.2	13.8±1.9	13.2±1.0
gemma3:12b	2.3±1.0	8.2±3.5	9.1±1.3	15.0±2.4	16.1±3.2	0.0±0.0	9.9±2.3	11.2±1.9	8.9±1.3
gpt-oss:20b	24.4±1.3	23.8±2.7	38.4±4.3	32.6±4.0	54.9±4.3	6.1±4.4	29.2±2.8	32.8±4.2	30.0±0.9
qwen2.5:72b	42.7±5.1	24.2±3.7	42.5±5.6	35.2±6.0	41.5±0.9	16.8±4.2	39.6±5.4	47.8±1.5	35.5±1.8
deepseek-V3:671b	16.0±2.3	11.2±2.4	21.0±2.7	7.9±1.1	23.9±3.6	0.0±0.0	19.1±1.8	28.4±3.4	15.1±0.5
<i>RQ2.b</i>									
mistral:7b	0.2±0.4	2.3±0.9	4.0±3.9	0.3±0.7	1.0±0.7	7.2±1.4	1.7±2.4	1.4±0.0	2.2±0.4
qwen2.5:7b	11.3±3.2	2.5±1.7	6.1±1.9	9.5±1.4	9.8±3.0	12.5±3.3	21.0±4.2	9.4±1.9	9.9±0.9
apertus-v1.1:8b	7.0±3.1	7.8±2.0	3.3±2.3	13.0±1.9	3.0±2.5	3.0±1.5	6.8±4.5	9.2±2.6	6.8±0.4
llama3.1:8b	27.8±3.4	27.6±1.5	24.4±4.8	29.8±3.1	24.2±3.3	30.8±2.1	36.4±5.4	25.5±4.7	28.1±1.5
gemma2:9b	14.6±2.4	9.6±1.6	7.4±2.4	5.9±1.7	10.3±0.9	13.8±4.5	15.7±3.2	10.9±1.7	11.0±1.3
gemma3:12b	3.5±1.7	13.5±1.0	16.2±1.8	11.9±1.2	29.0±1.3	14.8±1.1	16.7±0.5	11.1±0.9	14.5±0.5
gpt-oss:20b	68.9±2.7	69.1±2.8	63.9±6.7	64.7±2.5	95.2±2.0	42.7±5.1	65.2±5.0	80.0±4.9	69.0±1.0
qwen2.5:72b	48.2±2.8	37.7±2.0	58.4±2.4	40.9±1.7	53.9±3.4	41.3±7.6	36.0±4.0	52.9±3.3	46.0±0.5
deepseek-V3:671b	29.6±2.3	32.0±4.9	45.8±6.0	31.0±4.0	62.3±0.3	14.4±1.4	27.6±2.1	38.1±5.4	35.0±1.3

Table 2: Per-dimension flip rate (mean $\pm$ std), across all nine models, for RQ2.a and RQ2.b. Dimension names abbreviated: Comm. = Communicating, Dec. = Deciding, Disagr. = Disagreeing, Eval. = Evaluating, Lead. = Leading, Pers. = Persuading, Sched. = Scheduling, Trust. = Trusting. A line plot visualization is available in Figure 6, see Appendix.

6 Discussion and Conclusion

In this paper, we explored the sensitivity of large language models to group-conditioned cultural norms, using nationality as the relevant grouping variable. In doing so, we introduced the first step toward the “Culture Map of Chatbots”, an open benchmark operationalizing *The Culture Map* (Meyer, 2016) eight cultural dimensions into 104 human-verified base scenarios and four task-specific datasets covering behavior classification (RQ1.a, RQ1.b) and plausibility judgment (RQ2.a, RQ2.b). We evaluated nine open-weight models spanning 7B to ~671B parameters, and we found that providing nationality context improves accuracy when classifying people’s behavior in a workplace context. We also find that, for most models, nationality context impacts the models’ judgment about the plausibility of scenario-driven behavior descriptions, although not always in a positive manner.

We report results that are relevant to pluralism-aware evaluation; we trust that our roadmap for a “Culture Map of Chatbots”, audit methodology, and evaluation metrics support future work extending this group-conditioned evaluation paradigm to other grouping variables and cultural frameworks.

Future work should also test whether the effects we observe generalize across reasoning models, extend group-conditioned evaluation to grouping variables beyond nationality, and further investigate whether the near-ceiling accuracy some models

reach on RQ1.a reflects genuine cultural reasoning or reliance on explicit textual markers, using scenarios with more implicit phrasing.

Another natural extension of our work is a predictive setup that evaluates Person B’s behavior from Person A’s behavior, given both nationalities.

Limitations

Some models reach near-ceiling accuracy on certain dimensions in RQ1.a (e.g., Trusting, Scheduling) despite our quality audit (Section 3.2). We cannot fully rule out that some surviving scenarios still rely on comparatively explicit behavioral markers rather than requiring genuine cultural reasoning. Testing with more implicit phrasing is left to future work.

Our reasoning-versus-scale finding rests on a single reasoning model (*gpt-oss:20b*). The remaining eight models are of comparable parameter scale (7B–72B, with the exception of *deepseek-V3:671b*) and share a broadly similar training paradigm (large-scale pretraining followed by instruction tuning). A consistent pattern across these eight is therefore weaker evidence than it would be across multiple reasoning models, and we cannot rule out that some of *gpt-oss:20b*’s behavior reflects factors specific to this model rather than reasoning capability in general.

We do not override sampling parameters for any model (Section 4.1), so comparisons across models are not fully controlled for decoding configuration.

Our benchmark operationalizes cultural dimensions as binary (*high/low*), following Meyer’s framework. This is itself a simplification: real cross-cultural variation is unlikely to reduce to two discrete poles, and a graded or multi-way representation may better capture group-conditioned norms in future extensions of this work.

Finally, our results rely on nationality as the only grouping variable, and on a single source (Meyer’s eight dimensions) for defining cultural norms. Other social groupings and other theoretical frameworks may yield different conclusions about models’ sensitivity to group-conditioned norms.

Ethical Considerations

This research focuses on evaluating the understanding of large language models on cross-cultural workplace norms. It relies on nationality as an observable proxy for cultural background, which risks essentializing culture as fixed and homogeneous. Our labels describe group-level tendencies drawn from the (Meyer, 2016) research, not prescriptive claims about any individual, and should not be used outside this evaluative purpose.

References

- Sandhini Agarwal, Lama Ahmad, Jason Ai, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.
- Lora Aroyo and Chris Welty. 2015. [Truth is a lie: Crowd truth and the seven myths of human annotation](#). *AI Magazine*, 36(1):15–24.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Antske Fokkens, Aditi Iyer, Sergiu Massanet, Rebecca Passonneau, Massimo Poesio, and Maarten Sap. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*.
- Jeanne M. Brett, Kristin Behfar, and Mary C. Kern. 2006. Managing multicultural teams. *Harvard Business Review*, 84(11):84–91.
- Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. [How people use ChatGPT](#). Working Paper 34255, National Bureau of Economic Research.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.
- DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Gemma Team, Morgane Rivière, Shreya Pathak, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Hao Guan. 2025. Context-aware semantic ambiguity resolution in cross-cultural dialogue understanding. *Journal of Sustainability, Policy, and Practice*, 1(2):136–147.
- Shanshan Guo, Siyuan Jiang, Qing He, Yude Xiao, Jinyu Liang, Bo Yude, Ming He, Shuo Tao, and Le Zhang. 2025. [Do large language models truly understand cross-cultural differences?](#) *arXiv preprint arXiv:2512.07075*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Tanay Kumar, Shreya Gautam, Aman Chadha, Vinija Jain, and Francesco Pierri. 2026. Personality shapes gender bias in persona-conditioned llm narratives across english and hindi: An empirical investigation. *arXiv preprint arXiv:2604.23600*.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. [Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art](#). *Transactions of the Association for Computational Linguistics*, 13:652–689.
- Geng Liu, Carlo Alberto Bono, and Francesco Pierri. 2025b. Comparing diversity, negativity, and stereotypes in chinese-language ai technologies: an investigation of baidu, ernie and qwen. *PeerJ Computer Science*, 11:e2694.
- Geng Liu, Li Feng, Carlo Alberto Bono, Songbo Yang, Mengxiao Zhu, and Francesco Pierri. 2026. Evaluating chinese large language models: The influence of persona assignment on stereotypes and safeguards. *ACM Transactions on Intelligent Systems and Technology*, 17(5):1–33.
- Erin Meyer. 2016. *The Culture Map (INTL ED): Decoding How People Think, Lead, and Get Things Done Across Cultures*. PublicAffairs.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.

- Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. [Cultural commonsense knowledge for intercultural dialogues](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1774–1784.
- Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. [CULTURALLY YOURS: A reading assistant for cross-cultural content](#). In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 208–216, Abu Dhabi, UAE. Association for Computational Linguistics.
- Francesco Pierri, Anna Bernasconi, Flavio Giobergia, Claudio Savelli, Elena Baralis, Luca Cagliero, and Stefano Ceri. 2025. A conceptual map for exploring the landscape of large language models. *IEEE Access*.
- Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Project Apertus. 2025. Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*.
- Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. [NormAd: A framework for measuring the cultural adaptability of large language models](#). 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 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.
- Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025. [Reading between the lines: Can LLMs identify cross-cultural communication gaps?](#) 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 8043–8067, Albuquerque, New Mexico. Association for Computational Linguistics.
- Lingfeng Shen, Xu Cai, Yue Long, Imran Razzak, Guanhua Chen, and Shoaib Jameel. 2026. [CALM: Culturally self-aware language models](#). *arXiv preprint arXiv:2601.03483*.
- Günter K. Stahl, Martha L. Maznevski, Andreas Voigt, and Karsten Jonsen. 2010. Unraveling the effects of cultural diversity in teams: A meta-analysis of research on multicultural work groups. *Journal of International Business Studies*, 41(4):690–709.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. [Learning from disagreement: A survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- An Yang, Baosong Yang, Beichen Zhang, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. [Culture is not trivia: Sociocultural theory for cultural NLP](#). *arXiv preprint arXiv:2502.12057*.
- Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. [NormBank: A knowledge bank of situational social norms](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

Appendix

Prompt Templates

RQ1.a

```
def prompt_RQ1_a(context, question, mapping):  
    return f"""  
    Task: Analyze the business scenario and  
    ↪ classify each person's behavior accordingly.  
  
    Context: {context}  
  
    Questions:  
        Person A: {question["person_a"]}  
        Person B: {question["person_b"]}  
  
    Options  
  
    High  
        Label: {mapping["high"]["label"]}  
        Definition:  
        {mapping["high"]["definitions"]}  
  
    Low  
        Label: {mapping["low"]["label"]}  
        Definition:  
        {mapping["low"]["definitions"]}  
  
    CRITICAL FORMAT INSTRUCTIONS:  
        - You MUST output only valid JSON.  
        - Do NOT include explanations.  
        - Do NOT include reasoning.  
        - Do NOT include additional text.  
        - The values MUST be exactly "high" or "  
    ↪ low" (lowercase).  
        - Any other value is invalid.  
  
    The output format is strictly:  
  
    {  
        "person_a": "high",  
        "person_b": "low"  
    }  
    """
```

RQ1.b

```
def prompt_RQ1_b(context, question, mapping,  
    ↪ nationality_a, nationality_b):  
    return f"""  
    Task: Analyze the business scenario and  
    ↪ classify each person behavior accordingly.  
  
    Context: {context}  
  
    Nationality :  
        Person A : {nationality_a}  
        Person B : {nationality_b}  
  
    Questions:  
        Person A: {question["person_a"]}  
        Person B: {question["person_b"]}  
  
    Options
```

```
    High  
        Label: {mapping["high"]["label"]}  
        Definition:  
        {mapping["high"]["definitions"]}  
  
    Low  
        Label: {mapping["low"]["label"]}  
        Definition:  
        {mapping["low"]["definitions"]}  
  
    CRITICAL FORMAT INSTRUCTIONS:  
        - You MUST output only valid JSON.  
        - Do NOT include explanations.  
        - Do NOT include reasoning.  
        - Do NOT include additional text.  
        - The values MUST be exactly "high" or "  
    ↪ low" (lowercase).  
        - Any other value is invalid.  
  
    The output format is strictly:  
  
    {{  
        "person_a": "high",  
        "person_b": "low"  
    }}  
    """
```

RQ2.a

```
def prompt_RQ2_a(context, question,  
    ↪ nationality_a, nationality_b):  
    return f"""  
    Task: Judge whether each person's behavior  
    ↪ in this cross-cultural business scenario is  
        culturally plausible given their nationality.  
    ↪  
  
    Context: {context}  
  
    Nationality :  
        Person A : {nationality_a}  
        Person B : {nationality_b}  
  
    Questions:  
        Person A: {question["person_a"]}  
        Person B: {question["person_b"]}  
  
    CRITICAL FORMAT INSTRUCTIONS:  
        - You MUST output only valid JSON.  
        - Do NOT include explanations.  
        - Do NOT include reasoning.  
        - Do NOT include additional text.  
        - The values MUST be exactly "plausible"  
    ↪ or "implausible" (lowercase).  
        - Any other value is invalid.  
  
    The output format is strictly:  
  
    {{  
        "person_a": "plausible",  
        "person_b": "implausible"  
    }}  
    """
```

RQ2.b

```
def prompt_RQ2_b(context, question, mapping,
↳ nationality_a, nationality_b):

    return f"""
    Task: Judge whether each person's described
↳ behavior is culturally plausible given their
↳ nationality.

    Context: {context}

    Nationality:
        Person A: {nationality_a}
        Person B: {nationality_b}

    Cultural dimension being tested:

        High ({mapping["high"]["label"]}):
        {mapping["high"]["definitions"]}

        Low ({mapping["low"]["label"]}):
        {mapping["low"]["definitions"]}

    A person's behavior is "plausible" if it
↳ matches what is typically expected from
↳ someone of
    their nationality on this cultural dimension,
↳ and "implausible" if it contradicts it.

    Questions:
        Person A: {question["person_a"]}
        Person B: {question["person_b"]}

    CRITICAL FORMAT INSTRUCTIONS:
        - You MUST output only valid JSON.
        - Do NOT include explanations.
        - Do NOT include reasoning.
        - Do NOT include additional text.
        - The values MUST be exactly "plausible"
↳ or "implausible" (lowercase).
        - Any other value is invalid.

    The output format is strictly:

    {{
    "person_a": "plausible",
    "person_b": "implausible"
    }}
    """
```

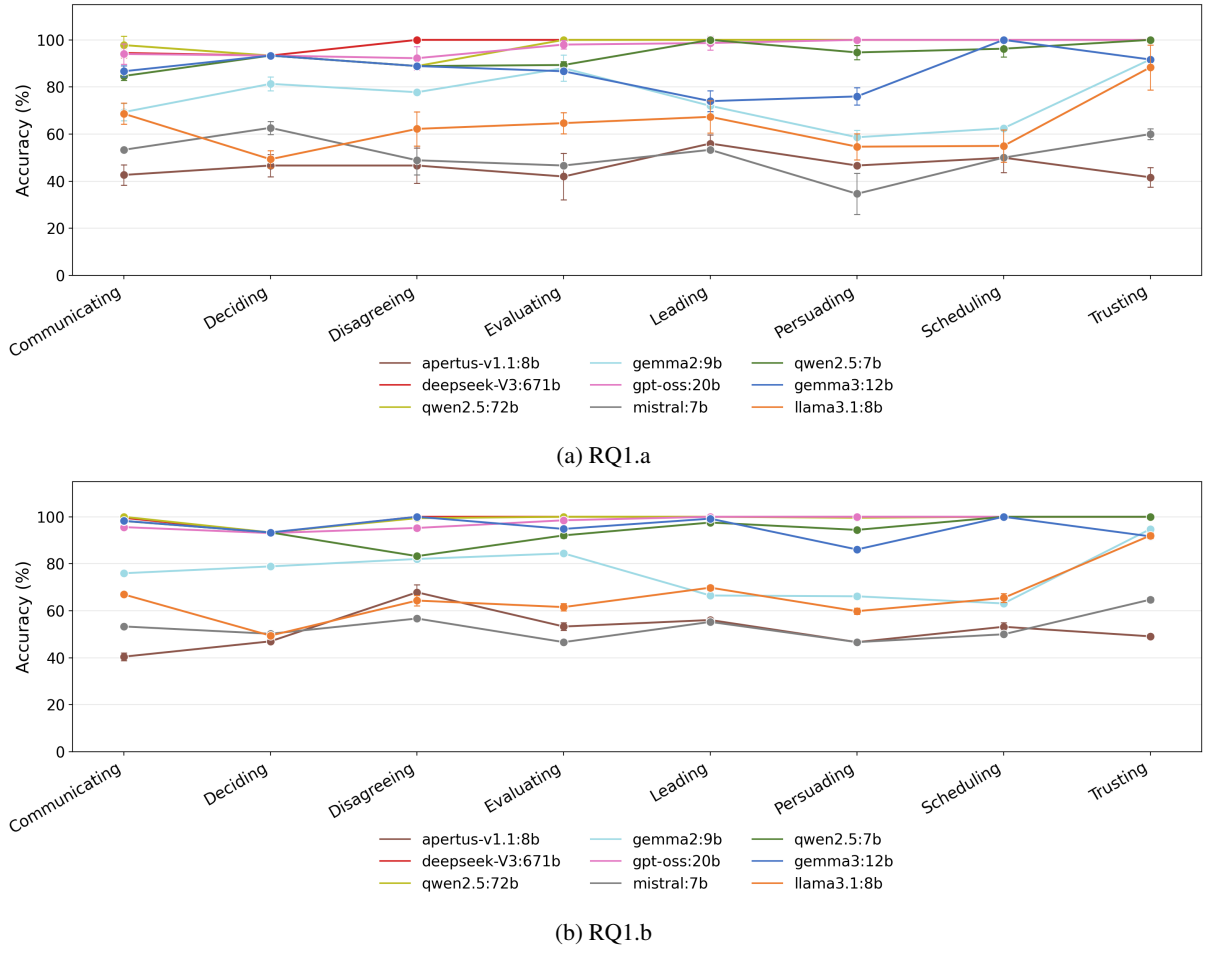
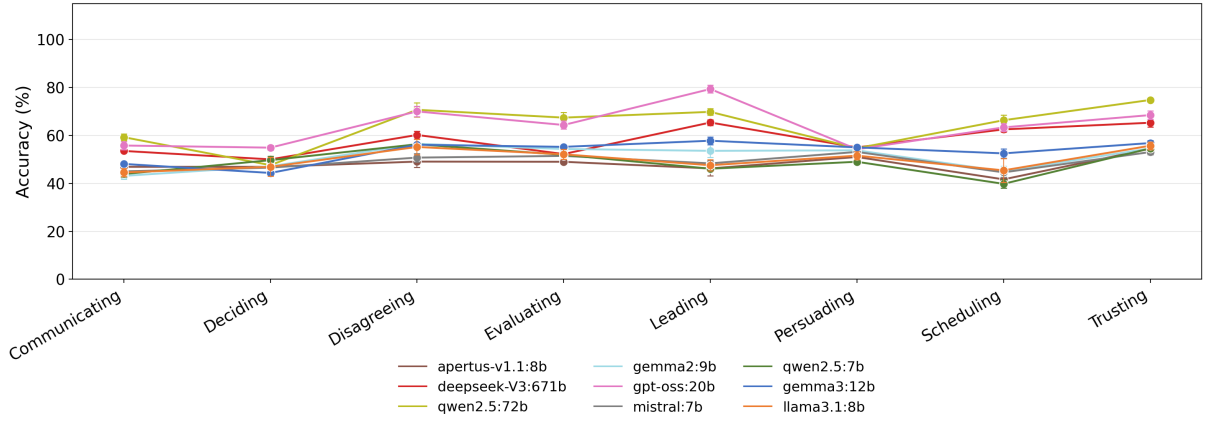


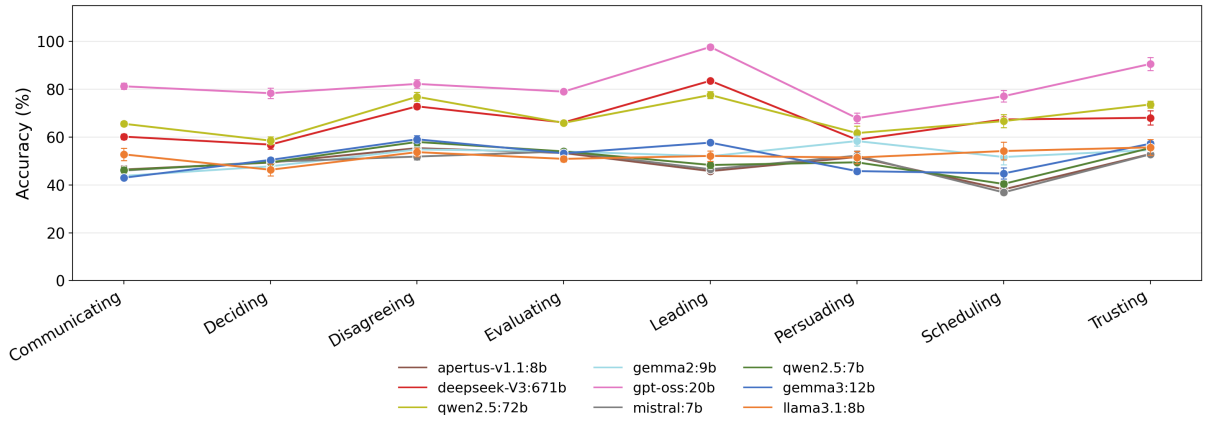
Figure 4: Per-dimension classification accuracy (mean $\pm$ std) for RQ1.a (w/o nationality) and RQ1.b (w/ nationality), across all nine models.

Model	Comm.	Dec.	Disagr.	Eval.	Lead.	Pers.	Sched.	Trust.	Global
<i>RQ2.a</i>									
mistral:7b	45.0 $\pm$ 0.4	46.6 $\pm$ 0.2	50.7 $\pm$ 1.8	51.4 $\pm$ 0.8	48.3 $\pm$ 1.2	53.1 $\pm$ 0.6	44.6 $\pm$ 0.5	53.1 $\pm$ 1.1	49.2 $\pm$ 0.4
qwen2.5:7b	43.9 $\pm$ 1.2	49.7 $\pm$ 1.6	56.3 $\pm$ 0.8	51.9 $\pm$ 1.0	46.1 $\pm$ 0.0	49.0 $\pm$ 0.9	39.8 $\pm$ 1.7	54.6 $\pm$ 0.8	48.9 $\pm$ 0.3
apertus-v1.1:8b	46.9 $\pm$ 1.2	46.9 $\pm$ 2.0	49.1 $\pm$ 2.4	49.0 $\pm$ 0.6	46.2 $\pm$ 3.1	51.0 $\pm$ 2.8	41.7 $\pm$ 3.7	54.4 $\pm$ 2.1	48.3 $\pm$ 0.9
llama3.1:8b	44.6 $\pm$ 1.7	46.9 $\pm$ 3.9	55.2 $\pm$ 3.2	52.1 $\pm$ 2.1	47.6 $\pm$ 2.3	51.6 $\pm$ 1.7	45.4 $\pm$ 5.0	55.7 $\pm$ 1.6	49.7 $\pm$ 0.6
gemma2:9b	43.1 $\pm$ 1.4	47.1 $\pm$ 1.6	56.5 $\pm$ 0.7	54.3 $\pm$ 0.7	53.6 $\pm$ 2.7	53.8 $\pm$ 0.5	45.0 $\pm$ 2.1	54.6 $\pm$ 1.1	51.0 $\pm$ 0.3
gemma3:12b	48.1 $\pm$ 0.6	44.3 $\pm$ 0.2	56.1 $\pm$ 0.5	55.2 $\pm$ 0.6	57.8 $\pm$ 1.6	55.0 $\pm$ 0.0	52.5 $\pm$ 1.9	56.8 $\pm$ 1.1	53.0 $\pm$ 0.4
gpt-oss:20b	55.8 $\pm$ 1.7	54.9 $\pm$ 1.0	70.0 $\pm$ 2.1	64.3 $\pm$ 1.7	79.3 $\pm$ 1.6	54.2 $\pm$ 0.8	63.3 $\pm$ 1.5	68.5 $\pm$ 1.7	63.3 $\pm$ 0.4
qwen2.5:72b	59.3 $\pm$ 1.3	47.6 $\pm$ 1.8	70.7 $\pm$ 3.0	67.4 $\pm$ 2.2	69.8 $\pm$ 1.4	54.6 $\pm$ 1.2	66.3 $\pm$ 2.2	74.8 $\pm$ 0.4	62.9 $\pm$ 0.3
deepseek-V3:671b	53.5 $\pm$ 0.3	50.0 $\pm$ 1.1	60.2 $\pm$ 1.6	52.2 $\pm$ 0.6	65.4 $\pm$ 1.2	55.0 $\pm$ 0.0	62.5 $\pm$ 1.0	65.3 $\pm$ 1.8	57.4 $\pm$ 0.3
<i>RQ2.b</i>									
mistral:7b	46.0 $\pm$ 0.2	49.7 $\pm$ 0.5	51.9 $\pm$ 1.3	54.0 $\pm$ 0.2	46.6 $\pm$ 0.2	52.2 $\pm$ 2.0	36.9 $\pm$ 0.9	52.8 $\pm$ 0.0	49.2 $\pm$ 0.2
qwen2.5:7b	46.2 $\pm$ 1.2	49.4 $\pm$ 0.8	58.0 $\pm$ 1.5	54.0 $\pm$ 1.0	48.3 $\pm$ 1.4	49.4 $\pm$ 0.7	40.4 $\pm$ 1.1	55.4 $\pm$ 0.8	50.2 $\pm$ 0.3
apertus-v1.1:8b	46.4 $\pm$ 1.4	49.3 $\pm$ 1.4	55.4 $\pm$ 0.8	53.3 $\pm$ 1.9	45.8 $\pm$ 0.8	51.7 $\pm$ 0.7	38.1 $\pm$ 0.9	52.9 $\pm$ 0.8	49.4 $\pm$ 0.4
llama3.1:8b	52.8 $\pm$ 2.5	46.3 $\pm$ 2.5	53.7 $\pm$ 1.9	50.9 $\pm$ 1.2	52.1 $\pm$ 2.2	51.4 $\pm$ 2.5	54.2 $\pm$ 3.7	55.7 $\pm$ 3.3	51.8 $\pm$ 0.8
gemma2:9b	43.8 $\pm$ 1.0	47.8 $\pm$ 1.0	54.8 $\pm$ 1.8	54.0 $\pm$ 0.7	52.0 $\pm$ 0.5	58.3 $\pm$ 2.1	51.7 $\pm$ 3.2	54.6 $\pm$ 2.0	51.9 $\pm$ 0.7
gemma3:12b	43.0 $\pm$ 0.5	50.4 $\pm$ 0.9	59.1 $\pm$ 1.5	53.2 $\pm$ 0.7	57.7 $\pm$ 0.6	45.8 $\pm$ 1.0	44.8 $\pm$ 2.3	57.2 $\pm$ 1.1	51.2 $\pm$ 0.6
gpt-oss:20b	81.2 $\pm$ 1.3	78.3 $\pm$ 2.1	82.2 $\pm$ 1.8	79.0 $\pm$ 0.7	97.7 $\pm$ 1.0	67.9 $\pm$ 2.2	77.1 $\pm$ 2.4	90.6 $\pm$ 2.7	81.8 $\pm$ 0.5
qwen2.5:72b	65.6 $\pm$ 1.0	58.5 $\pm$ 1.6	76.9 $\pm$ 1.9	65.9 $\pm$ 0.8	77.6 $\pm$ 1.4	61.7 $\pm$ 2.9	66.7 $\pm$ 2.8	73.6 $\pm$ 1.4	67.8 $\pm$ 0.9
deepseek-V3:671b	60.2 $\pm$ 1.2	56.9 $\pm$ 1.8	74.1 $\pm$ 1.1	65.0 $\pm$ 1.1	83.3 $\pm$ 0.3	58.3 $\pm$ 0.6	66.7 $\pm$ 0.6	70.1 $\pm$ 0.7	65.9 $\pm$ 0.6

Table 3: Per-dimension plausibility-judgment accuracy (mean $\pm$ std) for RQ2.a and RQ2.b, across all nine models. Corresponding line plots are provided in Figure 6.

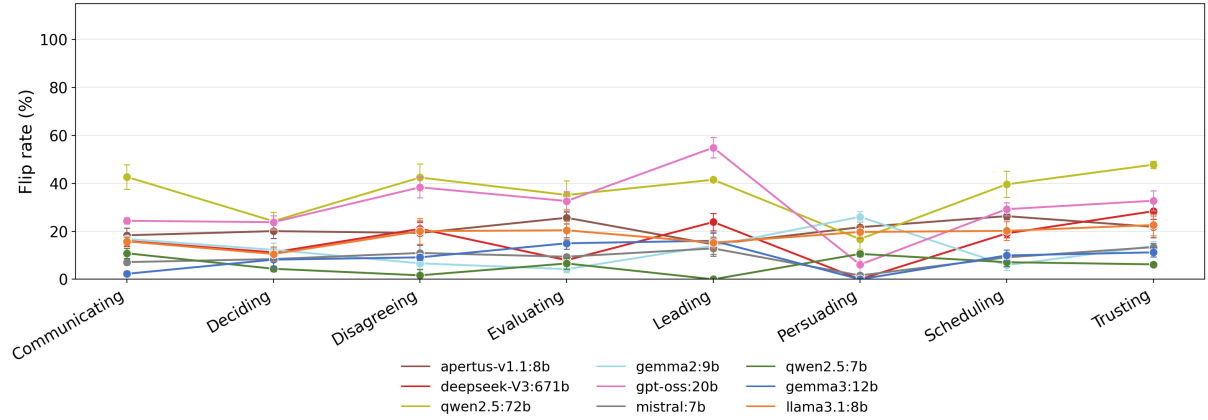


(a) RQ2.a

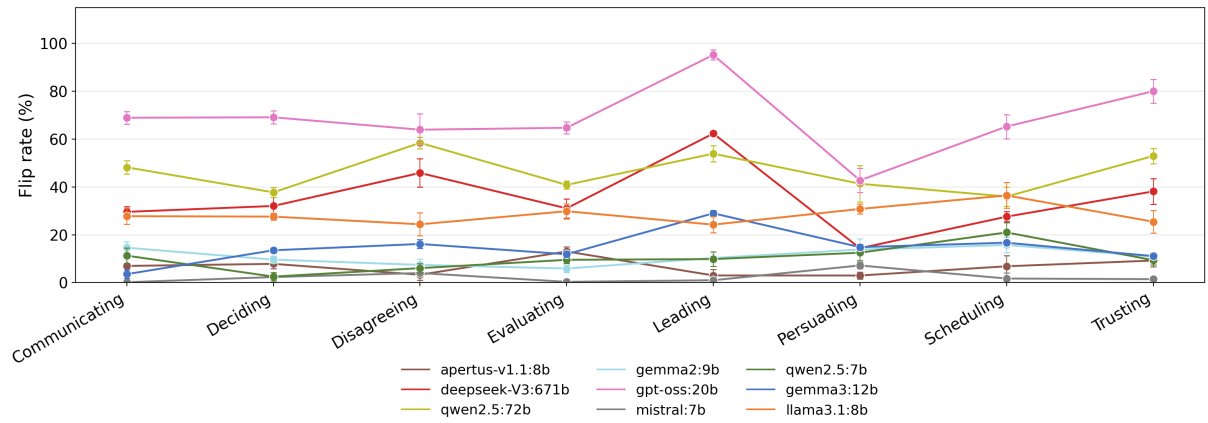


(b) RQ2.b

Figure 5: Per-dimension classification accuracy (mean $\pm$ std) for RQ2.a (w/o dimension definitions) and RQ2.b (w/ dimension definitions), across all nine models.



(a) RQ2.a



(b) RQ2.b

Figure 6: Per-dimension plausibility-judgment flip rate (mean $\pm$ std) for RQ2.a (w/o the dimension definition) and RQ2.b (w/ the dimension definition), across all nine models.