

FrameNet-Cultures: A Benchmark for Evaluating LLMs via Cross-Cultural Frame Semantics

Neda Jamshidi^{1,2}, Anders Søgaard¹, Monica Bianchini²

¹ University of Copenhagen, Denmark

² University of Siena, Italy

neda.jamshidi@student.unisi.it, soegaard@di.ku.dk, monica.bianchini@unisi.it

Abstract

Large language models (LLMs) exhibit cultural biases, yet existing benchmarks rely on closed-form, domain-specific questionnaires. We introduce FRAME-NET-CULTURES, a benchmark for evaluating cultural alignment in LLMs based on Fillmore-style frame semantics. Using the *EveryCulture* encyclopedia, we construct a lexicon of 18 cultural frames (e.g., greeting, child-rearing) across 20 countries, treating it as a structured reference for comparison rather than a definitive representation of contemporary societies. For each frame, we prompt five major LLMs—ChatGPT-5, Gemini-2.5-Flash, Mistral-Large, Qwen-3-Max, DeepSeek-V3.2—three times to generate open-ended instantiations, which are manually annotated and binarized. We measure alignment with country- and continent-level profiles using normalized Hamming distance, and validate cultural recognizability through human evaluation of generated dialogues. Under culture-neutral prompting, outputs align most closely with European profiles, followed by Asian and American ones, indicating a consistent cross-model pattern. With culture-specific prompting, models shift toward the target regions, aligning most strongly with Africa for Ethiopia and with Asia for India. FRAME-NET-CULTURES is the first open-ended benchmark for cultural alignment relying on frame semantics. Data, prompts, and annotations are publicly available at <https://github.com/neda-jamshidi/FrameNet-Cultures>.

1 Introduction

Large language models (LLMs) routinely generate culturally rich descriptions of social practices such as food, greeting, kinship, and religion. Despite this apparent fluency, we lack rigorous methods to evaluate whether these outputs reflect authentic human cultures or instead produce homogenized representations that systematically align with dominant regions, particularly Europe (Blodgett et al.,

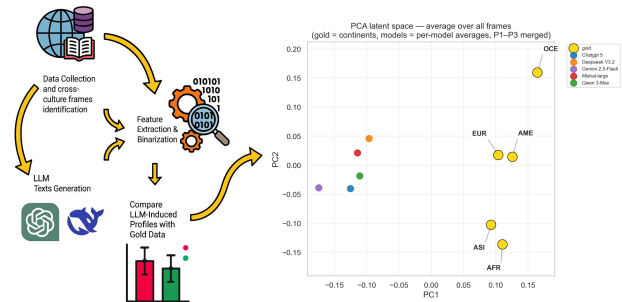


Figure 1: Benchmark construction and evaluation pipeline, from frame extraction and model generation to annotation, feature binarization, and PCA-based analysis (see Sections 3 and 4).

2020; Bender et al., 2021; Hershcovich et al., 2022). Current evaluation paradigms struggle to distinguish genuine cross-cultural variation from surface-level cultural mimicry driven by training priors, making it difficult to assess whether LLMs capture meaningful cultural diversity. Existing approaches rely on closed-form question answering or predefined cultural dimensions (Wang et al., 2024; Chiu et al., 2024), limiting their ability to capture *latent cultural structures* emerging in open-ended generation (Pawar et al., 2024; Shen et al., 2024). Latent cultural structure shapes how models generalize cultural knowledge even when culture is not explicitly queried. As a result, we lack methods for directly comparing model-induced cultural representations with ethnographic reference structures in a quantitative and interpretable way.

Research Questions. This study addresses the overarching question:

Problem Statement

To what extent do LLMs reproduce, distort, or homogenize cultural structures from ethnographic sources, and how do these representations align with dominant regions and shift under different prompting strategies?

Specifically, we explore four research questions:

- **RQ1:** How do LLMs represent cultural practices under open-ended, culture-neutral prompting?
- **RQ2:** How closely do model-induced cultural profiles align with ethnographic references at country and continent levels?
- **RQ3:** What systematic patterns of regional bias and homogenization emerge across models and cultural domains?
- **RQ4:** To what extent can explicit cultural conditioning shift these representations toward target cultures?

Task Framing. FRAMENET-CULTURES evaluates open-ended cultural structure: models generate free-form instantiations of abstract frames grounded in frame semantics (Fillmore, 1982; Baker et al., 1998). Latent cultural representations are inferred from these generations and compared with ethnographic references. This contrasts with (i) closed-form cultural benchmarks that rely on predefined dimensions or explicit statements, and (ii) evaluation settings where cultural attributes are directly specified. In FRAMENET-CULTURES, models are not given feature inventories or cultural profiles; cultural structure must emerge from generation. Under culture-conditioned prompting, target cultures are specified while preserving the same format, enabling controlled analysis without direct leakage.

Approach and Results. We introduce FRAMENET-CULTURES, an open-ended benchmark for evaluating latent cultural structures in LLMs. We extract and binarize cultural features for 20 countries across 18 frames from the *EveryCulture* encyclopedia and compare them to model-induced profiles generated for each frame using three prompts by five models—**ChatGPT-5**, **Gemini-2.5-Flash**, **Mistral-Large**, **Qwen-3-Max**, and **DeepSeek-V3.2** (Figure 1). Under culture-neutral prompting, all models exhibit homogenization and align most closely with European cultural profiles (mean normalized Hamming distance ≈ 0.40 – 0.43 across models), followed by Asian and American profiles, while African and Oceanian regions show weaker alignment (typically ≈ 0.46 – 0.48 across models). At the country level, lower-alignment countries such as Ethiopia and India exhibit substantially higher distances (often > 0.5). When

explicitly prompted to align with target cultures selected based on low baseline alignment (India and Ethiopia), representations shift away from European clusters toward the corresponding regions. Under India-aligned prompting, distances for India decrease to approximately 0.35–0.40, with Asia becoming the nearest continent (approximately 0.38–0.43). Under Ethiopia-aligned prompting, distances for Ethiopia decrease to approximately 0.34–0.36, with Africa becoming the nearest continent (approximately 0.36–0.38). This indicates consistent but partial prompt controllability alongside persistent regional anchoring. These findings are supported by human evaluation of generated dialogues, which shows the highest cultural recognizability under structured, culture-conditioned prompting.

Contributions. Our contributions are: (1) FRAMENET-CULTURES, the first open-ended benchmark for evaluating *latent* cultural structures in LLMs using frame semantics; (2) an ethnographically grounded dataset of binary cultural profiles for 20 countries across 18 frames; (3) a unified evaluation framework for comparing model-induced cultural representations under multiple prompting regimes; (4) a quantitative analysis of cultural alignment, bias, and homogenization using normalized Hamming distance, continent-level aggregation, and PCA; and (5) the public release of all data, prompts, annotations, and code for reproducibility.

Paper Structure. Section 2 reviews related work, Section 3 details dataset creation, Section 4 reports experimental results, Section 5 presents conclusions and future work, and Section 6 discusses the limitations of the research.

2 Related Work

Cross-cultural NLP and culturally situated knowledge. NLP systems encode culturally situated assumptions shaped by English-dominated training data, limiting cross-cultural generalization (Hershcovich et al., 2022; Blodgett et al., 2020; Bender et al., 2021). This has motivated interest in *cultural commonsense*, which varies across societies. Empirical work shows substantial variation in LLMs’ ability to express such knowledge (Shen et al., 2024), and that culturally specific elements often hinder mutual understanding, which LLMs only partially capture (Saha et al., 2025). Several datasets construct cultural knowledge via LLM prompting (MANGO) or large-scale extrac-

tion (CCSK) (Nguyen et al., 2024a,b), while other work focuses on linguistic diversity (Alwajih et al., 2025) or cultural adaptation in dialogue (Havaladar et al., 2025). These approaches rely primarily on model-generated or task-driven annotations, limiting comparison with independent human-authored evidence. In contrast, we ground evaluation in the ethnographic *EveryCulture* encyclopedia. Frame semantics provides an abstraction for cross-cultural comparison (Fillmore, 1982; Baker et al., 1998).

Benchmarks for cultural question answering. Cultural knowledge is commonly evaluated via question answering. CDEval and CulturalBench use multiple-choice or human-authored questions (Wang et al., 2024; Chiu et al., 2024), while BLEnD and MCEval extend this to multilingual settings and bias analysis (Myung et al., 2024; Huang et al., 2025). Larger-scale multilingual and multimodal benchmarks follow similar paradigms (Vayani et al., 2024). Region-specific datasets such as SaudiCulture and multi-agent settings like CulturePark likewise focus on factual recall or stylistic adaptation rather than structural cultural representation (Ayash et al., 2025; Li et al., 2024).

Cultural value alignment and adaptability. Another line of work compares LLM outputs to sociological survey data, showing alignment with values characteristic of English-speaking and Northern European countries (Tao et al., 2024). Prompting strategies can increase apparent cultural alignment (AlKhamissi et al., 2024), while analyses based on Hofstede’s dimensions reveal systematic biases (Masoud et al., 2024). NORMAD evaluates cultural adaptability via norm-based judgments (Rao et al., 2025), but focuses on decision-level compliance rather than latent structure. Related work further studies cultural bias and reward-model behavior (Mak and Luo, 2025; Zhang et al., 2025).

Ethnographic resources and surveys. Structured cultural datasets such as eHRAF, D-PLACE, and WALS provide valuable anthropological resources (Kirby et al., 2016; Haspelmath et al., 2005), but are rarely used to evaluate generated cultural structure. Recent surveys emphasize grounding evaluation in human-authored knowledge (Hershovich et al., 2022; Pawar et al., 2024).

Our contribution. Prior work relies on model-generated annotations (Nguyen et al., 2024a,b; Alwajih et al., 2025), question-answering benchmarks (Wang et al., 2024; Chiu et al., 2024; Ayash et al.,

2025), or value alignment analyses (Tao et al., 2024; AlKhamissi et al., 2024; Masoud et al., 2024; Rao et al., 2025). In contrast, we ground cultural representations in ethnographic data and directly compare them to model-induced profiles under open-ended prompting, enabling analysis of latent cultural structure.

3 Benchmark and Dataset Construction

Our goal is to construct a culturally grounded benchmark that (i) captures cross-societal variation across cultural domains, (ii) relies exclusively on human-authored ethnographic sources, and (iii) enables systematic comparison between LLM-induced representations and documented cultural structures. The pipeline consists of five stages (Sections 3.1–3.5).

3.1 Ethnographic Source and Country Selection

We use the *EveryCulture* encyclopedia,¹ a curated anthropological resource providing narrative descriptions of cultural practices. Its relatively consistent socio-level accounts make it suitable for frame-based annotation. From this source, we select **20 countries** to ensure geographic, linguistic, and cultural diversity across Africa, Asia, Europe, Oceania, and the Americas. The full list is provided in Appendix A.

3.2 Frame Identification

We define **18 cultural frames** corresponding to recurring domains of everyday social life. Following frame semantics (Fillmore, 1982), each frame is treated as a structured conceptual scenario (e.g., greeting, food, marriage) described by culturally relevant attributes. Frames are derived through systematic close reading of *EveryCulture* entries, retaining domains that recur across all selected countries to ensure comparability. Selection is guided by two criteria: (i) recurrence across all selected countries, (ii) independence from country-specific terminology. No LLMs are used in frame discovery or country selection. The resulting inventory includes domains such as *food*, *greeting*, *marriage*, *religion* and *kinship*. The full list is provided in Appendix B.

3.3 Feature Extraction and Binarization

For each frame, we manually define **4–11 features** capturing salient cultural attributes. These correspond to FrameNet-style frame elements,

¹<https://www.everyculture.com/>

adapted to represent culturally grounded realizations. Feature extraction is performed by closely reading the ethnographic text for each frame separately across all selected countries, and identifying recurring, culturally meaningful attributes. Features are selected based on three criteria: (i) recurrence within frame-specific textual descriptions across countries, (ii) explicit grounding in the source text, and (iii) definability as observable structural attributes. This ensures that features are both empirically grounded and consistently identifiable across cultures. Example features include food-related attributes such as ingredients (e.g., rice) and preparation methods (e.g., frying), as well as greeting behaviors (e.g., handshaking, kissing) and child-rearing practices (e.g., extended family participation). Compared to the original FrameNet resource, which contains approximately 1,200 frames with 5–20 frame elements each, our representation is intentionally compact to facilitate cross-cultural comparison and quantitative analysis.

Binarized ethnographic annotation. Each frame is encoded as a country-specific binary vector, where **1 indicates presence** and **0 indicates absence**. This encoding enables direct comparison across societies and between human-authored and model-generated descriptions. A feature is annotated as present if it is explicitly stated or can be clearly inferred from the *EveryCulture* descriptions, and absent otherwise. Explicit negation and non-mention are treated equivalently as absence, ensuring a conservative and consistent mapping from narrative text to structured representations. Annotation follows a fixed guideline based on semantic interpretation rather than surface matching, with rules for handling lexical variation, implicit realizations, and ambiguity (Appendix C). This process yields ethnographic profiles for **18 frames** across **20 countries**, resulting in **360 instances** (20 countries \times 18 frames), which serve as the reference (gold) set for evaluation (Appendix D). All frame identification, feature extraction, and primary annotation are performed by a single annotator² with expertise in natural language processing and linguistics, using standardized guidelines. To assess robustness, an independent second annotator with a background in linguistics and prior annotation experience reproduced the binarized annotations for

²The annotator is an author of this paper. Annotation was conducted voluntarily and without financial compensation.

two frames (*food* and *marriage*). The annotations showed high agreement, indicating that the protocol is reproducible under the provided guidelines.

3.4 Prompt Design and Model Sampling

To elicit model-induced cultural representations, we query five LLMs: **ChatGPT-5** (OpenAI, 2025a), **Gemini-2.5-Flash** (Google DeepMind, 2024), **Mistral-Large** (Mistral AI, 2024), **Qwen-3-Max** (Alibaba Cloud, 2024), and **DeepSeek-V3.2** (DeepSeek AI, 2025). We generate data under two prompting regimes.

1. Culture-neutral prompts. For each frame, we design **three open-ended prompts** that request generic, society-agnostic descriptions of how a cultural practice typically unfolds, without naming any country or cultural feature. Prompts follow three templates: **scene-based** (a detailed observational depiction of a typical instance), **narrative-opening** (the beginning of a story centered on a key moment), and **generalized-pattern** (a high-level description of how the practice operates across society). This design enables systematic analysis of prompting effects and the robustness of model-induced cultural representations (Appendix E). Each of the **5 models** is prompted with **3 prompts** per frame across **18 frames**, yielding **$5 \times 3 \times 18 = 270$ texts** in the culture-neutral setting. Appendix F provides example prompts and outputs for the *marriage* frame.

2. Outlier-alignment prompts. To assess cultural adaptability, we use prompts that explicitly request alignment with India and Ethiopia, selected as outliers due to their low alignment under culture-neutral prompting. We use exactly the same prompt templates as in the culture-neutral setting, with the only modification being the explicit inclusion of the target country name. This ensures a controlled comparison between neutral and culturally conditioned generation. For each target culture (India and Ethiopia), each of the **5 models** is prompted with **3 prompts** across **18 frames**, yielding **$5 \times 3 \times 18 = 270$ texts per culture**. This results in **270 texts for India** and **270 texts for Ethiopia**. Together with the **270 culture-neutral texts**, this yields a total of **810 generated texts**. Example prompts and model outputs are provided in Appendices G and H.

3.5 Binarization of Model-Generated Texts

All **810 model-generated texts** are binarized using the same feature schema and guidelines as the ethnographic data, by the same annotator, ensuring

consistency across datasets. Each text is mapped to a binary feature vector, where **1 indicates presence** and **0 absence**, following the same annotation criteria (Section 3.3). This yields a shared feature space for direct comparison between human cultural profiles and LLM-induced representations. The ethnographic dataset comprises **360 instances**, serving as the reference set for evaluation.

4 Experiments

We evaluate binarized model-induced cultural structures generated under two prompting regimes across 18 frames against ethnographic gold representations constructed for 20 countries across all frames, encoded using the same feature schema. We observe three consistent patterns. First, under culture-neutral prompting, model-induced representations exhibit cultural homogenization and align most closely with European profiles. Second, this Europe-centered alignment persists across aggregation levels (frame, country, and continent), indicating a systematic multi-scale regional bias. Third, under explicit cultural conditioning prompts, representations shift toward the intended target cultures. To characterize these effects, we analyze (i) country- and continent-level alignment, (ii) frame-level structure, (iii) prompt sensitivity, and (iv) global cultural geometry in a shared feature space. All comparisons use ethnographic gold vectors derived from *EveryCulture*. We further complement these analyses with a human evaluation based on generated dialogues to assess whether the structural representations we define are recognizable as culturally grounded.

4.1 Evaluation Metrics and Aggregation

We quantify alignment using **normalized Hamming distance** (Hamming, 1950). For a frame with k binary features, let $x \in \{0, 1\}^k$ denote the ethnographic gold vector and $y \in \{0, 1\}^k$ the model-induced vector:

$$d_{\text{norm}}(x, y) = \frac{1}{k} \sum_{i=1}^k \mathbf{1}[x_i \neq y_i],$$

with values in $[0, 1]$, where lower values indicate closer alignment. For each *model* \times *prompt* \times *frame*, distances are computed for all generations and aggregated using the *mean*, *minimum*, and *maximum*. To analyze higher-level structure, we aggregate representations at the country and continent levels by averaging binary feature vectors.

We further apply Principal Component Analysis (PCA) (Jolliffe, 2002). Let $X \in \mathbb{R}^{n \times d}$ denote the matrix of binary feature vectors, where n is the number of samples and d is the number of features. PCA computes a low-dimensional projection $Z = XW$, where W contains the principal components. We use this projection to visualize cultural representations in a low-dimensional space and analyze clustering, drift, and homogenization between ethnographic and model-induced profiles.

4.2 Country-Level Analysis

We begin with the **270 model-generated instances represented in the binarized feature space** under culture-neutral prompting and analyze alignment at the country level by comparing model-induced and ethnographic profiles using normalized Hamming distance, computed for each *model* \times *prompt* \times *country* combination and summarized across frames. A consistent pattern emerges across all models and prompts: European countries exhibit lower distances (higher alignment), while African and Oceanian countries show higher distances, indicating weaker alignment. This pattern is stable across prompts, suggesting that it is not driven by specific prompt formulations. The gap is quantitatively pronounced: Ethiopia consistently exhibits mean distances above 0.5 across models, while European countries fall within a substantially lower range (approximately 0.37–0.44).

Prompts collapsed per model. To further assess robustness to prompt variation, we aggregate prompts per model to obtain a *Country* \times *Model* comparison based on mean distance. The same alignment pattern persists after aggregation, with Europe remaining the closest region and Ethiopia consistently appearing as a high-distance outlier. This indicates that variation across prompts is small relative to variation across countries, and that alignment differences are primarily driven by underlying cultural representations rather than prompt design. Country-level heatmaps for both separate and aggregated prompt settings are provided in Appendix I, reporting mean, minimum, and maximum normalized Hamming distances.

4.3 Continent-Level Alignment

We next analyze large-scale cultural structure by aggregating profiles at the continent³ level (Africa, Asia, Europe, the Americas, and Oceania). For each continent, we average country-level binary

³Transcontinental countries were grouped by dominant historical-institutional alignment.

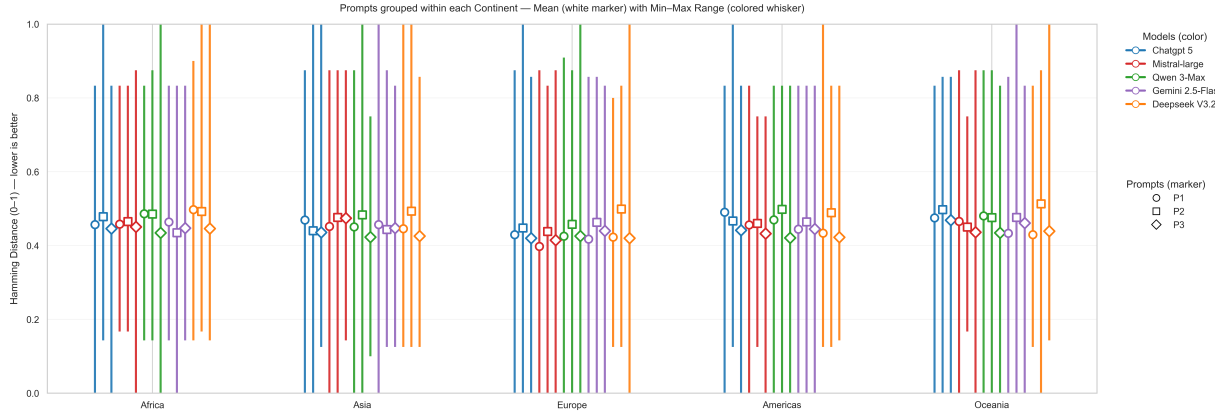


Figure 2: Continent-level normalized Hamming distance under culture-neutral prompting, grouped by prompt. Markers show mean distance across frames; whiskers indicate min–max ranges (0–1, lower is better).

feature vectors and compute normalized Hamming distances for each $model \times prompt \times continent$. Under culture-neutral prompting, European profiles consistently achieve the lowest mean distances (approximately 0.40–0.43 across models), indicating the closest alignment with model-induced representations. Asia and the Americas follow at comparable levels, while Africa and Oceania exhibit substantially higher distances (around 0.5), reflecting systematic divergence. This pattern is stable across models and prompts, indicating a shared Europe-centered structural bias. Detailed statistics (mean, min, max) are reported in Appendix J. Figure 2 visualizes these distances, with markers showing means and whiskers indicating min–max ranges. European profiles consistently have the lowest means (around 0.40 or lower), with minima in some cases approaching zero, indicating near-perfect alignment for certain frames. In contrast, African profiles show higher means (around 0.5) and rarely near-zero minima, indicating systematically weaker alignment across models.

Prompts collapsed per model. Aggregating prompts per model yields a $Continent \times Model$ comparison (Figure 3). European profiles consistently exhibit the lowest mean distances, while African and Oceanian profiles show the highest. This confirms that the observed bias persists after removing prompt-level variability (for more details see Appendix K).

4.4 Frame-Level Cultural Structure

Under culture-neutral prompting, we analyze domain-level alignment by computing mean normalized Hamming distance between model-induced and continent-level ethnographic profiles, averaged across models and prompts. This yields

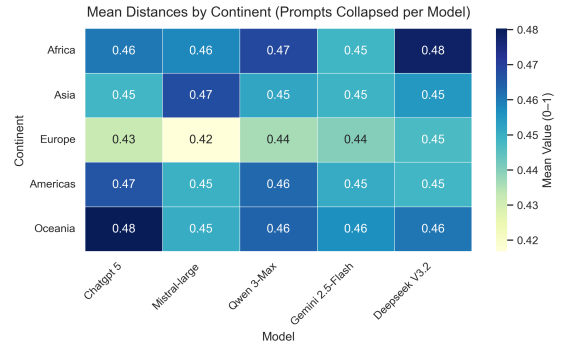


Figure 3: Continent-level mean normalized Hamming distance under culture-neutral prompting (prompts collapsed per model; lower is better).

a $Continent \times Frame$ matrix capturing the closest region per domain. Figure 4 shows the continent with minimum distance for each frame. Across 18 frames, Europe is most frequently the closest match (7/18, $p = 0.051$ under a binomial test with baseline $p = 0.2$), indicating a slight tendency toward Europe relative to the random baseline, followed by Asia, while Africa appears in only one frame. Domain-level patterns are also evident: frames such as *ethnic relations* and *gender roles* align most closely with Europe, whereas only a single *death* frame aligns more closely with Africa. These results confirm that the Europe-centered alignment observed at country and continent levels persists at the domain level. The minimal presence of African alignment further indicates systematic divergence from model outputs. Exact values and frame-level counts are reported in Appendix L.

4.5 Outlier-Alignment Conditioning

To assess whether model-induced representations can be steered toward underrepresented regions, we apply explicit cultural conditioning

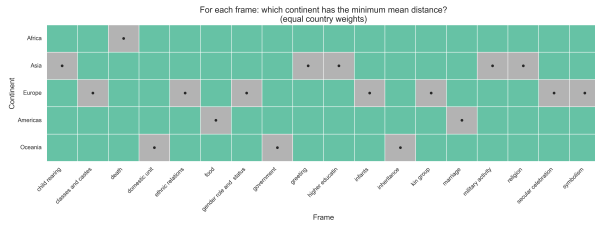


Figure 4: Continent achieving the minimum mean normalized Hamming distance for each cultural frame (all models and prompts aggregated).

prompts that explicitly name the target culture (**India** and **Ethiopia**), selected as outliers due to high baseline distances under culture-neutral prompting (exceeding 0.5 across models). We analyze shifts at the country, continent, and frame levels.

Country-level shifts. Explicit conditioning improves alignment with the target country. Under Ethiopia-specific prompting, model-induced representations are most closely aligned with the Ethiopian ethnographic profile across models (mean distance ranges between 0.34-0.36); a similar pattern holds for India (mean distance ranges between 0.35-0.4). This indicates that models can approximate target cultural structures when guided, although alignment may reflect broader regional similarity. Detailed results are reported in Appendix M.

Continent-level shifts. Under culture-neutral prompting, representations align most closely with Europe. Under India-specific prompting, alignment shifts toward Asia, and under Ethiopia-specific prompting, toward Africa across models. These shifts indicate that explicit cultural cues redirect representations away from the default Europe-centered bias, although the shift remains partial. Results are shown in Figures 5 and 6.

Frame-level shifts. We analyze, per frame, which continent achieves the minimum normalized Hamming distance and test significance using a binomial test ($N = 18$, baseline $p = 0.2$). Under culture-neutral prompting, Asia appears in 5/18 frames and Africa in 1/18 frames. Under India prompting, Asia is most frequent (7/18 frames, $p = 0.051$). Under Ethiopia prompting, Africa is most frequent (9/18 frames, $p = 0.004$). These results show consistent directional movement toward the intended regions, with a statistically significant effect for Africa. Overall, while default generation exhibits systematic bias, models remain responsive to explicit cultural conditioning across

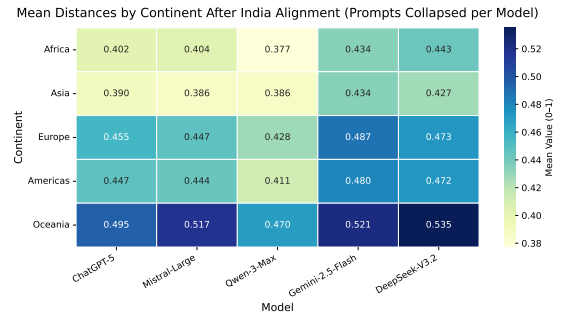


Figure 5: Continent-level mean normalized Hamming distance after India-alignment prompting (prompts collapsed per model; lower is better).

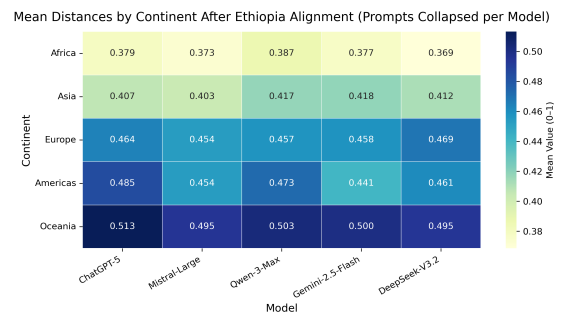


Figure 6: Continent-level mean normalized Hamming distance after Ethiopia-alignment prompting (prompts collapsed per model; lower is better).

all levels. Detailed frame-level results are provided in Appendix M.1.

4.6 Distributional and Prompt Sensitivity Analysis

Frame-level averages can obscure variability across individual generations. We therefore analyze per-sample normalized Hamming distance distributions, both at the model level (prompts collapsed) and per *model* \times *prompt*. Under culture-neutral prompting, distributions largely overlap across models and prompts, indicating low sensitivity to prompt wording. Median distances are around ≈ 0.43 for Prompts 1 and 3, while Prompt 2 more frequently reaches higher values (around ≈ 0.50) due to its narrative formulation. Under explicit cultural conditioning (India or Ethiopia), distributions retain a similar shape (mostly within 0.3–0.5) but become more concentrated around ≈ 0.43 , indicating reduced prompt sensitivity. In some models, Prompt 2 still shows higher medians (around ≈ 0.50), though less frequently and at lower levels than in the neutral setting. At the same time, lower distances shift from predominantly European profiles to the intended target regions. Overall, explicit cultural specification—not prompt wording—is the primary driver of represen-

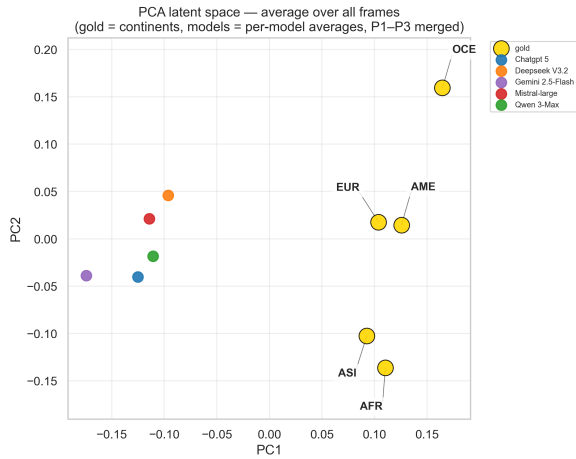


Figure 7: PCA projections averaged over frames, showing continent-level ethnographic profiles and model-induced representations under culture-neutral prompting.

tational change. Detailed distributions are reported in Appendix N.

4.7 Global Cultural Geometry via PCA

We analyze global cultural structure by projecting ethnographic and model-induced profiles into a shared low-dimensional space using PCA, which provides a qualitative visualization of high-dimensional relationships by preserving major variance directions, but is not an exact quantitative measure of alignment. Figure 7 shows projections averaged over frames with prompts merged under culture-neutral prompting. Ethnographic profiles form well-separated continental clusters, while model-induced profiles collapse into a compact central region closest to Europe and more distant from Africa and Oceania. This geometry reflects cultural homogenization and is consistent with the Europe-centered tendency observed earlier. We note that this alignment is defined relative to the constructed ethnographic representations and does not directly correspond to real-world cultural distributions. Prompt-wise projections are provided in Appendix O. Under outlier-alignment prompting, this geometry shifts: model-induced profiles move away from the European-centered region toward the target cultures. Figure 8 illustrates this shift for Ethiopia, where representations move toward the African region while remaining distinct from ethnographic profiles; a similar pattern is observed for India (Appendix P). These results indicate that conditioning repositions representations in latent space without fully recovering ethnographic structure. Aggregated-prompt PCA projections are reported in Appendix P.1.

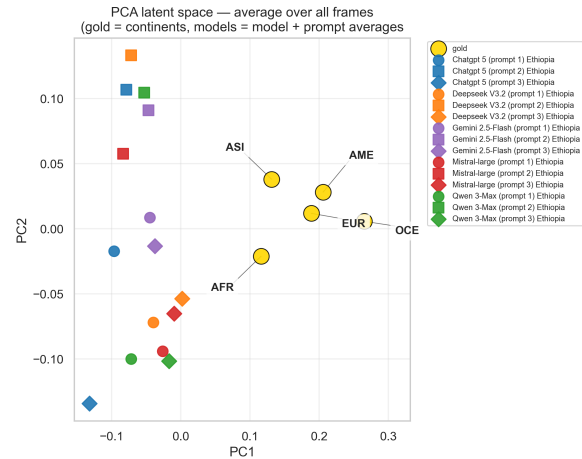


Figure 8: PCA projections under Ethiopia-alignment prompting, compared to continent-level ethnographic profiles.

4.8 Human Evaluation of Cultural Recognizability

To further evaluate the cultural validity of the constructed frames, we conduct a human evaluation based on dialogue generation. For each of the 18 frames, we generate natural, human-like everyday dialogues between two speakers (A and B), each consisting of six turns, with each turn limited to at most 15 words. For each frame, we use five LLMs and three prompting conditions (*Frame+Culture*, *Frame+Culture+Structure*, and *Frame+Structure+Binarized*) across five countries (Iran, Ethiopia, India, the United States of America, and Denmark), in addition to a culture-neutral *Frame-only* condition. In the *Frame* condition, the prompt specifies only a general everyday scenario (e.g., food or greeting) without any cultural cues, serving as a culture-neutral reference. In *Frame+Culture*, the same scenario is augmented with an explicit country mention (e.g., “in Iran”), encouraging the model to ground the dialogue culturally. In *Frame+Culture+Structure*, we further provide a structured list of culturally relevant elements derived from the frame definitions introduced in Section 3.3 to guide generation. In *Frame+Structure+Binarized*, binary constraints are imposed on feature presence (i.e., using the same binarized feature schema derived from these frame definitions) without explicitly naming the culture; we treat this condition as a baseline, as it enforces feature-level constraints without providing explicit cultural grounding. Across all frames, prompting conditions, countries, and models, this results in a total of **800 dialogues**. We use slightly different configurations for ChatGPT and Gem-

| Prompt | Iran | | | | Ethiopia | | | |
|-----------------------------|-------------|-------|-------|-------|-------------|-------|-------|-------|
| | A | B | C | D | A | B | C | D |
| Frame+Culture | 64.0 | 24.0 | 10.0 | 2.0 | 30.0 | 32.0 | 18.0 | 22.0 |
| Frame+Culture+Structure | 80.0 | 18.0 | 2.0 | 0.0 | 44.0 | 16.0 | 18.0 | 20.0 |
| Non-target (All conditions) | 7.4 | 12.94 | 23.42 | 56.24 | 7.85 | 18.62 | 35.69 | 37.85 |

Table 1: Human evaluation results (%). Structured prompting (Frame+Culture+Structure) substantially improves cultural recognizability (higher A) in both Iran and Ethiopia. Non-target dialogues receive low A scores, indicating low false-positive attribution.

ini, while keeping other models consistent. Dialogues are generated using five LLMs: **ChatGPT-5.2** (OpenAI, 2025b), **Gemini Pro** (Gemini Team and Google DeepMind, 2023), **Qwen 3-Max** (Alibaba Cloud, 2024), **DeepSeek v3.2** (DeepSeek AI, 2025), and **Mistral Large** (Mistral AI, 2024). The generated dialogues are presented in randomized order to native annotators recruited via Prolific⁴ (native annotators from Ethiopia, all fluent in English), as well as two native Iranian annotators, fluent in English, who are NLP researchers with annotation experience⁵, who were not informed about the prompting condition or intended target culture. For Ethiopia, four annotators each evaluated 225 dialogues, with a subset of 100 dialogues shared between two annotators to compute inter-annotator agreement. For Iran, two annotators each evaluated 450 dialogues, with 100 dialogues shared between them for agreement estimation. Annotators are asked to rate how confidently each dialogue reflects their own culture using a four-point scale: **A** (very sure: unmistakably my country’s culture), **B** (fairly sure: mostly fits), **C** (not sure: culturally ambiguous), and **D** (sure it is not mine: clearly mismatched). Among the evaluated data, 150 dialogues are associated with Iran and 150 with Ethiopia, while 650 dialogues correspond to non-Iran and non-Ethiopia cases, respectively. **Results show that explicit cultural grounding is necessary for cultural recognizability, while structural guidance further improves it.** Compared to the baseline (*Frame+Structure+Binarized*), which yields low A rates (30% for Iran and 10% for Ethiopia), both *Frame+Culture* and *Frame+Culture+Structure* substantially improve recognizability. The strongest performance is observed under *Frame+Culture+Structure*, where A ratings reach 80% for Iran and 44% for Ethiopia, compared to 64% and 30%, respectively, under

⁴<https://www.prolific.com>

⁵Ethiopian annotators were recruited and compensated via Prolific, while Iranian annotators participated voluntarily.

Frame+Culture. In contrast, dialogues not associated with the target culture receive low A rates, with only 7.4% of non-Iran dialogues labeled as Iranian and 7.85% of non-Ethiopia dialogues labeled as Ethiopian, indicating low false-positive rates and confirming that annotators do not systematically over-attribute cultural identity (Table 1). To assess annotation reliability, we compute inter-annotator agreement on 100 shared items per country. For Iran, agreement is 87.00% with Cohen’s $\kappa = 0.8129$ (*substantial to near-perfect*), while for Ethiopia it is 73.33% with $\kappa = 0.60$ (*moderate*). These results indicate that annotators can consistently identify culturally grounded outputs. Full prompts, example dialogues, annotation procedures, and complete results are provided in Appendix Q.

5 Conclusion & Future Work

We introduced FRAMENET-CULTURES, an open-ended benchmark for evaluating cultural alignment in LLMs using frame semantics grounded in ethnographic data. Our results show: **(RQ1)** Under culture-neutral prompting, LLMs produce culturally fluent but structurally homogenized outputs. **(RQ2)** These align most closely with European profiles at country and continent levels. **(RQ3)** This reflects a consistent regional bias across models, frames, and prompts, indicating a latent structural effect. **(RQ4)** Explicit cultural conditioning induces systematic but partial shifts toward target cultures without fully recovering ethnographic structure. Prompt variation has limited impact, while human evaluation shows that structural guidance improves cultural recognizability but requires explicit grounding. Overall, LLMs exhibit cultural homogenization, Europe-centered alignment, and limited adaptability. FRAMENET-CULTURES provides a framework for diagnosing these behaviors. Future work will extend coverage to more societies and frames, incorporate multilingual and multimodal data, and explore culturally grounded generation.

6 Limitations

In line with ACL policy, we outline the principal limitations of our work. We present these points as concrete shortcomings rather than future directions, refrain from adding citations, and avoid emphasizing the strengths of our approach.

Ethnographic source bias and coverage gaps. All cultural ground truth in FRAMENET-CULTURES is derived from the *EveryCulture* encyclopedia. While this resource provides structured, human-authored ethnographic descriptions, its coverage varies substantially across countries and cultural domains. Some entries are outdated, unevenly detailed, or omit practices that are culturally salient but insufficiently documented. As a result, certain features may be absent from the gold data despite their relevance, affecting both feature extraction and distance-based evaluation. Reliance on a single source also limits our ability to assess the robustness of extracted cultural structures across alternative documentation styles or datasets.

Temporal scope of cultural descriptions. The ethnographic material in *EveryCulture* reflects cultural practices documented over past decades. While some domains (e.g., food practices, ritual structures) tend to be relatively stable, others (e.g., gender roles, education, social norms) may evolve over time. Consequently, divergence between model outputs and gold profiles may partially reflect temporal shifts rather than true cultural misalignment.

Country-level abstraction. Cultural representations are constructed at the country level, implicitly treating each country as culturally homogeneous. This abstraction obscures intra-country variation related to region, ethnicity, class, religion, or urban-rural divides, and may mask culturally salient differences that exist below the national level.

Binary feature encoding. We represent cultural practices using manually defined binary features. While this enables systematic comparison and interpretability, it necessarily collapses graded, contested, or context-dependent practices into discrete values. In addition, the choice of feature inventory influences which cultural distinctions are made salient: alternative feature definitions or the inclusion of different attributes could yield different alignment patterns. Because encyclopedic descrip-

tions are not exhaustive, features not mentioned in the source text are treated as absent, which may penalize model outputs that include valid but undocumented practices. As a result, subtle distinctions, frequency effects, and internal variability within cultural domains are not fully captured.

Single-annotator subjectivity. All ethnographic and model-generated texts were annotated by a single annotator using a fixed feature schema. Although detailed guidelines were applied consistently and a small-scale validation on two randomly selected frames (food and marriage) yielded perfect agreement with an untrained annotator, feature assignment nonetheless involves interpretive judgment, particularly when descriptions are implicit, ambiguous, or metaphorical. The limited scope of this validation does not substitute for systematic multi-annotator labeling, and the absence of broader inter-annotator agreement analysis restricts our ability to quantify annotation uncertainty and incorporate diverse cultural perspectives, allowing individual biases or blind spots to persist.

Prompting assumptions. Our prompts elicit generic, society-independent descriptions and, in outlier settings, explicitly request alignment with a named culture. These controlled prompting conditions do not reflect many real-world usage scenarios, where cultural cues may be indirect, multilingual, or ambiguous. Consequently, observed prompt sensitivity may not generalize to interactive or naturally occurring contexts.

Model and language scope. All evaluated models are large, general-purpose LLMs primarily trained on high-resource languages. The findings may not generalize to smaller models, domain-specific systems, or models trained predominantly on non-English corpora. In addition, all experiments are conducted in English, limiting conclusions about cultural representation in other languages.

Metric limitations. Normalized Hamming distance assumes feature independence and equal weighting across features. Cultural practices, however, may be correlated or hierarchically structured, and missing a peripheral feature is penalized equally to missing a central one. While this metric provides a transparent measure of structural divergence, it does not capture narrative coherence, pragmatic appropriateness, or internal consistency of generated texts. Low distance does not guarantee

culturally appropriate generation, and high distance may reflect alternative but plausible realizations.

Limitations of PCA Analysis Our use of PCA is intended for qualitative visualization of global structure rather than precise quantitative evaluation. As a linear dimensionality reduction technique, PCA preserves only the dominant variance directions and may distort finer relationships present in the original high-dimensional space. Consequently, distances and relative positions in the projected space should not be interpreted as exact measures of cultural alignment. In addition, the projection is sensitive to aggregation choices (e.g., across prompts or frames), which can affect the covariance structure and thus the resulting geometry. Therefore, PCA results should be interpreted as illustrative of general trends—such as directional shifts and clustering behavior—and considered in conjunction with quantitative metrics (e.g., normalized Hamming distance) used in the main analysis.

Potential training data overlap. Because *EveryCulture* is publicly accessible and predates many modern LLM training corpora, some degree of overlap between training data and the source material cannot be ruled out. If present, such overlap could influence measured alignment. However, our evaluation focuses on structural patterns in feature space rather than surface-level similarity, and the extent to which memorization affects these patterns cannot be directly determined.

Interpretive scope of structural patterns. The observed continent-level clustering and European anchoring arise within the specific ontology, feature encoding, and distance metric used in this work. These patterns should therefore be interpreted as properties of the induced representational space rather than definitive sociocultural claims about real-world cultural similarity.

Environmental cost. All model inference and analysis were conducted using large-scale pre-trained models. Although no training or fine-tuning was performed, repeated generation across frames, prompts, and models incurs non-trivial computational cost and associated environmental impact.

Potential for misuse. FRAMENET-CULTURES is intended for research on cultural representation in language models. It is not validated for assessing real-world cultural competence, making normative judgments about societies, or ranking cultures. Use

of the benchmark for high-stakes or prescriptive applications would be inappropriate.

References

- Alibaba Cloud. 2024. Qwen-3-max. <https://qwenlm.github.io/blog/qwen3/>. Official release announcement, Accessed: 2026.
- Badr AlKhamissi, Muhammad ElNokrashy, and Mai et al. AlKhamissi. 2024. [Investigating cultural alignment of large language models](#). *arXiv preprint arXiv:2402.13231*.
- Fakhraddin Alwajih, Abdellah El Mekki, and Samar Mohamed et al. Magdy. 2025. [PALM: A culturally inclusive and linguistically diverse dataset for arabic LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.
- Lama Ayash, Hassan Alhuzali, and Ashwag et al. Alasmari. 2025. [Saudiculture: A benchmark for evaluating large language models' cultural competence within saudi arabia](#). *arXiv preprint arXiv:2503.17485*.
- Collin Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 86–90.
- Emily M. Bender, Timnit Gebru, and Angelina et al. McMillan-Major. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A critical survey of “bias” in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.
- Yu Ying Chiu, Liwei Jiang, and Bill Yuchen et al. Lin. 2024. [Culturalbench: A robust, diverse, and challenging cultural benchmark by human-ai cultural teaming](#). *arXiv preprint arXiv:2410.02677*.
- DeepSeek AI. 2025. Deepseek-v3.2. <https://api-docs.deepseek.com/news/news251201>. Official release announcement, Accessed: 2026.
- Charles J. Fillmore. 1982. Frame Semantics. *Linguistics in the Morning Calm*, pages 111–137.
- Gemini Team and Google DeepMind. 2023. [Gemini: A family of highly capable multimodal models](#). *arXiv preprint arXiv:2312.11805*.
- Google DeepMind. 2024. Gemini 2.5 flash. <https://deepmind.google/technologies/gemini/>. Official model page, Accessed: 2026.
- Richard W. Hamming. 1950. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160.
- Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. 2005. *The World Atlas of Language Structures*. Oxford University Press, Oxford.
- Shreya Havaldar, Young Min Cho, and Sunny et al. Rai. 2025. [Culturally-aware conversations: A framework & benchmark for LLMs](#). In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 220–229, Suzhou, China. Association for Computational Linguistics.
- Daniel Hershcovich, Stella Frank, and Heather et al. Lent. 2022. [Challenges and strategies in cross-cultural nlp](#). *arXiv preprint arXiv:2203.10020*.
- Shulin Huang, Linyi Yang, and Yue Zhang. 2025. [MCEval: A dynamic framework for fair multilingual cultural evaluation of LLMs](#). *arXiv preprint arXiv:2507.09701*.
- Ian T. Jolliffe. 2002. *Principal Component Analysis*, 2 edition. Springer.
- Kathryn R. Kirby, Russell D. Gray, and Simon J. et al. Greenhill. 2016. D-PLACE: A global database of cultural, linguistic, and environmental diversity. *PLOS ONE*, 11(7):e0158391.
- Cheng Li, Damien Teney, and Linyi et al. Yang. 2024. [Culturepark: Boosting cross-cultural understanding in large language models](#). *arXiv preprint arXiv:2405.15145*.
- Moon-Kuen Mak and Tiejian Luo. 2025. [A framework for evaluating cultural bias and historical misconceptions in LLM outputs](#). *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 5(3):100235.
- Reem I. Masoud, Ziquan Liu, and Martin et al. Ferianc. 2024. [Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions](#). *arXiv preprint arXiv:2309.12342*.
- Mistral AI. 2024. Mistral large. <https://mistral.ai/news/mistral-large>. Official release announcement, Accessed: 2026.
- Junho Myung, Nayeon Lee, Yi Zhou, and Jiho et al. Jin. 2024. [BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages](#). *arXiv preprint arXiv:2406.09948*.
- Thi Nguyen, Isar Nejadgholi, and Iryna Gurevych. 2024a. Cultural commonsense knowledge for intercultural dialogues. *arXiv preprint arXiv:2402.10689*.
- Thi Nguyen, Isar Nejadgholi, and Iryna Gurevych. 2024b. Extracting cultural commonsense knowledge at scale. *arXiv preprint arXiv:2210.07763*.
- OpenAI. 2025a. Chatgpt (gpt-5). <https://platform.openai.com/docs/models>. Model version: GPT-5, Accessed: 2026.

OpenAI. 2025b. Chatgpt (gpt-5.2). <https://platform.openai.com/docs/models>. Model version: GPT-5.2, Accessed: 2026.

Siddhesh Pawar, Junyeong Park, and Jiho et al. Jin. 2024. [Survey of cultural awareness in language models: Text and beyond](#). *arXiv preprint arXiv:2411.00860*.

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.

Siqi Shen, Lajanugen Logeswaran, and Moontae et al. Lee. 2024. [Understanding the capabilities and limitations of large language models for cultural commonsense](#). *arXiv preprint arXiv:2405.04655*.

Yan Tao, Olga Viberg, and Ryan S. et al. Baker. 2024. [Cultural bias and cultural alignment of large language models](#). *PNAS Nexus*, 3(9):pgae346.

Ashmal Vayani, Dinura Dissanayake, and Hasindri et al. Watawana. 2024. [All languages matter: Evaluating LMMs on culturally diverse 100 languages](#). *arXiv preprint arXiv:2411.16508*.

Yuhang Wang, Yanxu Zhu, and Chao et al. Kong. 2024. [Cdeval: A benchmark for measuring the cultural dimensions of large language models](#). *arXiv preprint arXiv:2403.12158*.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2025. [Evaluating and improving cultural awareness of reward models for LLM alignment](#). *arXiv preprint arXiv:2509.21798*.

A Selected Countries

The set of countries included in our analysis comprises: *Australia, Brazil, China, Denmark, Egypt, Ethiopia, France, Georgia, Germany, India, Iran, Italy, Japan, Mexico, New Zealand, Nigeria, Russia, Saudi Arabia, Turkey, and the United States of America*.

These countries are selected to maximize geographic, cultural, and socio-economic diversity across continents, including Europe, Asia, Africa,

Oceania, and the Americas. The selection covers a wide range of cultural traditions, languages, religions, and social practices, enabling robust cross-cultural comparison within a shared annotation framework. Additionally, all selected countries are consistently documented in the *EveryCulture* encyclopedia, ensuring comparability and sufficient ethnographic detail for frame-based annotation.

B Cultural Frames

The full set of cultural frames used in this study consists of the following 18 frames: *food, greeting, marriage, religion, secular celebrations, symbolism, ethnic relations, government, classes and castes, gender roles and status, domestic units, kin groups, inheritance, child rearing, infants, higher education, military activity, and death*.

These frames are selected to cover a broad spectrum of cultural practices, ranging from everyday interactions (e.g., food, greeting) to institutional and societal structures (e.g., government, education, kinship systems). The selection is grounded in the thematic organization of the *EveryCulture* encyclopedia, ensuring consistency across countries and sufficient ethnographic detail.

Importantly, the frames are chosen such that they support structured annotation through binary cultural features, enabling systematic comparison of cultural representations across societies and between human-authored and model-generated data within a shared feature space.

C Annotation Guidelines

This appendix describes the annotation protocol used to construct both the ethnographic gold profiles and the model-induced cultural representations. The goal of the protocol is to ensure consistent, reproducible mapping from narrative text to binary cultural features across all frames, countries, models, and prompting conditions.

C.1 Overview of the Annotation Pipeline

The construction of the benchmark proceeds in three stages: (i) frame identification, (ii) feature definition, and (iii) binarized annotation. Frames are first selected through close reading of *EveryCulture* entries, retaining only domains that recur across all selected countries to ensure cross-cultural comparability. For each frame, a set of features is then defined based on recurring, explicitly grounded, and structurally observable cultural attributes present across the selected countries. Fi-

nally, both ethnographic texts and model-generated outputs are annotated using a shared binary schema, where each feature is assigned a presence (1) or absence (0) value according to the principles described below. This unified pipeline ensures that all representations are directly comparable within a consistent feature space.

C.2 General Annotation Principle

Annotation is based on *semantic interpretation* rather than surface string matching. A cultural feature is annotated as present if the underlying cultural practice is explicitly stated or can be clearly inferred from the text, even when expressed using alternative terminology, culturally specific vocabulary, or indirect descriptions. Literal keyword matching is neither required nor sufficient for positive annotation.

C.3 Binary Label Definition

Each feature is assigned a binary value according to the following rules:

- **Present (1):** The feature is explicitly described or can be clearly inferred from the text.
- **Absent (0):** The text explicitly states that the feature does not occur, or provides no evidence for its presence.

Explicit negation and non-mention are treated equivalently and both result in an absence label.

C.4 Lexical Variation and Semantic Equivalence

Lexical variation is resolved through semantic equivalence. If a cultural practice is expressed using a regional variant, culturally specific term, or synonym, the corresponding feature is annotated as present. For example, references to *aloo* are treated as instances of potatoes in the food frame. More generally, culturally localized terminology is mapped to the appropriate abstract feature whenever the intended meaning is unambiguous.

C.5 Implicit Realization of Cultural Practices

Features are not required to be expressed using their canonical labels in order to be annotated as present. When a described practice clearly instantiates the underlying cultural concept, the corresponding feature is marked as present even if the explicit feature name does not appear in the text.

For example, in the *food* frame, references to named dishes that are conventionally prepared through slow cooking—such as *doro wat*

in Ethiopian cuisine—are annotated as instances of stewing, even when the term *stew* is not explicitly mentioned.

C.6 Generalization from Broader Descriptions

When a feature refers to participation of a specific social group or practice, broader descriptions encompassing that group are treated as positive evidence. For instance, in the *child-rearing* frame, references to extended family involvement are annotated as indicating grandparent participation unless the text explicitly excludes grandparents. This principle applies analogously across other frames involving kinship, caregiving, or social roles.

C.7 Conservative Resolution of Ambiguity

In cases where a cultural practice cannot be confidently inferred from the text, the feature is annotated as absent. The protocol prioritizes internal consistency and precision over recall, ensuring that positive labels reflect clear evidence rather than speculative interpretation.

C.8 Annotator Consistency and Scope

All annotations are performed by a single annotator—an author of this paper—following the same guidelines across all ethnographic sources and model-generated texts. The annotator is an NLP researcher with substantial experience in structured annotation and in interpreting linguistic and cultural representations. Annotation was conducted voluntarily and without financial compensation.

To assess annotation consistency, we additionally conducted a small-scale validation on two randomly selected frames (*food* and *greeting*), in which a second annotator—an NLP researcher with a linguistic background and prior annotation experience—independently applied the same guidelines. The annotation was conducted voluntarily and without financial compensation. Agreement was perfect on these frames under the applied guidelines.

To ensure methodological consistency, ethnographic gold profiles are re-annotated using the same criteria applied to model-generated outputs. This guarantees that all comparisons are performed within a shared and consistent representational framework.

| Rice | Potato | Stewing | Boiling | Frying | Fork | Spoon | Knife | Cheese | Alcohol Drink |
|------|--------|---------|---------|--------|------|-------|-------|--------|---------------|
| 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |

Table 2: Illustrative binary feature encoding for the *food* frame (Denmark), derived from the ethnographic excerpt *Food Practices in Denmark*.

D Illustrative Annotated Frame Example

This appendix provides a detailed illustrative example of an annotated cultural frame. We include a visual example showing the annotation of Danish food features, with the corresponding binary feature encoding reported in Table 2.

Ethnographic Excerpt: Food Practices in Denmark

Danes eat most of their meals at home and in private settings, although public dining places ranging from small hot dog stands to fancy restaurants are available and are used. A breakfast of coffee, bread, or cereal is eaten at home. Sunday breakfast commonly includes fresh bakery bread, boiled eggs, juice, tea or coffee, and the Sunday newspaper. Lunch at a work place, school, or institution is either homemade or available in kitchens or canteens, offering open sandwiches, hot meals, or a buffet table. It also may be bought at butcher shops, cafes, and sandwich bars. Open sandwiches are traditional, consisting of rye bread with salami, liver pâté, herring, roast pork, fried plaice, cod roe, cheese, chocolate, or fruit. Dinner at home traditionally consisted of an appetizer, a main course, and dessert. Soup, porridge and fish dishes were served but today are rarely eaten on a daily basis. A main course is traditionally composed of boiled potatoes, boiled vegetables such as green beans and cauliflower, and fried meat such as meat balls, cutlets, or roast pork served with brown gravy. Pizza, pasta, rice, chicken, and turkey have become common food items among young people. Imported fruit, vegetables, and spices are also common. Inns often dating back several centuries throughout the country offer traditional Danish food. Pizzerias are found in small towns and cities. In larger cities, there are Chinese, Italian, and Greek restaurants, along with fast-food establishments from America, the Middle East, and South America and restaurants that serve Danish open sandwiches (smørrebrød) and pastry. Food taboos include pet animals such as cats, dogs, and horses. The ecological movement and informed consumers have been mutually dependent since the 1970s. The demand for and production of organically grown foods have grown, and most supermarkets offer a range of organically grown vegetables, meat, and dairy products.

Food Customs at Ceremonial Occasions. Danes eat or drink at every social occasion, preferably traditional dishes, cakes, and drinks. However, the act of drinking and eating together is considered more important than what is actually consumed. Formal social occasions include birthdays, weddings, anniversaries, baptisms, confirmations, graduations, and funerals. Private parties held in community centers or restaurants are common. Hosts spend from one to six months' salary on a formal party for rent, food, drinks, and musicians. Holidays with special meals include New Year's Eve, Easter, Martin Mass, and Christmas. New Year's Eve traditionally is celebrated with boiled cod, Easter with elaborate lunches and roast lamb for dinner, and Martin Mass with roast goose. The traditional Christmas Eve dinner includes roast pork, roast duck, or goose stuffed with prunes, served with pickled red cabbage, white boiled potatoes, fried brown sugared potatoes, and thick brown gravy. Desserts include rice porridge and ris a la mande (rice porridge mixed with whipped cream, almonds, and vanilla and served with hot cherry sauce). At Christmas and Easter, special seasoned beers are sold. Christmas is celebrated by eating a traditional extravagant lunch and dinner that bring the family together.

E Prompt Templates

For each cultural frame, we design three open-ended prompts that elicit generic, society-agnostic descriptions of how a cultural practice typically unfolds, without naming any country or cultural feature. The prompts are designed to capture complementary aspects of cultural expression through different narrative styles.

Scene-based. This prompt requests a detailed, observational description of a typical instance of the cultural practice. It emphasizes concrete elements such as actions, interactions, gestures, language, and atmosphere, encouraging grounded and descriptive outputs.

Narrative-opening. This prompt asks the model to write the opening of a story centered on a key moment related to the cultural practice. It captures how the practice is introduced and framed in a narrative context, often highlighting salient events or transitions.

Generalized-pattern. This prompt requests a high-level description of how the cultural practice operates across society. It focuses on recurring patterns, social norms, and structural aspects, rather than specific individuals or events.

All prompts are constrained to remain open-ended and culturally unspecified, ensuring that any cultural signals in the generated text emerge from the model rather than being explicitly provided in the prompt.

F Illustrative Model-Generated Texts for the Marriage Frame

This appendix presents illustrative Gemini-generated examples for the *marriage* frame under three culture-neutral prompts. For each example, we report the prompt and the corresponding output. These examples are provided for qualitative illustration only.

F.1 Gemini, Prompt 1 (Culture-Neutral)

Prompt. Describe in detail how marriage unfolds within a society, as if you are observing the process from its beginning to its continuation in daily life. Capture the gestures, words, and atmosphere that define commitment and partnership. Write in a novel-like, observational tone under 200 words, and you may include direct quotes.

Generated text.

The societal mechanism of marriage begins with a subtle shift: the proposal. A bended knee, a hushed, “Will you?” transforms private affection into public intent. This moment, often trembling with anticipation and a shared smile, solidifies the transition. Next, the ritual: a blur of white lace, deep vows exchanged before witnesses. “To have and to hold, from this day forward. . .” The atmosphere is thick with expectation, the air scented with flowers and formal promises.

Then comes the quiet hum of daily life. The grand gestures fade, replaced by the minute choreography of partnership. It is the shared silence over morning coffee, the hand briefly resting on a shoulder in the kitchen, the soft consultation: “Did you remember to call the bank?” Commitment settles not in spoken oaths, but in the unspoken trust that they will navigate the mundane and the crises together, a constant, low-burning flame of shared identity.

F.2 Gemini, Prompt 2 (Culture-Neutral)

Prompt. Write the opening paragraph of a story that begins with a moment in which marriage takes shape — a decision, a union, or a turning point between two people. Let it feel like the start of a larger narrative that reveals how marriage functions and is understood in the society. Keep it under 200 words, and you may include brief dialogue.

Generated text.

The clock in the hall chimed the half-hour, its brass weights catching the sliver of moonlight, just as Elara set down the ivory teacup. “The dowry is agreed upon, then?” she asked, her voice steady, an unnerving counterpoint to the thrumming silence of the ancestral drawing-room. Across the mahogany table, Lord Theron, his profile etched by shadow and ambition, gave a slow, deliberate nod. “It is. Three estates, as we stipulated.”

This was not a romance beginning; it was the ratification of a treaty. In their society, marriage was not the consequence of love, but its scaffolding—a framework of

mutually beneficial alliances built upon land, titles, and a calculated exchange of futures. Elara looked not at his face, but at the gold signet ring on his finger, the symbol of the institution she was about to enter, and understood that the true complexity of their life together would lie not in their personal feelings, but in how skillfully they upheld the public architecture of their union. The wedding day was merely the announcement; the marriage itself would be the negotiation.

F.3 Gemini, Prompt 3 (Culture-Neutral)

Prompt. Describe, in a novel-like way, how marriage typically operates within a society. Do not focus on particular individuals; instead, narrate the general course — how unions form, are lived, and sometimes dissolve. Capture the blend of emotional, social, and legal dimensions that surround the institution. Keep it under 200 words and you may include short quoted phrases.

Generated text.

Marriage begins as an economic treaty cloaked in the finery of affection. Societies necessitate its structure, demanding that two individuals—often guided by familial strategy rather than raw desire—transition from separate entities to a sanctioned “we.” The legal decree, the State’s stamp of approval, elevates the private commitment to a public institution, cementing shared property and mutual obligation. For decades, the union functions as a quiet engine of stability: the shared labor of raising children, the whispered consultations on finances, the silent understanding that “we manage this together.” Commitment becomes less a feeling than a daily, repetitive action.

Yet, when the architecture weakens, the disintegration is rarely sudden. It starts with small gaps—unanswered questions, separate silences—before the legal apparatus is summoned once more to perform the slow, painful work of unwinding. The shared history is carved into discrete halves, property partitioned, and the “we” reverts, legally and socially, to

two separate “I”s, the societal contract fulfilled, then formally terminated.

G Example Prompt Formulations

This appendix provides example prompt formulations used in the *greeting* frame under different prompting regimes. We report one set of culture-neutral prompts and corresponding outlier-alignment prompts for Ethiopia and India to illustrate the parallel structure across conditions.

G.1 Culture-Neutral Prompts (Greeting Frame)

1. **Prompt 1.** Describe in detail the first moments when two people meet in a formal context, as if you are physically present and watching it unfold. Use sensory detail and write it in a novelistic style under 200 words. You may include direct quotes.
2. **Prompt 2.** Write the opening paragraph of a story that begins with two people meeting in a formal context. Let it feel like the start of a larger narrative, setting mood and tone as well as describing their first gestures and words. Keep it under 200 words and you may include direct quotes.
3. **Prompt 3.** Describe, in a novel-like way, how a formal greeting usually unfolds. Do not focus on two specific individuals, but narrate the typical flow of the first seconds in such an encounter as if you were telling a scene from life. Keep it under 200 words and you may include direct quotes.

G.2 Ethiopia-Aligned Prompts (Greeting Frame)

1. **Prompt 1.** Describe in detail the first moments when two people in Ethiopian society meet in a formal context, as if you are physically present and watching it unfold. Use sensory details and write in a novelistic style under 200 words. You may include direct quotes.
2. **Prompt 2.** Write the opening paragraph of a story that begins with two people in Ethiopian society meeting in a formal context. Let it feel like the start of a larger narrative, setting mood and tone as well as describing their first gestures and words. Keep it under 200 words and you may include direct quotes.

3. **Prompt 3.** Describe, in a novel-like way, how a formal greeting usually unfolds within the Ethiopian society. Do not focus on two specific individuals, but narrate the typical flow of the first seconds in such an encounter as if you were telling a scene from life. Keep it under 200 words and you may include direct quotes.

G.3 India-Aligned Prompts (Greeting Frame)

1. **Prompt 1.** Describe in detail the first moments when two people in Indian society meet in a formal context, as if you are physically present and watching it unfold. Use sensory details and write in a novelistic style under 200 words. You may include direct quotes.
2. **Prompt 2.** Write the opening paragraph of a story that begins with two people in Indian society meeting in a formal context. Let it feel like the start of a larger narrative, setting mood and tone as well as describing their first gestures and words. Keep it under 200 words and you may include direct quotes.
3. **Prompt 3.** Describe, in a novel-like way, how a formal greeting usually unfolds with in the Indian society. Do not focus on two specific individuals, but narrate the typical flow of the first seconds in such an encounter as if you were telling a scene from life. Keep it under 200 words and you may include direct quotes.

H Illustrative Examples of Model-Generated Texts

This appendix provides illustrative examples of model-generated texts for the *greeting* frame, produced under different prompting conditions. The examples are generated using the same prompt template (Prompt 3) under culture-neutral, Ethiopia-aligned, and India-aligned settings. Alongside these examples, Table 3 reports the corresponding binarized feature annotations, making explicit how each text instantiates greeting-related practices.

H.1 ChatGPT-5, Prompt 3 (Culture-Neutral)

A formal greeting often begins with the careful choreography of approach: two figures closing the space between them at a measured pace, postures adjusted to signal respect. There is a subtle pause, almost imperceptible, in which both seem to register the other's presence

fully before a word is spoken. Eyes meet briefly—steady, not lingering—and then one voice breaks the silence.

“Good morning,” or “It’s a pleasure to meet you,” comes first, the tone chosen to balance warmth with restraint. A hand is extended, palm angled just so, the gesture carrying centuries of learned convention. The other responds in kind, the clasp brief, firm, and released almost as soon as it begins, leaving behind the faint memory of pressure.

Smiles may appear—measured, contained—as though permission for ease has not yet been granted. Sometimes the exchange includes a small bow of the head, or the faintest lean forward, acknowledging the shared gravity of the moment. The setting, whether a polished office, a quiet corridor, or a ceremonial hall, seems to heighten the weight of these gestures. In those few seconds, roles are established, expectations set, and the invisible script of formality takes hold.

H.2 ChatGPT-5, Prompt 3 (Ethiopia-Aligned)

A formal greeting in Ethiopia often begins before a single word is spoken. Two people approach with measured steps, slowing just enough to signal respect. Their shoulders remain straight, eyes briefly lowered, and the air between them grows still, as if the moment itself asks for composure.

Hands meet first—usually a firm handshake, sometimes drawn slightly inward so their shoulders almost touch. It is a gesture both restrained and warm. A soft pause follows, a heartbeat of acknowledgment. The scent of frankincense from a nearby doorway or the faint aroma of fresh buna might drift through the scene.

Only then come the words, offered with calm clarity: “Enkwan dehna metah”—welcome—or “Selam,” simple and precise. The response is equally respectful: “Amesegenallo,” thank you.

A small nod, a slight leaning toward one another, and the tone of the encounter

is set. In those brief seconds, hierarchy, courtesy, and intention quietly arrange themselves, creating the foundation on which the rest of the interaction will unfold.

H.3 ChatGPT-5, Prompt 3 (India-Aligned)

A formal greeting in Indian society arrives with a quiet choreography. It begins at the threshold, where shoes hesitate and bodies subtly align themselves to the room. The air holds familiar traces—incense from earlier prayers, paper, warm tea. Backs straighten. Faces settle into attentive calm. Before words, there is recognition: a measured glance, a pause that signals respect.

Hands rise, palms pressed together in namaste, fingers pointing upward, the gesture gentle but deliberate. “Namaskar,” someone says, voice even, neither loud nor hurried. Sometimes the gesture gives way to a brief handshake—light, careful, released almost as soon as it begins. Titles are spoken fully, names handled with care, never rushed. Smiles remain soft, controlled, signaling courtesy rather than intimacy.

Space is carefully negotiated: no one steps too close, no one withdraws too far. Chairs are offered, water poured, the scrape of wood against stone marking the transition from arrival to presence. Only then does conversation begin, often circling weather or travel, as if allowing the room itself to agree that the meeting has properly started.

| Model | verbal_phrase | handshake | cheek_kiss | hug | eye_contact | physical_distance |
|--|---------------|-----------|------------|-----|-------------|-------------------|
| ChatGPT-5, Prompt 3 (Culture-Neutral) | 1 | 1 | 0 | 0 | 1 | 1 |
| ChatGPT-5, Prompt 3 (Ethiopia-Aligned) | 1 | 1 | 0 | 0 | 0 | 0 |
| ChatGPT-5, Prompt 3 (India-Aligned) | 1 | 1 | 0 | 0 | 1 | 1 |

Table 3: Binary feature realization for greeting-related practices under Prompt 3 across prompting conditions.

I Country-Level Alignment Analysis under Prompt Aggregation and Individual Prompts

This appendix reports country-level alignment analyses that complement the frame- and continent-level results in the main text. Model-induced cultural profiles are compared directly against ethnographic gold profiles at the level of individual countries. Results are presented under two settings: (i) analyses based on individual prompts, and (ii) analyses where prompts are aggregated per model. For each setting, we report mean, minimum, and maximum normalized Hamming distances. Mean normalized Hamming distance reflects typical alignment behavior (lower is better). Figure 9 shows mean alignment under culture-neutral prompting for each *model* × *prompt* combination. A consistent pattern emerges: European countries exhibit lower distances (stronger alignment), while countries such as Ethiopia and India show higher distances, indicating weaker alignment. Figure 10 reports mean alignment with prompts aggregated per model, yielding a single score per *Country* × *Model*. The same pattern persists under aggregation, confirming that the observed differences are not driven by prompt variation. Figures 11 and 12 present minimum normalized Hamming distances. A value of 0 indicates perfect alignment between model-induced and ethnographic representations for at least one frame or prompt. Figures 13 and 14 report maximum normalized Hamming distances, capturing worst-case divergence. A value of 1 indicates complete mismatch between model-generated and ethnographic representations.

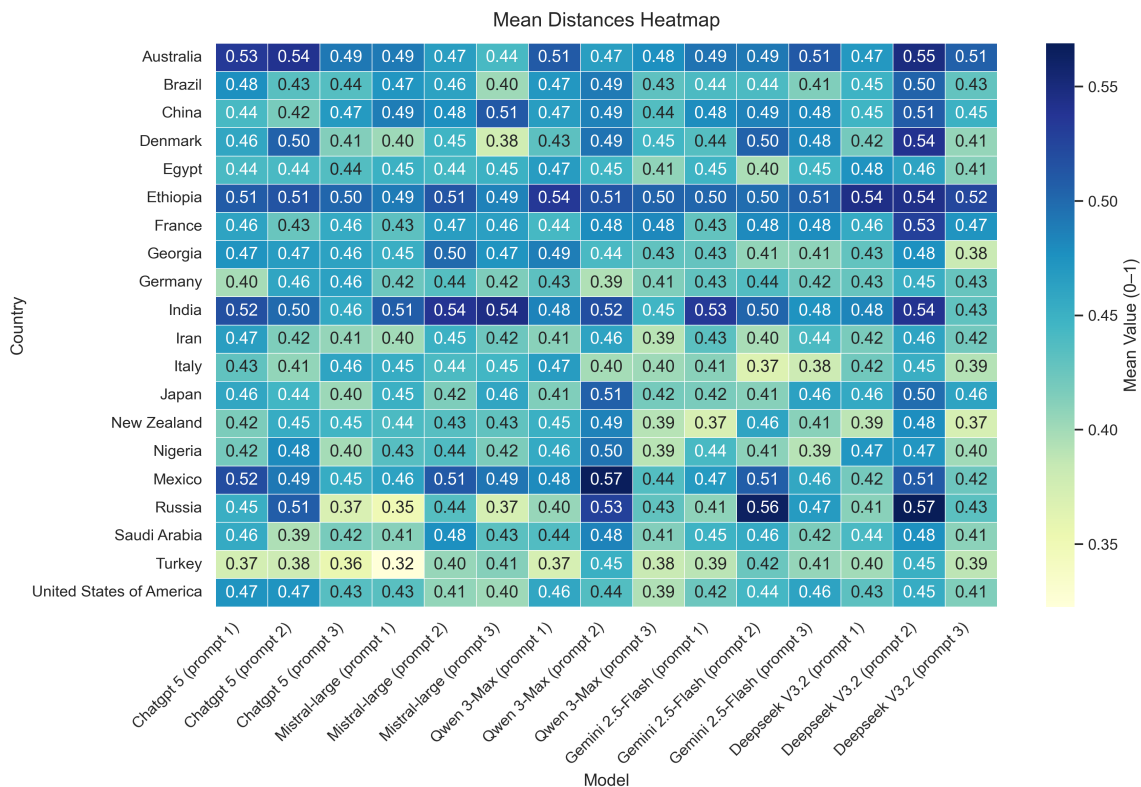


Figure 9: Country-level **mean** normalized Hamming distance for each *model* × *prompt* combination under culture-neutral prompting.

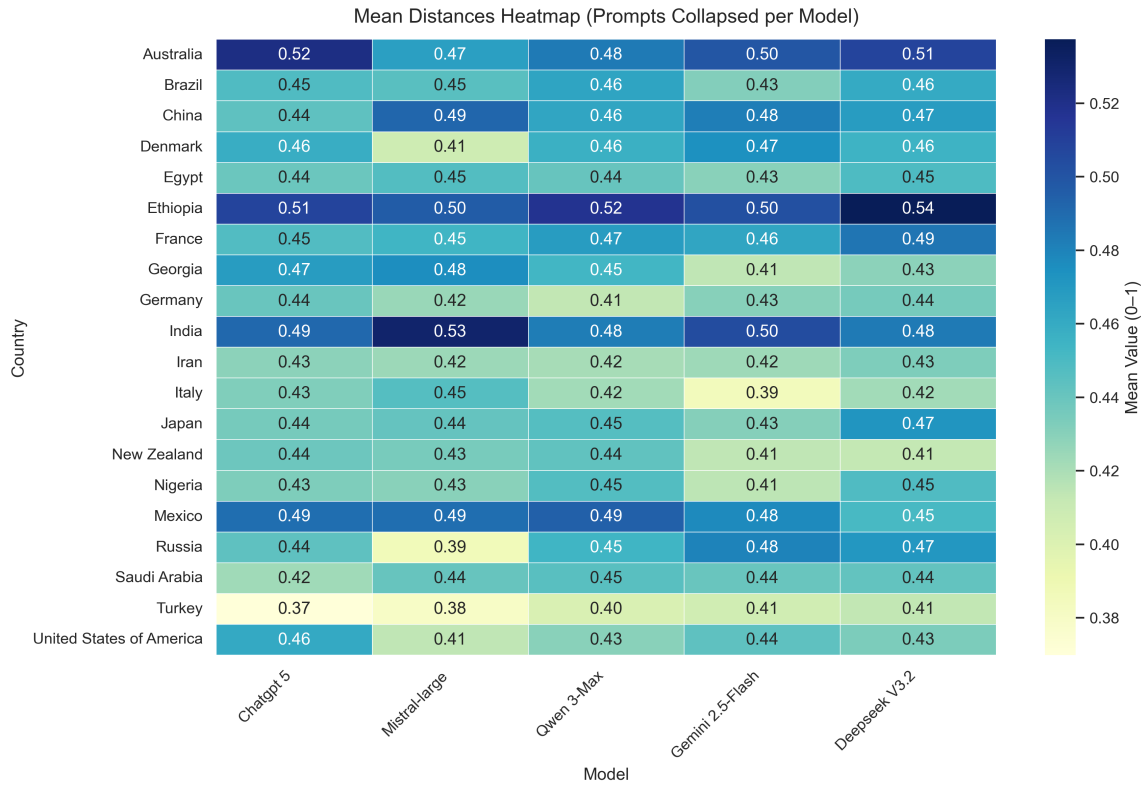


Figure 10: Country-level **mean** normalized Hamming distance per *Country* × *Model*, with frames and prompts collapsed.

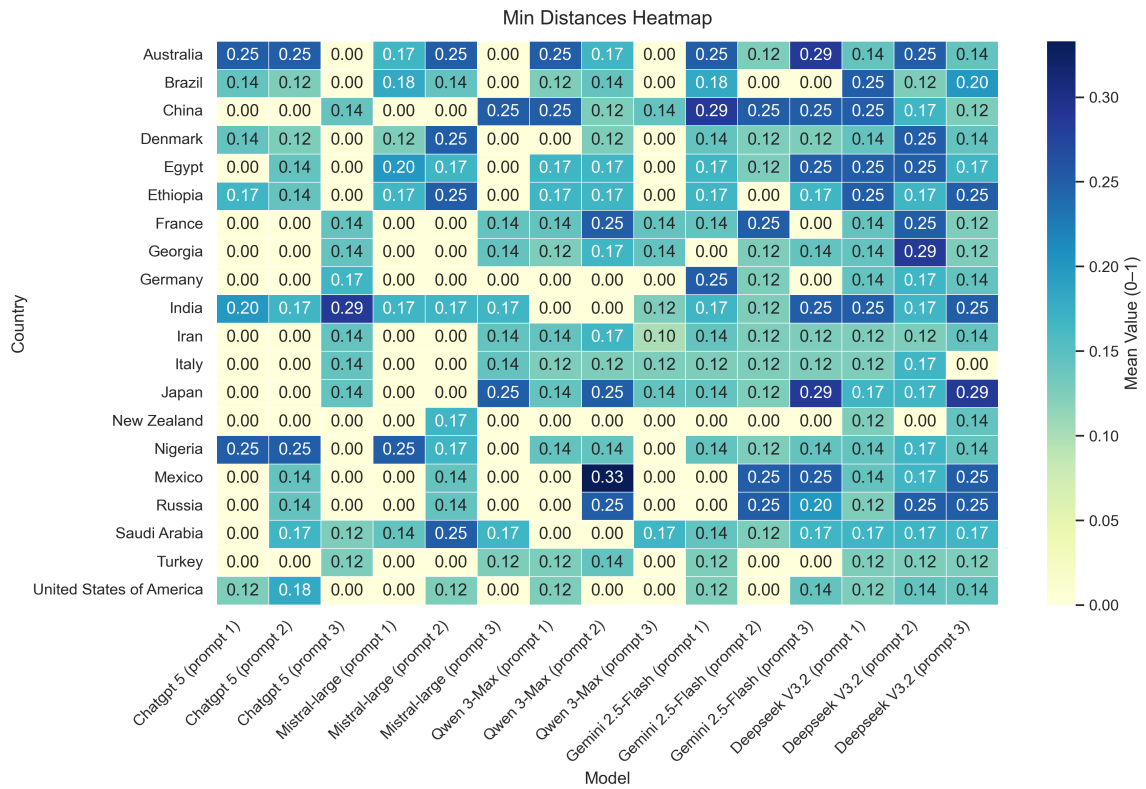


Figure 11: Country-level **minimum** normalized Hamming distance for each *model* × *prompt* combination.

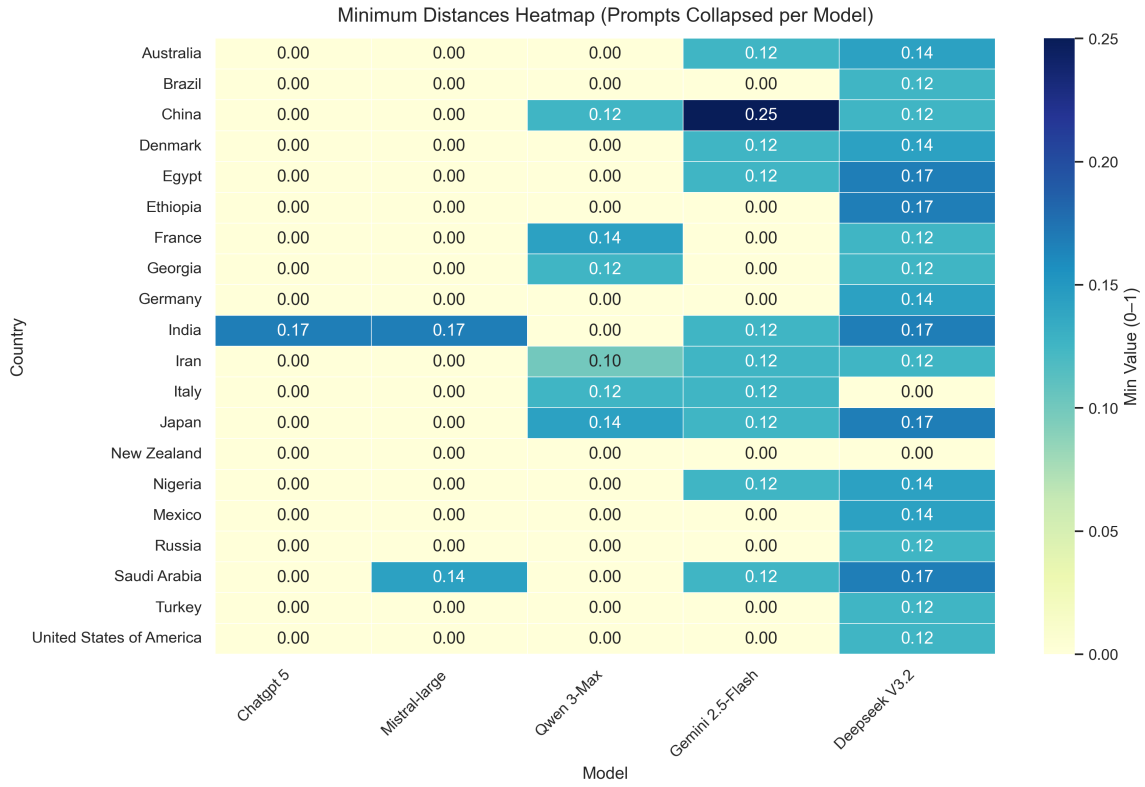


Figure 12: Country-level **minimum** normalized Hamming distance per *Country* × *Model*.

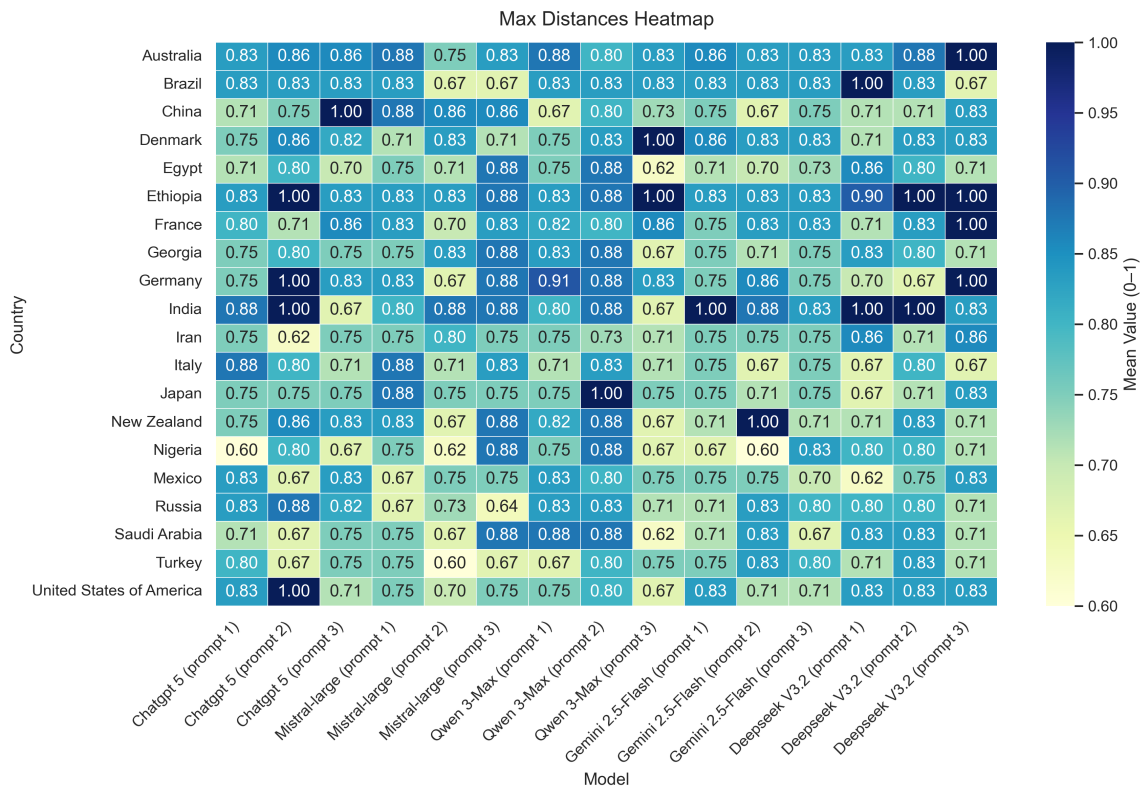


Figure 13: Country-level **maximum** normalized Hamming distance for each *model* × *prompt* combination.

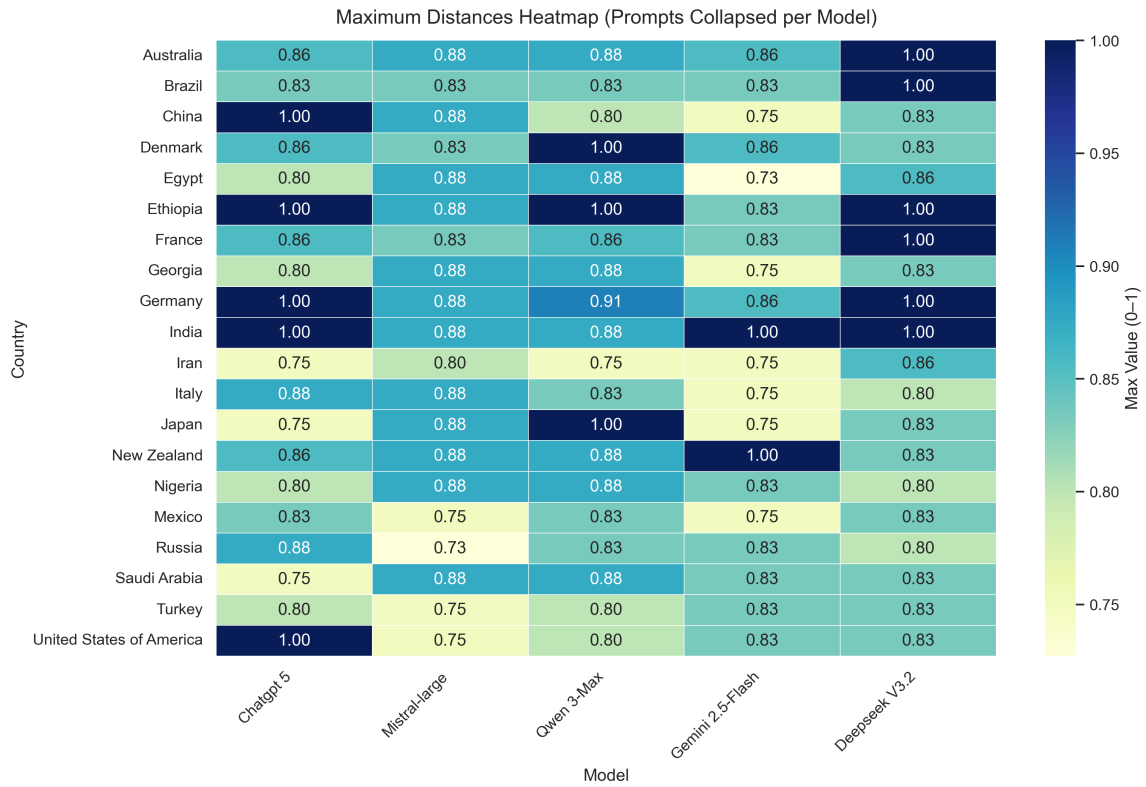


Figure 14: Country-level **maximum** normalized Hamming distance per *Country* \times *Model*.

J Continent-Level Alignment Analysis by Individual Prompts

This appendix presents continent-level alignment analyses computed separately for each prompt. For each prompt template, we report mean, minimum, and maximum normalized Hamming distances between model-induced cultural profiles and continent-level ethnographic gold profiles, computed across all frames and models. These statistics capture typical alignment behavior as well as best- and worst-case divergence for each continent under each prompt, without aggregating across prompts. Countries were assigned to continents using a fixed and consistent mapping. Transcontinental cases were grouped with the continent reflecting their dominant historical-institutional alignment, ensuring analytical consistency and avoiding the introduction of an additional Eurasian category. Figures 15, 16, and 17 report heatmaps of mean, minimum, and maximum distances for each prompt. Mean distances reflect typical alignment behavior (lower is better), minimum distances capture best-case alignment, and maximum distances reflect worst-case divergence. Across prompts, a consistent pattern is observed: European profiles achieve the lowest distances (strongest alignment), while African and Oceanian profiles exhibit higher distances, indicating weaker alignment.

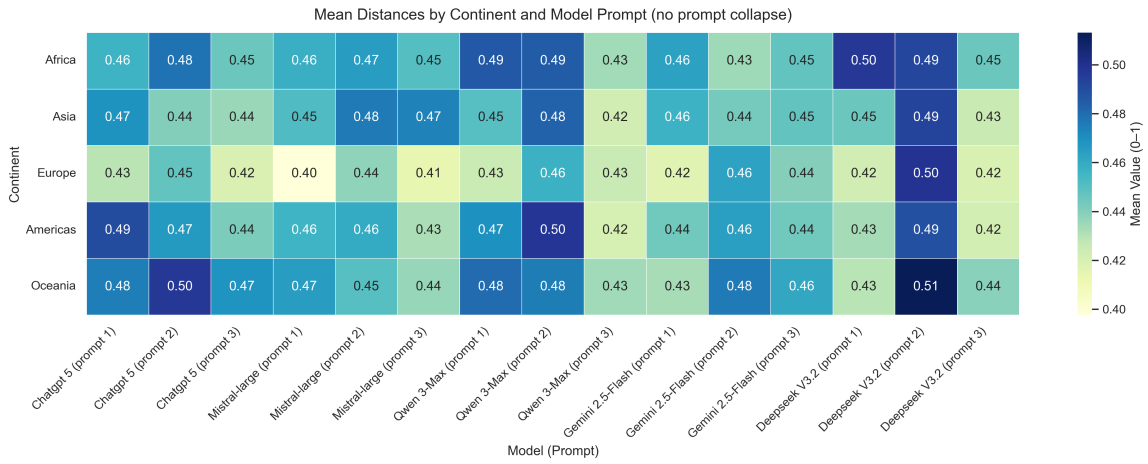


Figure 15: **Mean** normalized Hamming distance per Continent \times Prompt, computed separately for each prompt (not aggregated), and averaged across models and frames (lower is better).

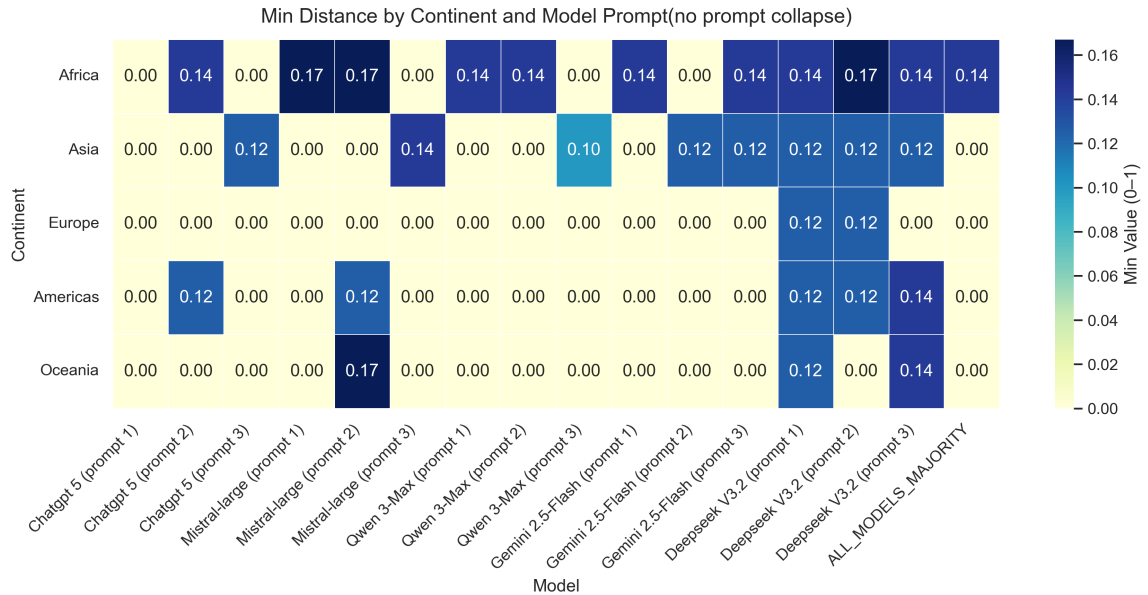


Figure 16: **Minimum** normalized Hamming distance per Continent \times Prompt, computed separately for each prompt (not aggregated), representing best-case alignment across models and frames.

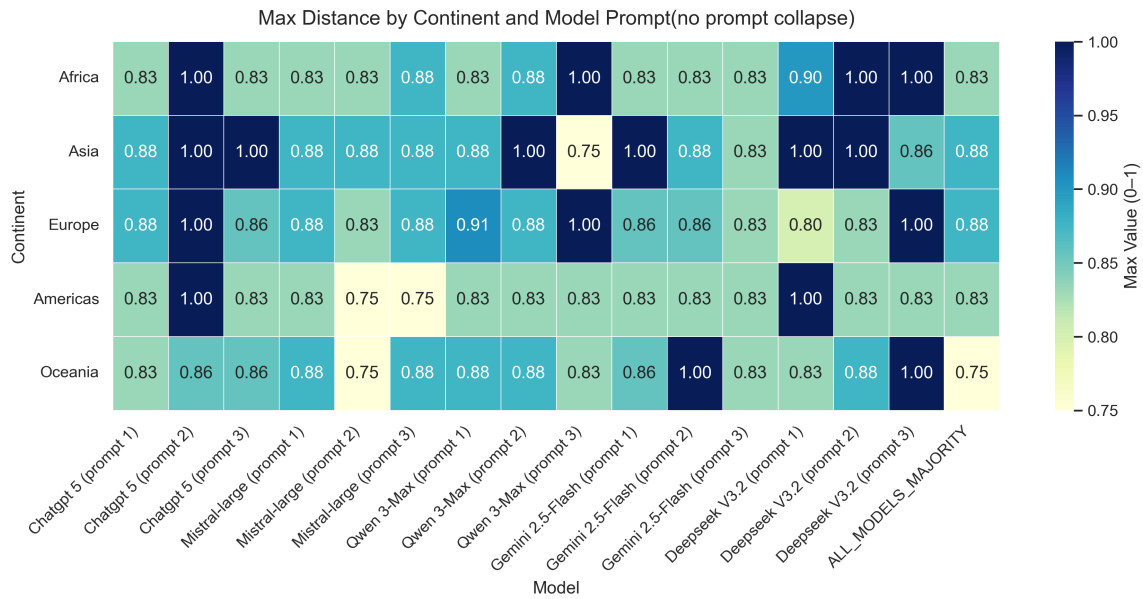


Figure 17: **Maximum** normalized Hamming distance per Continent \times Prompt, computed separately for each prompt (not aggregated), representing worst-case divergence across models and frames.

K Model-Level Continent Comparison with Prompt Aggregation

This appendix provides additional continent-level analyses with prompts aggregated per model, focusing on variability around the mean alignment patterns reported in the main text. Specifically, we report minimum and maximum normalized Hamming distances, together with an error-bar visualization summarizing the range of distances observed across frames. Figures 18 and 19 show heatmaps of minimum and maximum distances per $Model \times Continent$, capturing best-case alignment and worst-case divergence, respectively. Across both views, European profiles consistently achieve the lowest distances, while African and Oceanian profiles exhibit the highest distances, reinforcing the contrast between European profiles and those of Africa and Oceania observed in the mean-distance analysis. Figure 20 presents the corresponding error-bar plot, where markers denote mean distances (reported in the main text) and whiskers indicate the minimum and maximum distances across frames. This visualization confirms that the relative ordering of continents—Europe as closest and Africa and Oceania as most distant—remains stable when accounting for variability across frames.

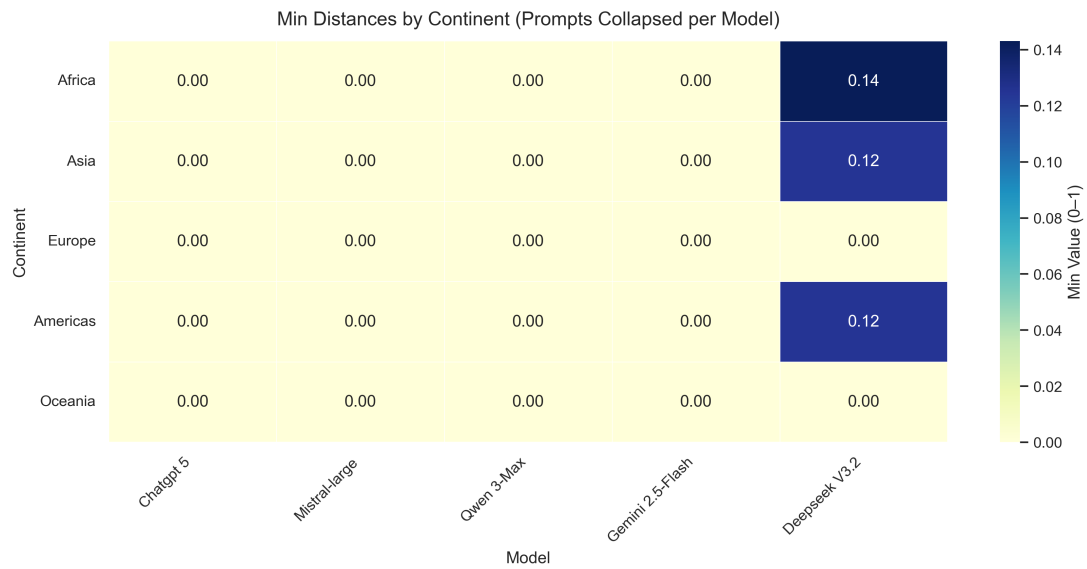


Figure 18: Minimum normalized Hamming distance per $Model \times Continent$ with prompts aggregated, representing best-case alignment across frames (lower is better).

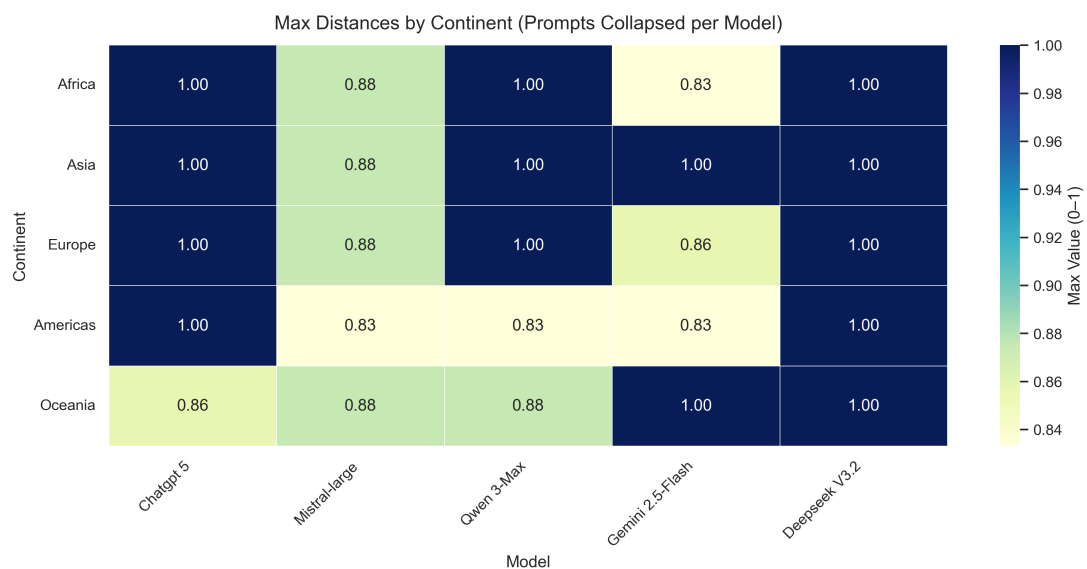


Figure 19: Maximum normalized Hamming distance per $Model \times Continent$ with prompts aggregated, representing worst-case divergence across frames (lower is better).

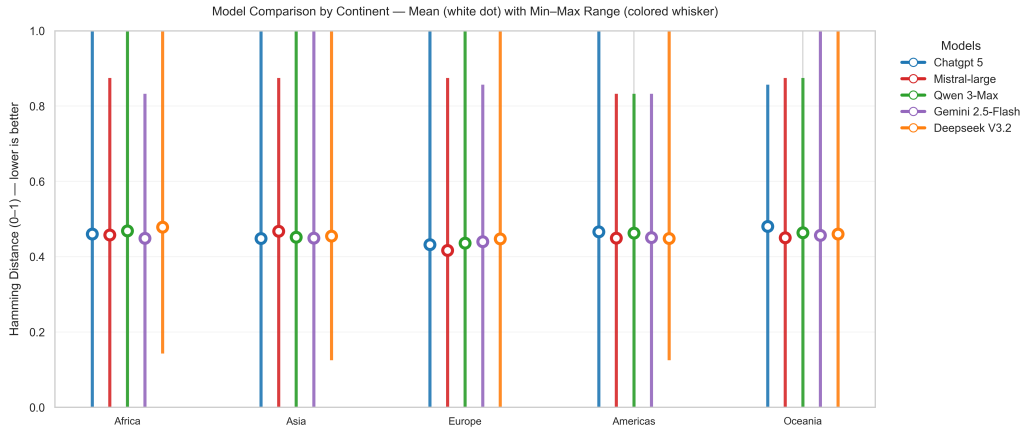


Figure 20: Model comparison by continent with prompts aggregated: mean normalized Hamming distance with minimum and maximum ranges across frames. Mean values are discussed in the main text; whiskers indicate variability.

L Closest Continent per Cultural Frame

This appendix reports detailed frame-level results underlying the argmin analysis presented in the main text. Figure 21 shows the full Continent \times Frame heatmap of mean normalized Hamming distances between model-induced cultural profiles and continent-level ethnographic gold profiles, averaged across all models and prompts. From this heatmap, we identify, for each cultural frame, the continent attaining the minimum mean distance. European profiles achieve the minimum distance in the largest number of frames (7), followed by Asian profiles (5). In contrast, African profiles achieve minimum alignment in only one frame, indicating the largest overall divergence from model-induced cultural representations. The remaining frames are distributed across other continents.

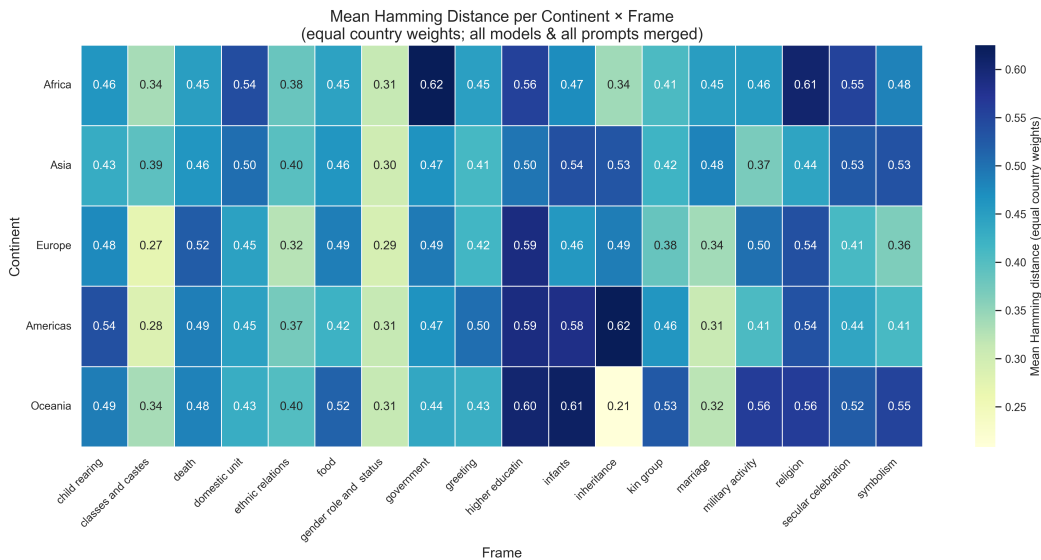


Figure 21: Mean normalized Hamming distance per Continent \times Frame, averaged across all models and prompts (lower is better).

M Country-Level Alignment Shifts Under Cultural Conditioning

This appendix presents country-level alignment under explicit cultural conditioning for the two target cultures, India and Ethiopia. Figure 22 shows the country-level normalized Hamming distances for India under culture-specific prompting. Under this setting, model-induced representations become closely

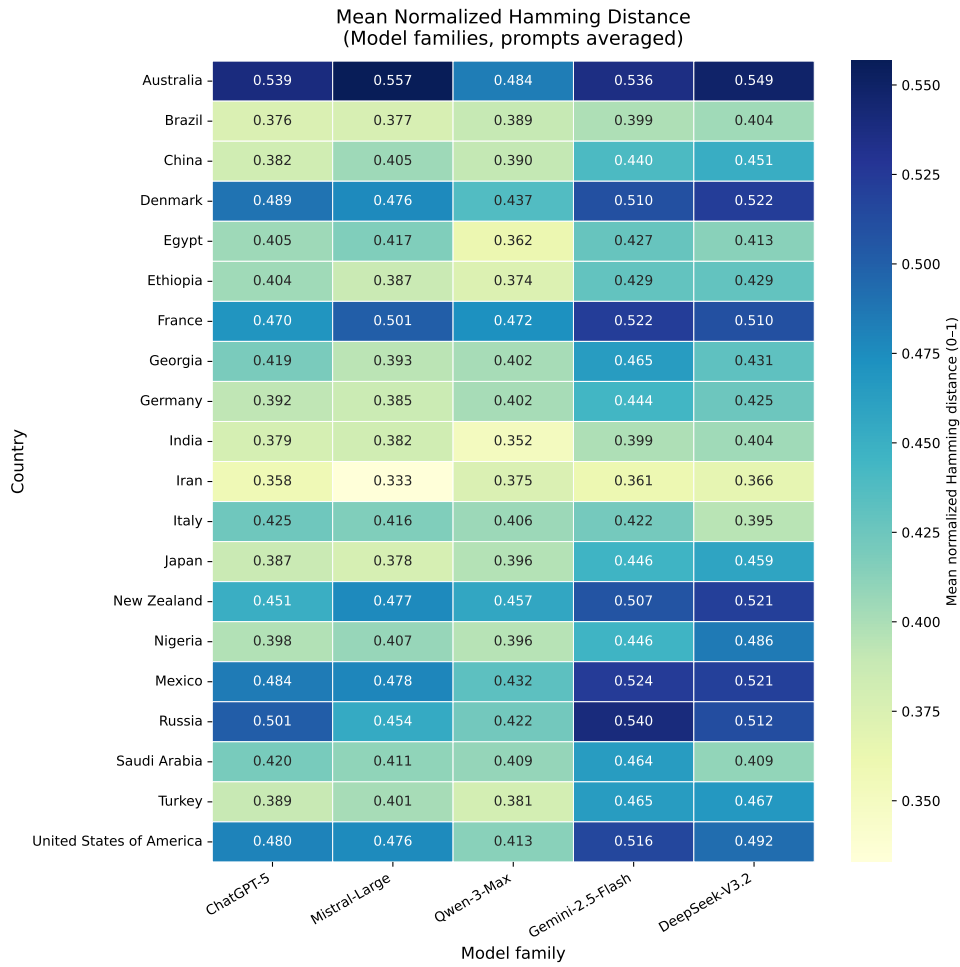


Figure 22: Country-level normalized Hamming distance for India under culture-specific prompting.

aligned with the Indian ethnographic profile across all models. Interestingly, the aligned representations also show proximity to culturally related regions such as Iran, suggesting that models capture broader regional similarities rather than strictly country-specific distinctions under explicit cultural conditioning.

Figure 23 shows the corresponding results for Ethiopia. Under culture-specific prompting, model-induced representations shift toward the Ethiopian ethnographic profile, and the closest alignment is consistently achieved with Ethiopia across all models. These results demonstrate that explicit cultural conditioning can effectively steer model-induced representations toward the intended target cultures at the country level, although some degree of regional smoothing may still persist.

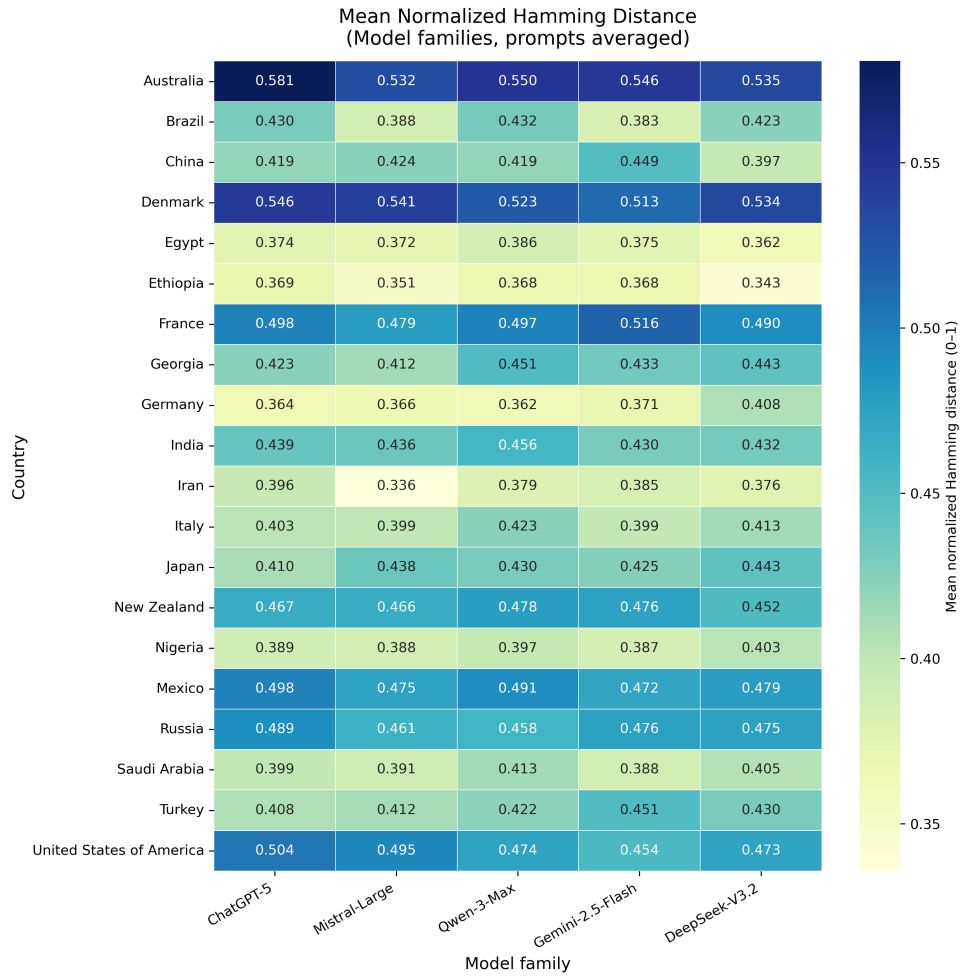


Figure 23: Country-level normalized Hamming distance for Ethiopia under culture-specific prompting.

M.1 Frame-level shifts

At the frame level, we analyze how often each continent achieves the minimum normalized Hamming distance per frame. Significance is evaluated using a binomial test ($N = 18$, baseline $p = 0.2$). Under India-specific prompting, Asia is most frequently selected as the closest continent. As shown in Figure 24, Asia achieves the minimum in 7 out of 18 frames. The corresponding heatmap in Figure 25 provides the exact distance values across all frames. Under Ethiopia-specific prompting, Africa becomes the closest continent in 9 out of 18 frames (Figure 26), with detailed distances shown in Figure 27. These results indicate consistent directional shifts toward the intended regions at the frame level. When evaluated statistically, the shift toward Africa under Ethiopia conditioning is significant ($9/18$, $p = 0.004$), while the shift toward Asia under India conditioning is weaker but still above the random baseline ($7/18$). In contrast, under neutral prompting, Europe is most frequently selected but only marginally above chance ($p = 0.051$), indicating a weaker and less stable bias.

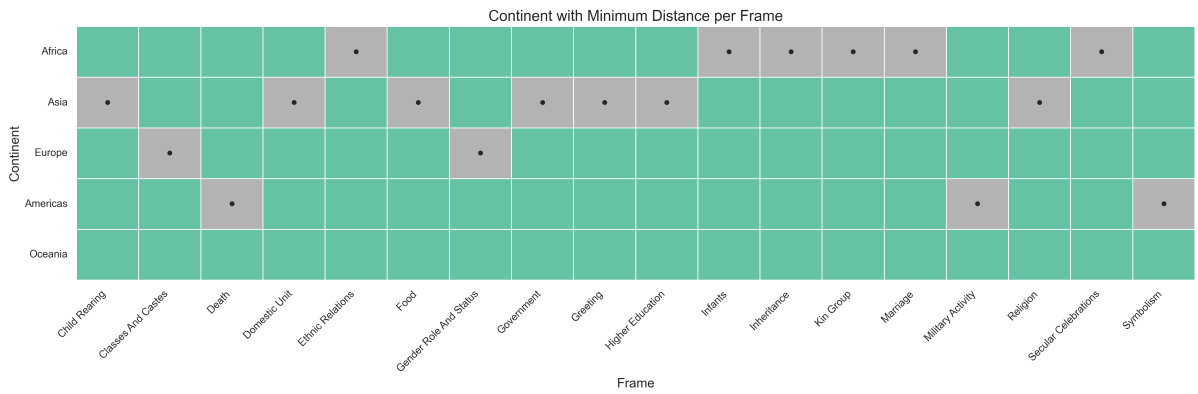


Figure 24: Continent achieving minimum normalized Hamming distance per frame under India-specific prompting (Asia: 7/18 frames).

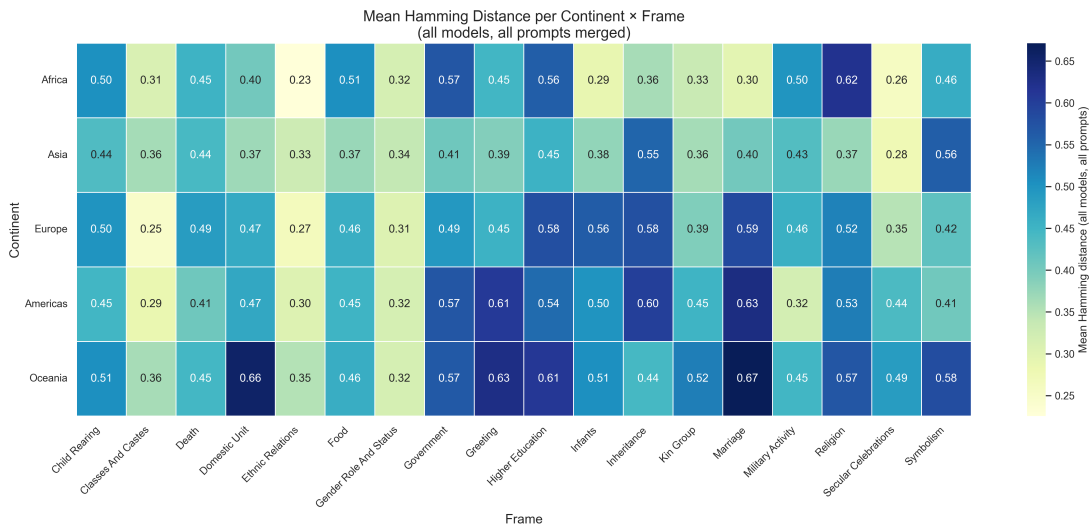


Figure 25: Frame-level normalized Hamming distance heatmap under India-specific prompting.

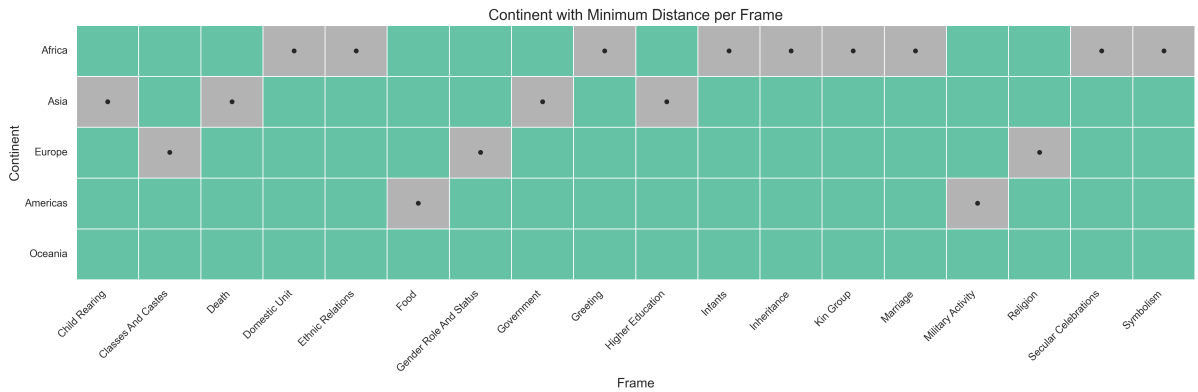


Figure 26: Continent achieving minimum normalized Hamming distance per frame under Ethiopia-specific prompting (Africa: 9/18 frames).

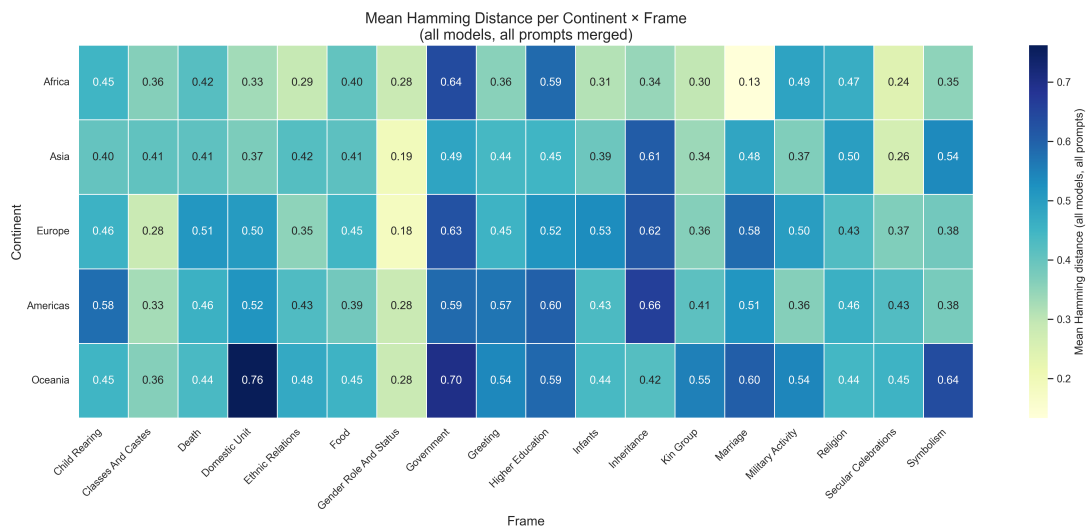


Figure 27: Frame-level normalized Hamming distance heatmap under Ethiopia-specific prompting.

Table 4 reports continent-level summary values under culture-neutral, India-aligned, and Ethiopia-aligned prompting. For each continent and prompting regime, values are computed as the average of the model-specific mean normalized Hamming distances across the five evaluated models, with prompts collapsed and distances averaged over all frames. This aggregation provides a compact numerical comparison between baseline and culturally conditioned generation. Under culture-neutral prompting, Europe achieves the lowest mean distance (0.436), indicating the strongest alignment. Under Ethiopia-aligned prompting, Africa attains the lowest distance (0.377), while under India-aligned prompting, Asia achieves the lowest value (0.404). These results demonstrate that, although models default to European alignment under neutral conditions, explicit cultural conditioning shifts alignment toward the intended target regions.

Table 4: Continent-level summary of normalized Hamming distance under different prompting regimes. Values report the average of model-specific mean normalized Hamming distances computed across the five evaluated models (lower is better).

| Continent | Neutral | Ethiopia-aligned | India-aligned |
|-----------|--------------|------------------|---------------|
| Europe | 0.436 | 0.460 | 0.458 |
| Asia | 0.454 | 0.416 | 0.404 |
| Americas | 0.456 | 0.462 | 0.450 |
| Oceania | 0.462 | 0.501 | 0.507 |
| Africa | 0.464 | 0.377 | 0.412 |

N Distributional Analyses and Prompt Sensitivity

This appendix reports distributional analyses of per-sample normalized Hamming distances to assess variability across individual model generations and sensitivity to prompt formulation, under both culture-neutral and outlier-alignment prompting. Figure 28 shows model-wise distance distributions under culture-neutral prompting with prompts collapsed. All five models exhibit highly similar distributions, indicating broadly comparable behavior. Mistral shows a slightly lower median distance (approximately 0.43), but the difference is small and does not approach perfect alignment. Overall, no model consistently outperforms the others under neutral prompting. Figures 29 and 30 report aggregated (prompt-collapsed) distributions under India- and Ethiopia-alignment prompting. Under India alignment, models again behave very similarly, with DeepSeek exhibiting a marginally lower median, though differences remain minor. Under Ethiopia alignment, the aggregated distributions are nearly identical across all models, indicating effectively uniform behavior when explicit cultural conditioning is applied. Prompt-specific distributions are shown in Figures 31, 32, and 33. Across settings, Prompt 2 consistently exhibits higher distances, particularly under culture-neutral prompting, while this effect is attenuated under Ethiopia alignment. Prompt 2 corresponds to a narrative-opening formulation, which initiates a story without situating it in a specific society. This narrative framing encourages generic, story-driven continuations and introduces additional degrees of freedom unrelated to concrete cultural practices, increasing variability and structural drift. In contrast, scene-based and generalized-pattern prompts more directly elicit descriptions of typical practices, yielding more stable structural representations—especially when explicit cultural alignment is provided.

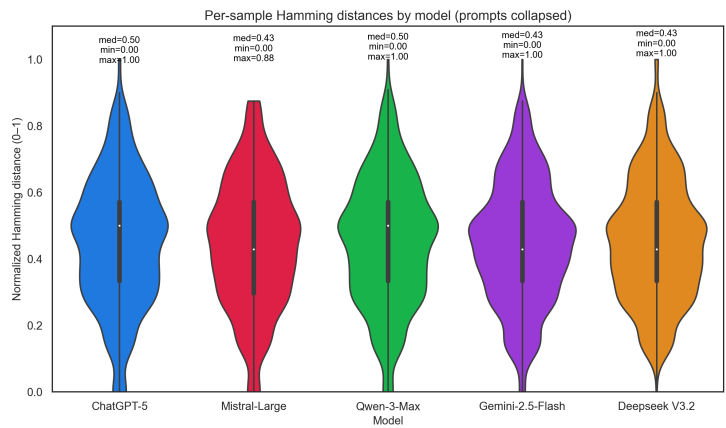


Figure 28: Distribution of per-sample normalized Hamming distances (0–1) for each model under culture-neutral prompting (prompts collapsed).

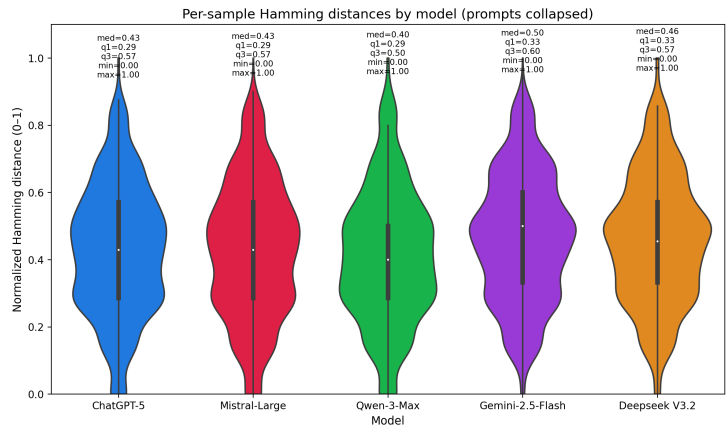


Figure 29: Distribution of per-sample normalized Hamming distances (0–1) for each model under India-alignment prompting (prompts collapsed).

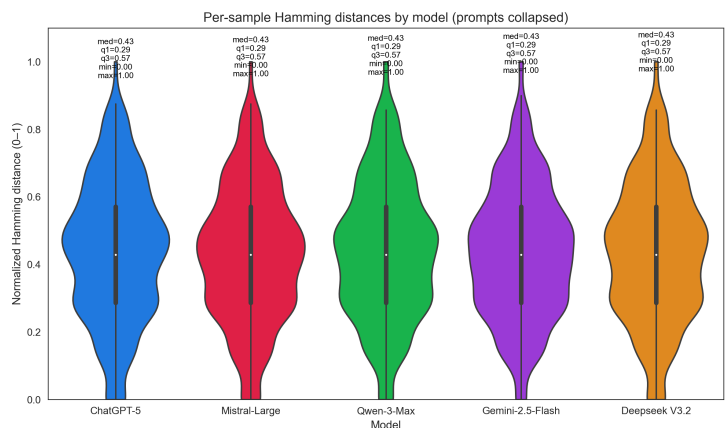


Figure 30: Distribution of per-sample normalized Hamming distances (0–1) for each model under Ethiopia-alignment prompting (prompts collapsed).

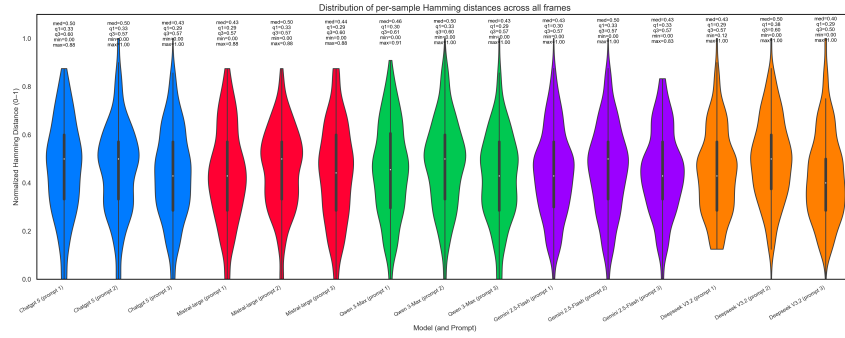


Figure 31: Distribution of per-sample normalized Hamming distances (0–1) for all $model \times prompt$ combinations under culture-neutral prompting.

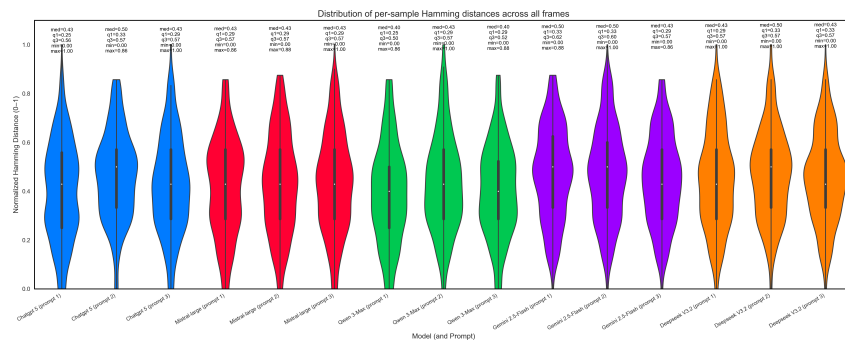


Figure 32: Distribution of per-sample normalized Hamming distances (0–1) for all $model \times prompt$ combinations under India-alignment prompting.

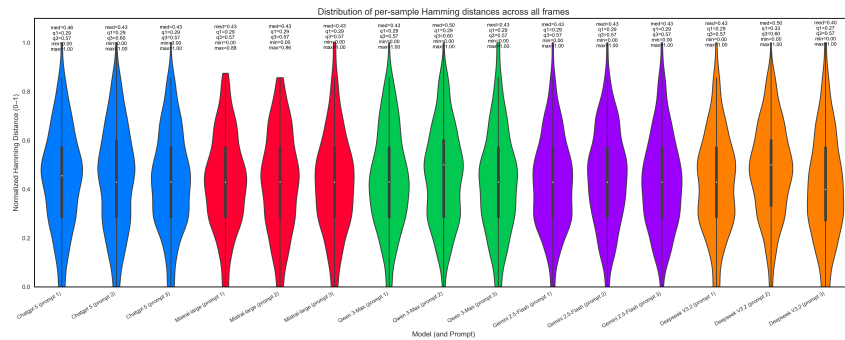


Figure 33: Distribution of per-sample normalized Hamming distances (0–1) for all $model \times prompt$ combinations under Ethiopia-alignment prompting.

O PCA Analysis Under Neutral Culture Across Separate Prompts

Figure 34 shows PCA projections computed separately for each prompt (P1–P3) under neutral culture prompting, with representations averaged over all frames. While projections vary slightly across prompts, model-induced profiles consistently collapse into a compact region of the latent space, remaining closer to European ethnographic profiles than to African and Oceanian regions. These patterns are consistent with the homogenization trend observed in the aggregated analysis in the main text.

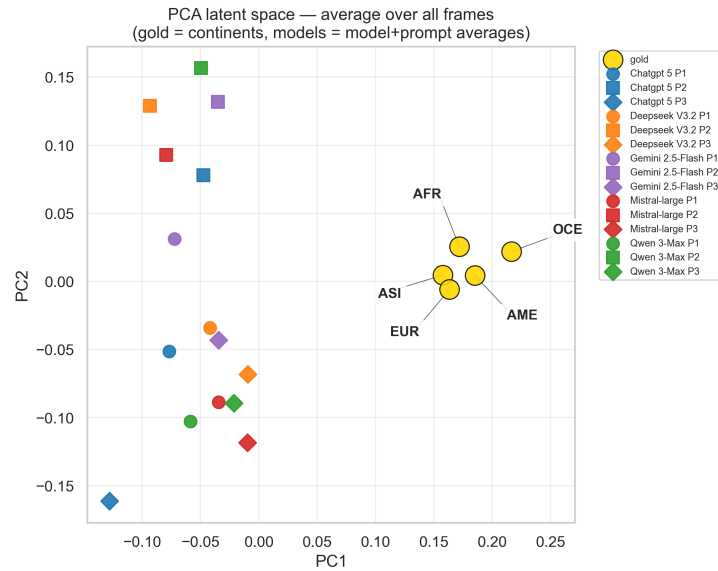


Figure 34: PCA latent space across separate prompts (P1–P3) under neutral culture prompting, averaged over all frames.

P PCA Analysis for India Alignment Across Separate Prompts

To complement the main-text analysis, we report PCA projections for India-alignment prompting in Figure 35. For each $model \times prompt$ combination, representations generated under India-alignment prompting are averaged across all 18 frames, yielding one point per model and prompt. These representations are shown separately for each prompt and projected together with continent-level ethnographic gold profiles. As shown in Figure 35, relative to the neutral-culture baseline, India-aligned model profiles shift away from the European-centered region and move closest to Asian gold clusters, followed by African clusters. While the model-induced profiles do not overlap with the continent-level gold points, the observed geometric shift indicates a consistent directional effect of explicit India-alignment prompting across models and prompts.

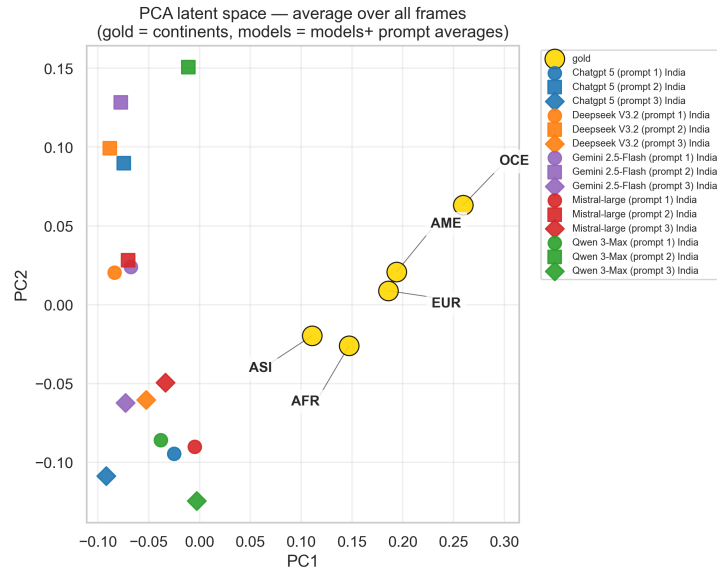


Figure 35: PCA latent space under India-alignment prompting, with separate prompts shown and representations averaged over all 18 frames, compared to continent-level ethnographic gold profiles.

P.1 PCA for Ethiopia and India with Aggregated prompts

Figures 36 and 37 report PCA projections for Ethiopia and India with prompts aggregated. For each country, model-induced profiles generated under outlier-alignment prompting are averaged across prompts and projected together with continent-level ethnographic gold profiles using the same PCA procedure as in the main text.

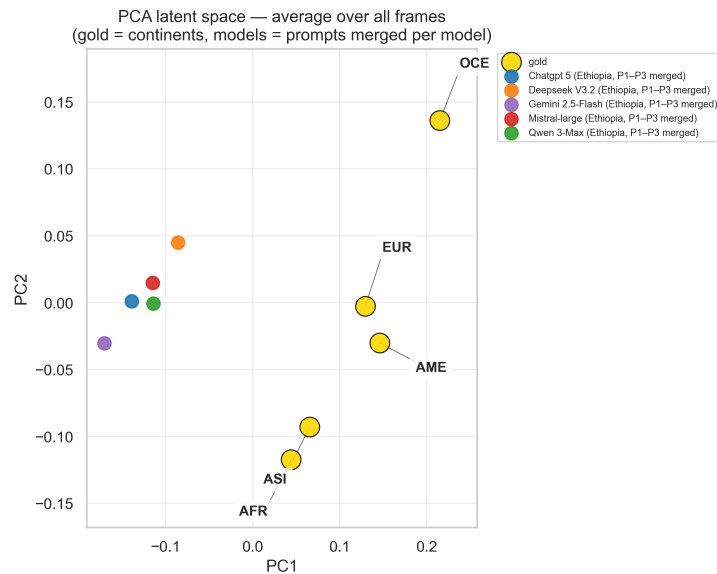


Figure 36: PCA projection for Ethiopia alignment with prompts aggregated, compared to continent-level ethnographic gold profiles.

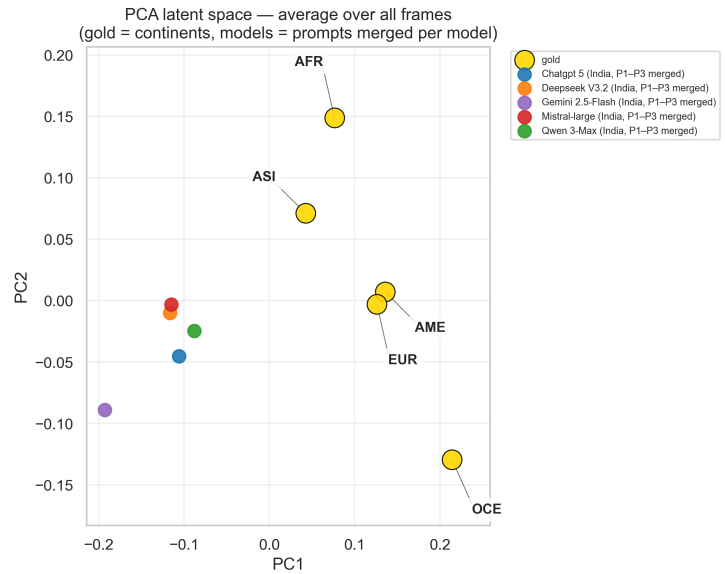


Figure 37: PCA projection for India alignment with prompts aggregated, compared to continent-level ethnographic gold profiles.

Separate vs. Aggregated PCA Projections Differences between projections computed separately per prompting condition and those computed after prompt aggregation arise from the order of aggregation. PCA is sensitive to the covariance structure of the data, which changes when representations are averaged across prompts prior to dimensionality reduction. Projections computed separately preserve within-model variation induced by different prompts, occasionally revealing proximity to target cultural regions (e.g., Africa for Ethiopia), whereas aggregated projections reflect the dominant central tendency after prompt-induced variation is reduced. Despite these differences, the overall alignment tendencies remain consistent: under neutral-culture prompting, model-induced profiles are closest to European regions, with the Americas typically appearing nearby, while African and Oceanian regions remain more distant. This relative proximity between Europe and the Americas persists even under culture-aligned prompting. Under Ethiopia- and India-alignment prompting, representations shift toward African and Asian regions, respectively, but without disrupting the broader continental geometry. Although the exact positioning varies across prompts and continents, these directional patterns are preserved across both separate and aggregated projections. We emphasize that PCA provides a low-dimensional visualization of high-dimensional relationships, and thus these geometric patterns should be interpreted as qualitative trends rather than exact quantitative measurements.

Q Full Dialogue Prompts, Example Dialogues, Annotation Procedures and examples

Q.1 Prompt Templates

We provide an example of the prompts used for the *Food* frame. All other frames follow the same structure with frame-specific scenarios and features.

Frame (culture-neutral) Prompt:

Generate a natural dialogue.

Scenario:

Two people are eating a regular weekday meal at home and chatting about the daily food.

Instructions:

- Exactly 6 turns (3 per person)
- Max 15 words per turn

- The conversation should sound natural and human, like everyday life

Output format:

- Write the dialogue in English
- Use exactly 6 turns
- Alternate speakers:
Speaker A:
Speaker B:
Speaker A:
Speaker B:
Speaker A:
Speaker B:
- Each turn must be on its own line

Frame+Culture (example: India) Prompt:

Generate a natural dialogue.

Scenario:

Two people are eating a regular weekday meal at home and chatting about the daily food.

Instructions:

- Exactly 6 turns (3 per person)
- Max 15 words per turn
- The conversation should reflect Indian culture
- Do NOT explicitly mention the country or culture name (e.g., "India", "Indian")
- Sound natural and human, like everyday life

Output format:

- Same as above

Frame+Culture+Structure (example: India) Prompt:

Generate a natural dialogue.

Scenario:

Two people are eating a regular weekday meal at home and chatting about the daily food.

Food frame structure:

Rice
Potato
Stewing
Boiling
Frying
Garlic
Fork
Spoon
Knife
Cheese
Alcohol drink

Instructions:

- Exactly 6 turns (3 per person)
- Max 15 words per turn

- The conversation should reflect Indian culture
- Do NOT explicitly mention the country or culture name
- It is NOT necessary to use all elements above; only include those that fit naturally
- Mention food items and actions naturally in conversation
- Do NOT refer to elements as a list or feature set
- Avoid vague expressions; keep the dialogue concrete and clear
- Sound natural and human

Output format:

- Same as above

Frame+Structure+Binarized (baseline) Prompt:

Generate a natural dialogue.

Scenario:

Two people are eating a regular weekday meal at home and chatting about the daily food.

Food frame features (binary):

Rice = 1
 Potato = 0
 Stewing = 0
 Boiling = 0
 Frying = 0
 Garlic = 0
 Fork = 0
 Spoon = 0
 Knife = 0
 Cheese = 0
 Alcohol drink = 0

Hard rules:

- Features with value 1 MUST appear naturally
- Features with value 0 MUST NOT appear or be implied
- Do NOT refer to features explicitly
- Mention food naturally in everyday conversation
- Avoid vague placeholders

Instructions:

- Exactly 6 turns (3 per person)
- Max 15 words per turn
- Sound natural and human

Output format:

- Same as above

Q.2 Annotation Guidelines

Annotators for Ethiopia were recruited via the Prolific platform (<https://www.prolific.com>) and were native speakers of the target culture, fluent in English. Annotators for Iran were two native Iranian NLP researchers, fluent in English, who conducted the annotation voluntarily and without financial compensation. The annotation was conducted through an online interface, where dialogues were presented in randomized order. Annotators were not informed about the prompting condition or the intended target

culture of each dialogue. Each annotator was asked to evaluate how confidently a given dialogue reflects their own culture, based on the following four-point scale:

- **A — Very sure:** The dialogue unmistakably reflects my country's culture.
- **B — Fairly sure:** The dialogue mostly fits my country, although some elements could apply to other cultures.
- **C — Not sure:** The dialogue is culturally ambiguous and could belong to multiple countries.
- **D — Not mine:** The dialogue clearly does not reflect my country's culture and appears mismatched or inappropriate.

Annotators were instructed to rely on their cultural intuition and everyday knowledge, rather than over-analyzing isolated details. Each dialogue was rated independently. For Ethiopia, four annotators participated, each annotating 225 dialogues. A subset of 100 dialogues was shared among annotators to measure inter-annotator agreement. For Iran, two annotators participated, each annotating 450 dialogues, with 100 shared dialogues used for agreement analysis. The presentation order of dialogues was fully randomized to avoid ordering effects. The full annotation interface and instructions are included for reproducibility.

Q.3 Examples of Generated Dialogues with Corresponding Prompts and Ratings

Table 5 shows the full prompts used for the *Food* frame (Iran), while Table 6 presents example generated dialogues and corresponding human ratings. Annotator judgments reflect the degree of culturally specific signals present in each dialogue. The *Frame+Culture* example receives a **C** rating, as it includes culturally plausible elements such as rice and drinking tea after meals, which are common in Iranian culture but not uniquely specific, also appearing in other cultural contexts (e.g., South Asia). The *Frame+Culture+Structure* example is rated **A**, as it includes highly culture-specific elements such as *khoresh*, which is strongly associated with Iranian cuisine. The presence of such distinctive cultural markers, combined with coherent usage, makes the dialogue clearly identifiable as Iranian. In contrast, the *Frame+Structure+Binarized* example receives a **D** rating, as the generated dialogue exhibits clear cultural inconsistencies. Although the model includes required features, their combination does not reflect authentic Iranian practices. For example, rice is typically steamed rather than boiled in Iranian cuisine, and wine is not commonly consumed in everyday meals. These mismatches indicate that the model is mechanically satisfying feature constraints without capturing the underlying cultural structure, resulting in an output that is perceived as culturally incorrect.

Table 5: Full prompts for the *Food* frame (Iran).

| Prompt Type | Full Prompt |
|---|---|
| Frame+Culture (Iran) | <p>Generate a natural dialogue.</p> <p>Scenario: Two people are eating a regular weekday meal at home and chatting about the daily food.</p> <p>Instructions: - Exactly 6 turns (3 per person) - Max 15 words per turn - Reflect Iranian culture - Do NOT mention the country explicitly - Sound natural and human</p> <p>Output format: - Write in English - 6 alternating turns (A/B) - Each turn on its own line</p> |
| Frame+Culture+Structure (Iran) | <p>Generate a natural dialogue.</p> <p>Scenario: Two people are eating a regular weekday meal at home.</p> <p>Food frame structure: Rice, Potato, Stewing, Boiling, Frying, Garlic, Fork, Spoon, Knife, Cheese, Alcohol drink</p> <p>Instructions: - Exactly 6 turns (3 per person) - Reflect Iranian culture - Do NOT mention country name - Use only elements that naturally fit - Avoid vague or abstract references - Keep dialogue natural</p> <p>Output format: Same as above</p> |
| Frame+Structure+Binarized (Iran) | <p>Generate a natural dialogue.</p> <p>Scenario: Two people are eating a regular weekday meal at home.</p> <p>Food frame features (binary): Rice=1, Potato=0, Stewing=1, Boiling=1, Frying=1, Garlic=0, Fork=1, Spoon=0, Knife=1, Cheese=1, Alcohol=1</p> <p>Hard rules: - Include features with value 1 - Exclude features with value 0 - Do NOT mention rules explicitly - Keep dialogue natural</p> <p>Instructions: - Exactly 6 turns - Max 15 words per turn</p> <p>Output format: Same as above</p> |

Table 6 presents example generated dialogues and corresponding human ratings.

| Prompt Type | Generated Dialogue (ChatGPT-5.2) | Rating |
|---|---|--------|
| Frame+Culture (Iran) | <p>Speaker A: The rice smells amazing, reminds me of dinners back home.</p> <p>Speaker B: Yeah, simple food after work feels comforting.</p> <p>Speaker A: My mom always said weekday meals should be warm and filling.</p> <p>Speaker B: True, nothing fancy, just something that brings everyone together.</p> <p>Speaker A: Let's have some tea after, it helps unwind.</p> <p>Speaker B: Good idea, evenings feel calmer that way.</p> | C |
| Frame+Culture+Structure (Iran) | <p>Speaker A: The rice smells great; did you stew the chicken this morning?</p> <p>Speaker B: Yes, with turmeric and onions; I fried garlic earlier.</p> <p>Speaker A: I'll grab a spoon; this khoresh needs soaking with rice.</p> <p>Speaker B: Careful, the potatoes boiled fast and soften the sauce.</p> <p>Speaker A: Weeknights feel better with homemade food like this.</p> <p>Speaker B: After dinner, let's have tea and cheese, like always.</p> | A |
| Frame+Structure+Binarized (Iran) | <p>Speaker A: The stewed chicken with boiled rice smells comforting after work.</p> <p>Speaker B: I fried onions earlier; the sauce tastes richer with cheese.</p> <p>Speaker A: Pass a fork and knife; the rice pairs well with the stew.</p> <p>Speaker B: A glass of wine helps unwind during a weekday dinner.</p> <p>Speaker A: I like how boiling softened the rice without losing texture.</p> <p>Speaker B: Frying finished the stew nicely; cheese melts beautifully on top.</p> | D |

Table 6: Example generated dialogues and human ratings for the *Food* frame (Iran).