Not What, But How: A Communicative Audit of LLM Response Framing

Anonymous ACL submission

Abstract

Large language models (LLMs) are being increasingly used to answer subjective, information-seeking questions, where users are sensitive to how responses are communicated, not just whether the answers are correct. Existing LLM evaluations for subjective cultural queries largely focus on factual correctness, ignoring *how* the response is framed. To this end, we introduce `FRANZ`, an automated `FRAmework for respoNse characteriZation` to conduct communicative audit of LLM responses along four dimensions: cultural positioning, use of generalizing language, anthropomorphic cues, and adherence to conversational maxims. To enable this evaluation, we contribute `SQUARE` - a corpus of 376k subjective questions sourced from 57 subreddits, and mapped to 7 countries and 19 question categories. We demonstrate `FRANZ`'s applicability by scoring responses from three open-weight LLMs. We observe that LLMs show statistically significant differences in the frequency with which they employ each response characteristic. Unlike single-dimensional audits, `FRANZ` reveals that insider positioning and anthropomorphism are positively coupled, with the degree of coupling varying by country, providing a diagnostic lens for identifying framing divergences.

1 Introduction

For subjective queries, variations in how LLM responses are framed and communicated critically shape how users interpret and act on model outputs (Kumar et al., 2026; Cohn et al., 2024; Wan et al., 2026; Kuang and Bicchieri, 2024). An underexplored dimension of this framing is Cultural Positioning Bias (CPB) (Wan et al., 2026; Tao et al., 2024), which refers to the tendency to adopt insider positioning for some communities, while defaulting to generic outsider framings for others.

While existing work analyzes CPB in isolation (Xu et al., 2025), **insider positioning** involves adopting a group's communicative identity, leading it to interact with a richer set of response characteristics. Without insider context, LLMs may compensate through broad-stroke **generalizing language**, risking stereotyping (Beukeboom, 2025; Davani et al., 2024). Insider positioning also relies on interpersonal **anthropomorphic expressions**. Despite extensive studies on anthropomorphism in LLMs (Ibrahim et al., 2026a; Maeda and Quan-Haase, 2024), how it varies across cultures remains underexplored. Complementing these, **Gricean maxims** (Grice, 1975) offer a distinct lens on communication effectiveness. The maxims assess whether a response is informative, concise, relevant, and structured (Keenan, 1976; Wierzbicka, 1991). *While these four response characteristics have been studied in isolation, their systematic and unified evaluation is unexamined.* To this end, we design `FRANZ` (`FRAmework for respoNse characteriZation`): a sociolinguistically grounded framework that operationalizes four distinct yet interrelated facets of response framing. This shifts the focus from *what* LLM outputs encode to *how* they are framed (Reinhart et al., 2025).

Cultural benchmarks that evaluate cloze-form, single-ground-truth factual knowledge (Kabir et al., 2025), lack the open-ended queries needed to study response framing. As users from diverse backgrounds do not ask the same question, it is difficult to assess how LLM responses are framed for culturally situated queries. We therefore turn to Reddit, where communities host diverse discussions (Antoniak et al., 2024; Oddný et al., 2023) and members draw on lived experiences and shared norms that outsiders may lack (Xu et al., 2025). We construct `SQUARE`, a corpus of 376k `Subjective QUeries Across REddit` that we map to 7 countries and 19 culturally[1] grounded question categories. We employ

[1]We adopt country-level subreddits as a proxy for culture.

FRANZ to evaluate responses for SQUARE generated by `Llama-3.1-8B-it` (Llama), `Gemma-3-12b-it` (Gemma), and `Ministral-3-14B-it` (Mistral). The setup is summarised in Figure 1.

Our analysis reveals the following: **LLMs exhibit model-specific signatures**. Upon examining each response characteristic independently (Sections 6.1 & 6.3) – Mistral shows the strongest tendency to employ insider positioning as well as generalizing language; Gemma displays the best adherence to maxims; and Llama appears most constrained. These rankings persist across country-category subsets we examine, pointing to systematic model-level tendencies. Moreover, **response characteristics co-occur selectively** (Sections 6.2 & 6.3). For example, an empathetic tone is more likely to be detected when a response is insider-oriented, but this tendency varies within countries with no relative ranking among LLMs. Our results highlight systematic differences among LLMs (Section 7), leading to uneven user experience based on the model they engage with. For developers and auditors, FRANZ provides a diagnostic lens to identify when response framing diverges via a multidimensional evaluation[2].

2 Related Work

Cultural evaluation of LLMs Cultural awareness of LLMs has been widely studied (Adilazuarda et al., 2024; Pawar et al., 2025a) through commonsense (Myung et al., 2024; Chiu et al., 2025; Wang et al., 2024b), value surveys (Dwivedi et al., 2023; Zhao et al., 2024; Kabir et al., 2025), demographic context (Wang et al., 2024a; Li et al., 2024b; Lovenia et al., 2024; Thellmann et al., 2024), etc. These benchmark datasets and evaluation frameworks (Shi et al., 2024; Myung et al., 2024; Kumar et al., 2025), judge whether an LLM surfaces correct cultural knowledge against a single ground truth, leaving aside *how* the answer is framed. In subjective settings, the response framing matters as much as the content itself (Reinhart et al., 2025). We complement this content-focused tradition by characterizing response framing.

Anthropomorphism in LLMs The presence of anthropomorphic language signaling agency, emotion, empathy, and personhood has become pervasive (Jones et al., 2025). Extensive work has cataloged anthropomorphic cues in LLM responses (Bhattacharyya, 2015; Ibrahim et al., 2026a,b; Cheng et al., 2025). Research also documents how anthropomorphic framing affects users' trust, persuasion (Cohn et al., 2024; Maeda and Quan-Haase, 2024; Cheng et al., 2025, 2026), and perception (Xiao et al., 2025). Critically, though the presence of anthromorphism is culturally moderated (Ge et al., 2024; Schimmelpfennig et al., 2026), existing work rarely captures variation across culture. Our work bridges this by analyzing anthropomorphic cues across seven countries and nineteen culturally grounded categories.

Communicative framing: positioning, generalizing, maxims. Building on the anthropological tradition of *emics* vs *etics* (Headland et al., 1990), Wan et al. (2026) operationalized cultural positioning as the extent to which an LLM adopts an insider stance (inciting language that signals membership in the referenced community). Complementary work examines how LLM-mediated knowledge production is adapted by users of insider and outsider groups (Xu et al., 2025). Persona-steering similarly surfaces variations when LLMs are asked to inhabit specific demographic stances (Liu et al., 2024; Chen et al., 2025; Pawar et al., 2025b).

Meanwhile, generalizing language captures whether responses make group-level claims about individuals. Their examination requires dedicated annotation resources (Davani et al., 2024) and linguistic-theoretic grounding (Beukeboom, 2025). Here, conversational maxims (Grice, 1975) supply a cooperative-quality lens that has been operationalized for dialogue (Krause and Vossen, 2024). But cross-cultural pragmatics has long shown that maxim adherence is calibrated under shared assumptions (Keenan, 1976; Wierzbicka, 1991; Sobhani and Saghebi, 2014). While these dimensions have largely evolved in parallel, FRANZ unifies them into four interacting facets of response framing and measures their co-occurrence at scale.

3 Framework

FRANZ (FRAmework for respoNse characteriZation) annotates an LLM-generated response along the four dimensions in Figure 1. We arrive at the four by tracing how *cultural positioning* (Wan et al., 2026) surfaces in text. The anchoring characteristic is **insider positioning**, which directly encodes the perspective through which the response is framed. In the absence of insider positioning, models may compensate with high-level **generalizing language** (Beukeboom,

[2]Code and data will be made public upon acceptance.

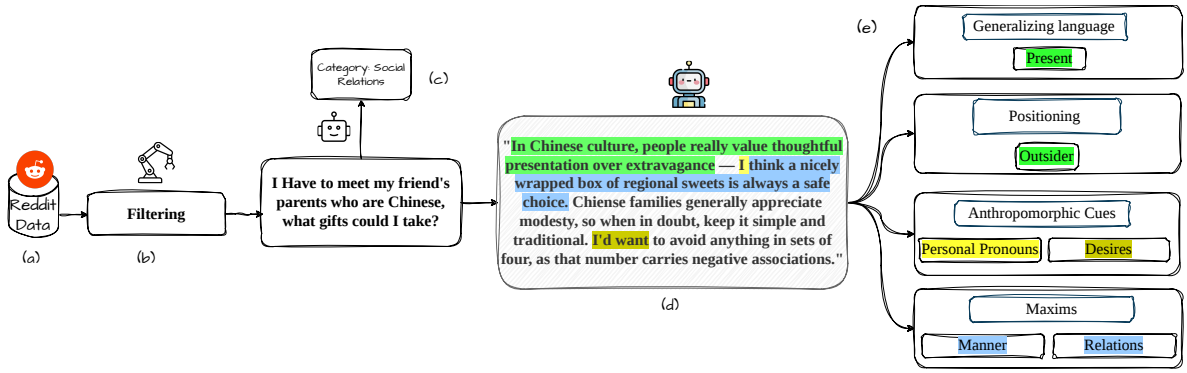


Figure 1: **Overview of the pipeline.** LLMs generate responses to subjective questions in SQUARE (a–c; Section 4); FRANZ then scores each response along four characteristics viz. cultural positioning, generalizing language, anthropomorphism, and maxim adherence (d–e; Section 3).

2025; Davani et al., 2024). Conversely, when an insider stance surfaces, it is typically enacted through interpersonal **anthropomorphic phrasing** such as agency, emotion, and relational stance (Ibrahim et al., 2026a; Maeda and Quan-Haase, 2024). Together, these three aspects trace, at a low level, how positioning is inhabited or evaded in text. To complement these, we incorporate a broader dimension of response framing through **conversational maxims** (Grice, 1975; Krause and Vossen, 2024). Maxims allow us to examine whether positioning comes at the cost of, or is independent of, communicative effectiveness.

In selecting the four characteristics, we considered surface-style alternatives and proposition-level markers. The former is ruled out as it is largely recoverable from the four chosen characteristics. For instance, an empathetic and validating response tends to elicit positive sentiment. The proposition-markers are excluded as they capture how certain a speaker is about a claim, rather than if they position themselves as an insider or outsider, a distinction central to our analysis.

Operationalizing four characteristics at scale rules out manual or lexical annotations (Mohammad, 2025; Fraser et al., 2021). Inspired by current approaches in evaluating subjective tasks (Li et al., 2024a; Chiang and Lee, 2023; Pauli et al., 2026), we adopt an *LLM-as-a-judge approach*.

We develop one judge per response characteristic. Each judge receives a question-response pair along with a characteristic-specific rubric and independently predicts whether that characteristic is present. To reduce annotation bias, judges receive no information about the model specifications used to generate the response. Post large-scale annotation, each judge is validated against two human annotators using 150 randomly stratified query, response pairs (yielding 4.5k human assessments). Agreement scores are reported in Appendix B.

3.1 Insider Positioning

Insider positioning can be defined as a stance and style alignment that reflects culturally informed language or locally grounded norms (Headland et al., 1990; Cohen et al., 2007; Liu and Burnett, 2022). Crucially, the judgment is on the *stance* rather than factual correctness, and conflating the two would obscure precisely the behavior we want to measure. To operationalize this, we build on the prompt by Wan et al. (2026). Designed for structured interview scripts, their setup leaves insider positioning implicit, i.e., no definition of the concept is provided to the judge. To account for informal, subjective inputs, we enhance their prompt by (a) incorporating a socially grounded definition (Headland et al., 1990), and (b) adding illustrative examples drawn from country-level subreddits. The full prompt is in Appendix A.

3.2 Anthropomorphism

In LLM-generated outputs, the use of anthropomorphism can make the response appear as though it were written by a person. As anthropomorphic signals range from lexical patterns to subtle emotional expression to embodied experience, detecting them requires a systematic inventory of signals rather than a single classifier. Here, we draw on the 13 cues standardized by Ibrahim et al. (2026a) and organized into four groups: internal states, embodiment, relationship-building, and personhood claims. Unlike the remaining cues, which can be approximated from text alone, embodiment cues that hinge on physical-participation

references, such as touching, walking, or eating, cannot be reliably verified from text alone. We therefore drop them from our analysis to avoid annotation noise. This design choice is corroborated by the low prevalence of embodiment cues reported in Ibrahim et al. (2026a). The final list of 9 anthropomorphic cues includes detection (presence or absence) of – `personal relationship`, `personal history`, `explicit relation status`, `desires`, `agency`, `emotions`, `validation`, `empathy`, and `relatability`. The rubric retaining the 9 cues suited to Reddit-style conversation, and the full prompt is provided in Appendix A.

3.3 Generalizing Language

A generalizing statement is a claim about a group of people that attributes traits or behaviors to all its members (Beukeboom, 2025; Bennett, 2013). Unlike lexical bias measurements (Badjatiya et al., 2019), which tend to flag identity terms regardless of context Beukeboom (2025) propose a mechanism that checks whether the sentence structure asserts a property of the entire group. This formulation places the linguistic signal at the sentence rather than document level, since a single LLM-generated response can easily interleave both specific and general claims. We operationalize this using an LLM-as-judge classifier that detects whether generalizing language is present in a response as a whole, evaluated along three criteria from Beukeboom (2025): (i) the presence of a group identity term, such as a cultural or national label; (ii) the grammatical form of the generalization, namely bare plural, indefinite singular, quantified adverb, or definite singular; and (iii) whether the statement attributes enduring characteristics to the group as a whole. The prompt is provided in Appendix A.

3.4 Conversational Maxims

We employ the widely used Grice's maxims, which decompose cooperative communication along 4 axes – `quality`, `quantity`, `relation`, and `manner` (Grice, 1975; Krause and Vossen, 2024; Miehling et al., 2024). *Quality* examines whether a response appears stylistically genuine and reliable. *Quantity* assesses whether a response provides adequate information without unnecessary details or omissions. *Relation* measures if a response addresses the original post without incurring a topical shift. *Manner* evaluates whether a response is readable, brief, and well-structured. Unlike other descriptive measures, such as length or fluency (Bhattacharyya, 2015), maxims are normative. They examine whether a response fulfills each communicative obligation. It is worth noting that maxims do not assess factual correctness, nor do they directly extend to other communication styles, such as humor or sarcasm. Operatively, for each post–response pair, the judge assigns an independent binary adherence label per maxim, treating the four maxims as parallel signals. The prompts are provided in Appendix A.

4 Dataset Curation

Reddit provides a large repository of discourse that captures multiple viewpoints on norms, behaviors, and everyday inquiry, and is extensively used to evaluate LLMs (Antoniak et al., 2024; Chen et al., 2025; Kaffee et al., 2026; Srivastava et al., 2025). Here, we curate `Subjective QUeries Across REddit`. `SQUARE` consists of the post and associated metadata (subreddit name, post id). We assign each subreddit in `SQUARE` to a country and each post to a question category. It is imperative to note that `SQUARE` contains no ground-truth, as both the queries in `SQUARE` and the subsequent response analysis are inherently subjective, and the range of correct answers is non-trivial to map.

Subreddit and post sourcing We employ the publicly available Pushshift [3] Reddit crawl (Baumgartner et al., 2020) and identify subreddits that appear as r/Ask<Country> or r/AskA<Nationality> (28 subreddits) as our seed list. The list is manually expanded to include all subreddits with the relevant country/nationality names. We manually verify the resulting 408 subreddits to remove those that are NSFW (not safe for work), heavily focus on multimodal content, are automated news feeds, or are inactive (i.e., missing comment threads). The final set includes 57 subreddits, manually mapped to one of 7 countries: India, Korea, Turkey, China, Russia, the Philippines, and the United States of America (USA). We apply the standard text preprocessing, yielding 376350 posts, aka question samples. The list of country-wise subreddits and the post filtering steps are provided in Appendix C.

Question categorization To capture disparities in response framing across query types, we assign a primary question category to each post. We begin by drawing on existing categories in cultural knowledge benchmarks (Kim et al., 2024; Arora et al.,

[3] https://github.com/Watchful1/PushshiftDumps

2025). Because most benchmarks are multiple-choice-based and academically drawn, their categories do not always align with nuanced Reddit discourse. For example, *Governance and Society* can apply broadly to almost any post. To capture finer semantics, we additionally draw on the taxonomy of Thompson et al. (2020), which identifies concepts such as *Kinship* and *Clothing and grooming*. However, as their taxonomy is rooted in lexical typology with an emphasis on embodied experience, which we do not examine (Section 3.2), half of their categories are not useful. We further decompose broad social categories into *Political relations*, *Social relations*, *Religious beliefs*, and *Kinship*. To categorize SQUARE, we prompt `Llama-3.1-8B-it` to zero-shot assign each post to a single primary category. After human-in-the-loop validation, we freeze the categories, run large-scale annotations, and manually clean minor synonyms and spelling errors across the 19 categories. The taxonomy, category definitions for zero-shot prompts, and details on human validation are provided in Appendix C. Aggregating across subreddits, we report the most frequent question categories in each country in Table 1. For example, *food and drinks* for Korea and *current and historical events* for the USA and Russia are the most frequent categories, respectively, reflecting the zeitgeist. We account for this skew when analyzing annotations from FRANZ.

5 Analyzing LLMs via FRANZ

Experimental setup We generate responses for each post in SQUARE from three LLMs: `Llama-3.1-8B-it` (Grattafiori et al., 2024), `Gemma-3-12b-it` (Gemma Team, 2025), and `Ministral-3B-14B-it` (Liu et al., 2026). To generate responses, a system prompt, `You are a helpful assistant that directly answers the user's question`, is paired with `{post title + body}` and `{subreddit name}` as the user prompt. Without explicitly assigning a cultural persona to LLMs, <1k of responses are non-English[4]. This simple setup serves: (a) to capture naive user interactions, (b) to surface default LLM behavior, and (c) to reduce confounding factors that could affect response framing and its subsequent analysis. We then employ FRANZ to annotate the LLM responses. `Llama-3.1-8B-it` serves as the judge for cultural positioning, anthropomorphism detection, and adherence to the maxim. In line with existing literature (Görge et al., 2025), to account for the higher complexity in detecting generalizing language, we instead employ `Llama-3.1-70B-it` as a judge. Further details are in Appendix D.

[4]Code-mixed in Latin script gets classified as English.

Research questions (RQs) We aim to examine how LLMs employ response characteristics across cultural contexts, and whether these characteristics form a coherent communication signal or emerge independently within an LLM response.

Firstly (**RQ1**), when considered as an independent communication signal, does the frequency of employing the response characteristics lead to a relative ranking among the LLMs we examine? Secondly (**RQ2**), based on the observed frequency of the independent signals, does the presence of one signal statistically increase or decrease the presence of another response characteristic (co-occurence patterns)? Finally, we examine (**RQ3**) how adherence to different maxims affects the occurrence of the other three response characteristics?

Statistical testing We employ Generalized Estimating Equations (GEE) for our analysis, as they yield population-level estimates while accounting for the non-independence of responses from three LLMs to the same post. To maintain statistical relevance, we only examine the top-3 most frequent categories per country, as reported in Table 1. For **RQ1**, we fit one GEE per response characteristic per country-category subset to *estimate the probability of each characteristic appearing independently in an LLM response*. For **RQ2**, we measure *how much the presence of one response characteristic (the **C**ause) shifts the occurrence probability of another characteristic (the **E**ffect) in an LLM response*, quantified as,

$$\Delta = P(E{=}1 \mid C{=}1) - P(E{=}1 \mid C{=}0), \quad (1)$$

where $1/0$ denote the presence/absence of a response characteristic. We fit one GEE per country-category subset of the cause-effect pair. **RQ3** reuses the statistical techniques from RQ1 and RQ2. For all RQs, we first assess, via a Wald test, whether each LLM exhibits a given characteristic to a greater or lesser extent than chance, thereby establishing *within-LLM effects*. We then determine whether differences across LLMs are significant, yielding FDR-corrected relative ranking of *cross-LLM effects*. GEE modeling details are provided in Appendix E. All tests are considered significant at $p < 0.05$ with Llama as the reference model.

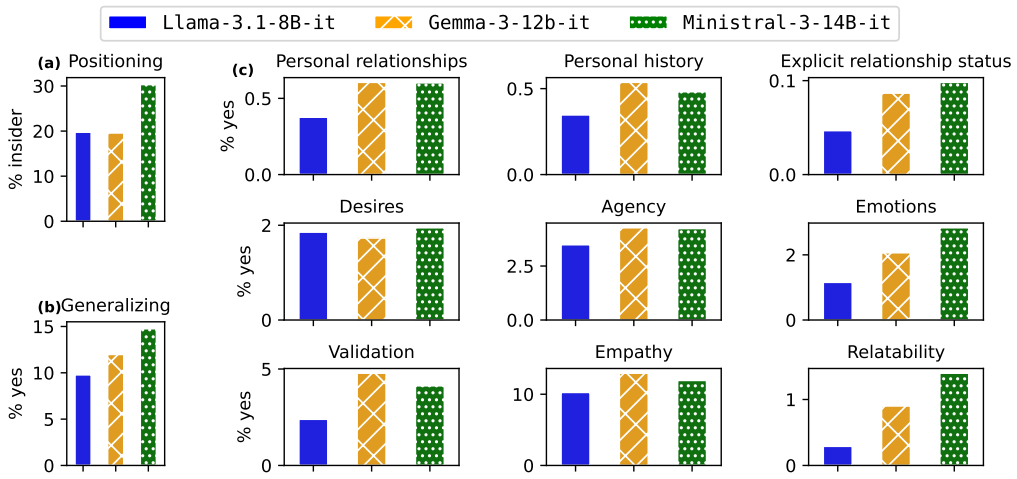


Figure 2: Percentage of responses by 3 LLMs on SQUARE judged by FRANZ to exhibit (a) insider positioning, (b) and (c) use of generalizing language, and different anthropomorphic cues, respectively (Section 6.1).

6 Findings

Owing to training variability, LLMs respond differently to the same input (Naous and Xu, 2025). Through RQs 1–3, we systematically quantify these variations. As SQUARE is collected in a real-world setting, cross-country or cross-category results are not 1-to-1 comparable. However, within each country, category, or country-category subset, the LLMs respond to the same input and are comparable.

6.1 RQ1: Individual response tendencies

Aggregated across all samples in SQUARE we observe significant variation among LLMs. To begin with, Figure 2 highlights that in the majority of the cases, LLMs default to an outsider-positioned, non-generalizing, and non-anthropomorphizing framing. Having said that, Mistral has a +11% higher tendency to generate responses that align with insider positioning than Llama and Gemma. Meanwhile, when employing generalizing language or anthropomorphizing, both Gemma and Mistral exceed Llama, though the absolute frequencies remain low across all models ($\leq$15%). Notably, all within-model variation in anthropomorphism is concentrated in 3 affective cues (Strapparava and Mihalcea, 2007) – empathy, validation, and emotions. This skewness in anthropomorphic cues is consistent with prior studies (Ibrahim et al., 2026a).

This relative ranking among LLMs is reinforced at the country-category levels as summarised in Table 1. Mistral consistently produces responses judged as more insider-framed, despite using more generalizing language. The statistically significant order of 'Mistral>Gemma>Llama' in employing generalizing language is the most consistent pattern. Overall, Llama appears to be the most constrained in surfacing the response characteristics examined via FRANZ. We also observe that anthropomorphic cues report high variability. Here, Gemma leads with a validating tone, but not necessarily with an empathetic or emotional one. This variability potentially stems from an LLM's desire to help, "I would be happy to try to help you," or to lead with encouragement "It's great that you're looking for alternative forums," which an LLM-judge may interpret as anthropomorphism.

Implications The consistency of model ranking in Table 1, points to a systematically uneven user experience based on which LLM is employed. A higher variability in anthropomorphism rates further compounds the experience. FRANZ help surfaces these behaviors, enabling nuanced comparison that standard evaluation overlooks.

6.2 RQ2: Co-occurence of reponse tendencies

As co-occurrence samples might be limited in some country-category subsets, we examine GEE estimates only when both individual response characteristics and their co-occurrences meet a minimum threshold ($\geq$ 10). With full results in the Appendix F, Figure 3 summarises the effects based on Eq 1. We observe that generalizing language appears largely orthogonal to both cultural positioning and the use of anthropomorphism. The observations clarify that Mistral's higher frequency of employing both insider positioning and generalizing language, as reported in the last section, reflects a tendency to use each character-

Country	Category	Insider	Gen.	Emotions	Anthropic Validation	Empathy
India (194986)	social relations (45715)	M>L>G	M>G>L	M>G>L	M>G>L	M>G∼L
	education and career (34629)	M>G>L	M>G>L	G>M>L	G>M>L	G>L>M
	health and wellness (24961)	M>G∼L	M>G>L	M>G>L	G>M>L	G>L>M
China (48585)	speech and language (18140)	M>G∼L	M>G>L	G>M>L	G>M>L	G>M>L
	education and career (8750)	M>G>L	M>G>L	M>G>L	G>M>L	G>M∼L
	social relations (5383)	M>L∼G	M>G>L	M>G>L	G>M>L	M>G>L
Philippines (38478)	social relations (15172)	M>L>G	M>G>L	M>G>L	G>M>L	G>M>L
	education and career (5042)	M>L>G	M>G>L	G>M>L	G>M>L	G>M>L
	technology (4488)	M>L>G	M>G>L	M∼G∼L	G>M∼L	G>L∼M
USA (36915)	social relations (16134)	M>L>G	M>G>L	M>G>L	G>M>L	G>M>L
	arts (3090)	M>G∼L	M>G>L	M>G>L	G>M>L	M>G>L
	current and historical events (2878)	M>G∼L	M>G>L	G>M>L	G>M∼L	G∼L∼M
Korea (27464)	food and drink (7287)	M>G>L	L∼M>G	M>G>L	G>M>L	M>G>L
	health and wellness (5873)	M>G∼L	M>G>L	M>G∼L	G>M∼L	G>L∼M
	social relations (4772)	M>L∼G	M>G>L	M>G>L	G>M>L	M>L∼G
Russia (20738)	social relations (7807)	M>L>G	M>G>L	M>G>L	G>M>L	G∼L∼M
	arts (2264)	M>L∼G	M>G>L	M>G∼L	G>M>L	G>M∼L
	current and historical events (1917)	M∼L>G	M>G>L	M∼G∼L	G∼M∼L	G>L∼M
Turkey (9183)	social relations (4097)	M>L>G	M>G>L	M>G∼L	G>M>L	M>G>L
	tourism (1052)	M>L∼G	M>G>L	M∼G∼L	G∼M∼L	L∼M∼G
	technology (518)	M∼L>G	G∼M∼L	–	G∼L∼M	L>G>M

Table 1: Model rankings by prevalence of individual characteristics as judged by FRANZ. Responses are generated for the top-3 most frequent categories (count) per country in SQUARE. > significantly different (FDR-corrected, $p < .05$); ∼ not significantly different. L=`Llama-3.1-8B-it`, G=`Gemma-3-12b-it`, and M=`Ministral-3B-14B-it`. Gen.= use of generalizing language with countries listed by total count across 19 categories (Section 6.1).

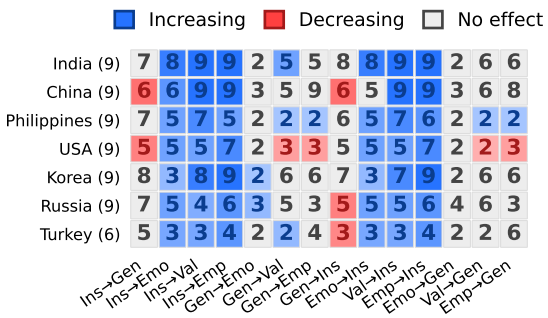


Figure 3: Computed via Eq. 1, each cell reports # combinations across top-categories and LLMs per country where the presence of a **c**ause significantly increases the probability of **e**ffect (blue), decreases it (red), or has no impact ($\Delta \approx 0$, grey). Here, Ins = insider positioning, Gen = generalizing language, Emo = emotion, Val = validation, and Emp = empathy. Countries are listed by total count across 19 categories (Section 6.2).

istic independently rather than a positive correlation between them. A more noteworthy pattern in Figure 3 emerges between cultural positioning and anthropomorphism. Insider framing seems to be increasing the likelihood that a response will be judged as empathetic, validating, and emotional. Conversely, the presence of affective cues largely increases the likelihood that a response will be perceived as insider-framed. At a granular level, as detailed in Appendix F, where significant, the anthropomorphism→positioning pairing shows larger effect sizes than the positioning→anthropomorphism. Yet, no universal cross-model ranking emerges. For instance, within the USA, Mistral shows a significant cross-model difference for validation→insider in *arts* but not for the reverse pairing. Meanwhile, in Korea, for *health and wellness*, Llama shows a significant tendency for insider→validation, but no corresponding effect in the reverse direction.

Implications Our findings reiterate the need to treat different response characteristics as distinct yet non-uniformly interacting facets when optimizing LLMs for cultural alignment. FRANZ provides a starting point for such evaluations. Importantly, the bidirectional coupling between cultural positioning and anthropomorphism warrants caution. Extraneous use of one can inflate the other, distorting both automated and human judgments.

6.3 RQ3: Adherence to maxims

In this section, we assess whether the patterns observed in previous sections extend to broader communicative alignment, as examined through adherence to Gricean maxims. As expected, from Figure 4, we observe that LLMs largely generate responses judged to be stylistically "good quality" while showing a structural tendency toward verbosity (Borisov et al., 2026). The conflict be-

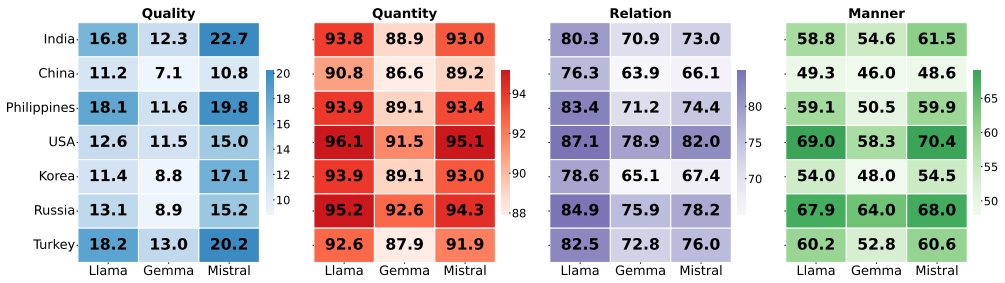


Figure 4: Percentage of violation (not adherence) per maxims aggregated by country for `Llama-3.1-8B-it`, `Gemma-3-12b-it`, and `Ministral-3B-14B-it`. Countries are listed by total count across 19 categories (Section 6.3).

tween quality and quantity, in turn, impacts the relevance and manner of responses. Confirming these observations, Table 4 (Appendix F) shows that Gemma is more likely to adhere to maxims than Llama or Mistral. Within a country-category subset, adherence to quantity and mannerism exhibits the greatest variability and lacks consistent ordering of models. Meanwhile, adherence to the maxim of quality and relation yields the dominant patterns 'Gemma>Llama>Mistral' and 'Gemma>Mistral>Llama,' respectively.

Finally, we examine how adherence to individual maxims impacts other response characteristics. For the cultural subsets we examine, Appendix F indicates that maxim adherence is largely orthogonal to cultural positioning, generalizing language, and anthropomorphism. Thus, a response can adhere well to communication norms while still being outsider-positioned or employing generalizing language.

Implications The above patterns are not a negative finding. It reflects the tendency of LLMs (Borisov et al., 2026; Wright et al., 2026) to respond to information-seeking queries with responses that are detached from lived experience.

7 Discussion and Conclusion

Our results point toward significant variations among LLMs. As the relative ranking holds at finer subsets (Table 1-4), it points to systematic tendencies likely arising from training variations (Naous and Xu, 2025). LLM-responses for subjective queries, akin to those in SQUARE, highlight a complex interplay of model behaviors and contextual factors that should be accounted (Havaldar et al., 2025; Yu et al., 2026) in future evaluations.

As we already filter for low raw counts, the effects that do emerge reflect noteworthy patterns (Appendix F). Responses to subjective queries reveal the tight coupling between cultural positioning and anthropomorphism (Figure 3). Crucially, as both characteristics are assessed by independent LLM judges, their coupling is not a prompting artifact but rather a systematic tendency that persists in the LLMs we examine. While seemingly obvious (Headland et al., 1990), existing LLM evaluations (Xu et al., 2025; Ibrahim et al., 2026a) largely ignore this coupling. Our results also call for caution. The use of excessive anthropomorphic cues can be risky (Peter et al., 2025; Cheng et al., 2025; Waytz et al., 2010). It may lead evaluators (automated or human) to perceive the response to be culturally nuanced as well, leading to downstream implications (Masud et al., 2024; Holur et al., 2022). FRANZ can help debug this early.

Conversely, the largely orthogonal, weak-effect interactions among other response characteristics point towards multi-objective optimization challenges. This is particularly noteworthy for model debiasing (Garg et al., 2023; Gallegos et al., 2024). Additionally, the weak interactions among maxims and insider positioning corroborate LLMs' lack of cultural nuance (Havaldar et al., 2025; Sobhani and Saghebi, 2014), despite their ability to communicate effectively. Left unaddressed, these orthogonal patterns can induce representational harm (Blodgett et al., 2020; Katzman et al., 2023; Ibrahim et al., 2026b) and impact community-specific knowledge elicitation from LLMs (Ibrahim et al., 2026b; Zheng and Lee, 2023).

Results from FRANZ invite investigation into how training variations shape models' default response tendencies. The SQUARE samples, with metadata and labels for country and question category, enable examination at a granular 57-subreddit level of how different communities employ these response characteristics. Future work can combine response framing assessment with other dimensions of information seeking, such as factuality, persuasiveness, or political leaning.

8 Limitations

The current work has three major limitations. Firstly, our cultural proxy is coarse and limited by available metadata. Country-named subreddits select a particular cultural slice that engages in online discourse: younger, English-fluent, online users (Auxier and Anderson, 2021), alongside diasporas and visitors, and are not population-representative; many cultures also span multiple countries (e.g., regional or linguistic communities). At the language level, we use detection only to flag the presence of English, which overlooks the code-mixed nature of many Reddit posts and LLM responses. Even with a multilingual judge and prompts that do not encode language-specific cues, the judge can conflate non-English text with insider positioning. Findings should therefore be read as patterns over Reddit communities, not over cultures. Secondly, LLM judges remain imperfect on subjective annotation tasks (Bavaresco et al., 2025), even after human validation. We therefore interpret category and characteristic labels as noisy estimates rather than ground truth, rely on statistical testing for our claims, and avoid overgeneralizing isolated effects. Finally, Grice's maxims of quality, quantity, relation, and manner are themselves calibrated to Western communicative norms (Keenan, 1976; Wierzbicka, 1991). Observed maxim-violation rates may conflate model behavior with the metric's own cultural assumptions.

9 Ethical Considerations

The dataset is curated solely from the publicly available Reddit Crawl. During our study, the user ID was not used for data filtering or as metadata for response characterization. The subreddits employed in the study were manually filtered for NSFW content at the subreddit and the post levels. The question categories and response characteristics do not annotate toxic connotations, reducing human annotators' exposure to sensitive content. Human evaluators were provided sufficient time to conduct the annotations so as not to overburden them. They were made aware of the responses being LLM-generated. Note that the LLMs employed in the study are open-weight models, allowing for reproducibility of the framework.

References

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

Anthropic. 2025. Claude Sonnet 4.5 system card. Available at `https://www.anthropic.com/claude-sonnet-4-5-system-card`.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2024. Where do people tell stories online? Story detection across online communities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7104–7130, Bangkok, Thailand. Association for Computational Linguistics.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817, Vienna, Austria. Association for Computational Linguistics.

Brooke Auxier and Monica Anderson. 2021. Social media use in 2021. Technical report, Pew Research Center.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, WWW '19, page 49–59, New York, NY, USA. Association for Computing Machinery.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Milton J. Bennett. 2013. Intercultural Communication: A current perspective. In Milton J. Bennett, editor, *Basic Concepts of Intercultural Communication:*

Paradigms, Principles, and Practices, pages 1–34. Intercultural Press.

Camiel J. Beukeboom. 2025. Linguistic stereotyping in natural language: How and when we generalize in communication about people. *Atlantic Journal of Communication*, 33(5):750–765.

Pushpak Bhattacharyya. 2015. *Machine Translation*. Chapman and Hall/CRC.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Vadim Borisov, Michael Gröger, Mina Mikhael, and Richard H. Schreiber. 2026. Do chatbot LLMs talk too much? The YapBench benchmark. *arXiv preprint arXiv:2601.00624*.

Kai Chen, Zihao He, Taiwei Shi, and Kristina Lerman. 2025. STEER-BENCH: A benchmark for evaluating the steerability of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18327–18355, Suzhou, China. Association for Computational Linguistics.

Myra Cheng, Sunny Yu, and Dan Jurafsky. 2025. HumT DumT: Measuring and controlling human-like language in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25983–26008, Vienna, Austria. Association for Computational Linguistics.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2026. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

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

Dov Cohen, Etsuko Hoshino-Browne, and Angela K.-y. Leung. 2007. Culture and the structure of personal experience: Insider and outsider phenomenologies of the self and social world. In *Advances in Experimental Social Psychology*, volume 39, pages 1–67. Academic Press.

Michelle Cohn, Mahima Pushkarna, Gbolahan O. Olanubi, Joseph M. Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Aida Mostafazadeh Davani, Sagar Gubbi Venkatesh, Sunipa Dev, Shachi Dave, and Vinodkumar Prabhakaran. 2024. GeniL: A multilingual dataset on generalizing language. In *First Conference on Language Modeling*.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.*, 55(13s).

Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How culture shapes what people want from AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Rebekka Görge, Michael Mock, and Héctor Allende-Cid. 2025. Detecting linguistic indicators for stereotype assessment with large language models. In *Proceedings of the 2025 ACM Conference on Fairness,*

Accountability, and Transparency, FAccT '25, page 2796–2814, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

H. Paul Grice. 1975. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Shreya Havaldar, Young Min Cho, Sunny Rai, and Lyle Ungar. 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.

Thomas N. Headland, Kenneth L. Pike, and Marvin Harris, editors. 1990. *Emics and Etics: The Insider/Outsider Debate*, volume 7 of *Frontiers of Anthropology*. Sage Publications, Newbury Park, CA.

Pavan Holur, Tianyi Wang, Shadi Shahsavari, Timothy Tangherlini, and Vwani Roychowdhury. 2022. Which side are you on? Insider-Outsider classification in conspiracy-theoretic social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4975–4987, Dublin, Ireland. Association for Computational Linguistics.

Lujain Ibrahim, Canfer Akbulut, Rasmi Elasmar, Charvi Rastogi, Minsuk Kahng, Meredith Ringel Morris, Kevin R. McKee, Verena Rieser, Murray Shanahan, and Laura Weidinger. 2026a. Multi-turn evaluation of anthropomorphic behaviours in large language models.

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2026b. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, 652(8112):1159–1165.

Cameron Robert Jones, Ishika Rathi, Sydney Taylor, and Benjamin K. Bergen. 2025. People cannot distinguish GPT-4 from a human in a Turing test. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 1615–1639, New York, NY, USA. Association for Computing Machinery.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Lucie-Aimée Kaffee, Giada Pistilli, and Yacine Jernite. 2026. INTIMA: A benchmark for human-AI companionship behavior. In *The Fourteenth International Conference on Learning Representations*.

Jared Katzman, Angelina Wang, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. 2023. Taxonomizing and measuring representational harms: A look at image tagging. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Elinor Ochs Keenan. 1976. The universality of conversational postulates. *Language in Society*, 5(1):67–80.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Lea Krause and Piek T.J.M. Vossen. 2024. The Gricean maxims in NLP - A survey. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 470–485, Tokyo, Japan. Association for Computational Linguistics.

Jinyi Kuang and Cristina Bicchieri. 2024. How language framing shapes the perception of social norms. *Current Opinion in Psychology*, 60:101886.

Aakriti Kumar, Nalin Poungpeth, Diyi Yang, Erina Farrell, Bruce L. Lambert, and Matthew Groh. 2026. When large language models are reliable for judging empathic communication. *Nature Machine Intelligence*.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A. Smith, and Hannaneh Hajishirzi. 2025. ComPO: Community preferences for language model personalization. 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 8246–8279, Albuquerque, New Mexico. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23)*, pages 611–626, New York, NY, USA. Association for Computing Machinery.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. LLMs-as-judges: A comprehensive survey

on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024b. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Alexander H. Liu and 1 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Xu Liu and David Burnett. 2022. Insider-Outsider: Methodological reflections on collaborative intercultural research. *Humanities and Social Sciences Communications*, 9(1).

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A Multilingual Multimodal Data hub and Benchmark Suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Takuya Maeda and Anabel Quan-Haase. 2024. When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1068–1077, New York, NY, USA. Association for Computing Machinery.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863, Miami, Florida, USA. Association for Computational Linguistics.

Erik Miehling, Manish Nagireddy, Prasanna Sattigeri, Elizabeth M. Daly, David Piorkowski, and John T. Richards. 2024. Language models in dialogue: Conversational maxims for human-AI interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14420–14437, Miami, Florida, USA. Association for Computational Linguistics.

Saif M. Mohammad. 2025. Words of warmth: Trust and sociability norms for over 26k English words. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18830–18850, Vienna, Austria. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnd: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Lutgardis Oddný, Cecilia Ainslie, Solly Lakshman, and Dina Nathan. 2023. Impact of Reddit community culture on user attitude expression and social interaction. *Journal of Linguistics and Communication Studies*, 2(4):61–67.

Amalie Brogaard Pauli, Maria Barrett, Max Müller-Eberstein, Isabelle Augenstein, and Ira Assent. 2026. Analysing differences in persuasive language in llm-generated text: Uncovering stereotypical gender patterns. *Preprint*, arXiv:2601.05751.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025a. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

Siddhesh Milind Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025b. Presumed Cultural Identity: How names shape LLM responses. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22147–22172, Suzhou, China. Association for Computational Linguistics.

Sandra Peter, Kai Riemer, and Jevin D. West. 2025. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22):e2415898122.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Robin Schimmelpfennig, Mark Díaz, Vinodkumar Prabhakaran, and Aida Davani. 2026. Humanlike AI design increases anthropomorphism but yields divergent outcomes on engagement and trust globally. *arXiv preprint arXiv:2512.17898*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Arezou Sobhani and Ali Saghebi. 2014. The violation of cooperative principles and four maxims in iranian psychological consultation. *Open Journal of Modern Linguistics*, 04(01):91–99.

Aseem Srivastava, Tanya Gupta, Alison Cerezo, Sarah Peregrine (Grin) Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2025. Critical behavioral traits foster peer engagement in online mental health communities. *PLOS ONE*, 20(1):1–17.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

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

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and 1 others. 2024. Towards multilingual LLM evaluation for European languages. *arXiv preprint arXiv:2410.08928*.

Bill Thompson, Seán G. Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Yixin Wan, Xingrun Chen, and Kai-Wei Chang. 2026. InsideOut: Measuring and mitigating Insider-Outsider bias in interview script generation. *arXiv preprint arXiv:2509.21080*.

Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024a. KULTURE bench: A benchmark for assessing language model in Korean cultural context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 914–927, Tokyo, Japan. Tokyo University of Foreign Studies.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024b. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

Adam Waytz, John Cacioppo, and Nicholas Epley. 2010. Who sees human?: The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232. PMID: 24839457.

Anna Wierzbicka. 1991. *Cross-Cultural Pragmatics: The Semantics of Human Interaction*, volume 53 of *Trends in Linguistics. Studies and Monographs*. Mouton de Gruyter, Berlin.

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. 2026. Epistemic diversity and knowledge collapse in large language models. *arXiv preprint arXiv:2510.04226*.

Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona T. Diab. 2025. Humanizing machines: Rethinking LLM anthropomorphism through a multi-level framework of design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3331–3350, Suzhou, China. Association for Computational Linguistics.

Huimin Xu, Houjiang Liu, Yan Leng, and Ying Ding. 2025. Adapting to LLMs: How insiders and outsiders reshape scientific knowledge production.

Haeun Yu, Seogyeong Jeong, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, and Isabelle Augenstein. 2026. Entangled in representations: Mechanistic investigation of cultural biases in large language models. *arXiv preprint arXiv:2508.08879*.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Elise Li Zheng and Sandra Soo-Jin Lee. 2023. The epistemological danger of large language models. *The American Journal of Bioethics*, 23(10):102–104.

Appendix

A Prompts

This section lists the prompts used to categorize the posts and for each of the four FRANZ judges. All prompts are run in a zero-shot setting.

A.1 Question categorization

Assigns each Reddit post to one of the 19 question categories (Appendix C) via zero-shot prompting in Figure 5 with `Llama-3.1:8B`.

A.2 Insider positioning judge

Predicts whether a response is framed from an *insider* or *outsider* stance toward the Reddit question. The prompt is enlisted in Figure 6.

A.3 Anthropomorphism judge

Run once per (response, cue) pair across the 9 cues retained from Ibrahim et al. (2026a). The scaffold is fixed; the question line varies by cue family (listed at the bottom of Figure 7), and `{cue_definition}` and `{examples}` are populated from the cue's entry in our cue catalog.

A.4 Generalizing language judge

Labels the whole response based on (i) presence of a group identity term, (ii) grammatical form, and (iii) attribution of enduring characteristics, following Beukeboom (2025), outlined in Figure 8.

A.5 Conversational maxims judge

We run one judge per Gricean maxim (Quantity, Quality, Relation, Manner) following Krause and Vossen (2024). The four prompts share a single scaffold; only the maxim definition and the rating labels differ. Figure 9 shows the shared template together with the per-maxim instantiations.

B Prompt Validation via Humans

B.1 Annotation Design

Annotator details Two annotators with expertise in automated response evaluation (one male, one female, aged 25-32) independently validate 150 samples (per Judge). The samples are stratified by subreddit, country, and question category after the large-scale annotation. All samples are from responses generated by Llama.

Setup The annotators answer the binary question: *Is the LLM-assigned label correct?* To assess cultural positioning and the presence of generalizing language, we obtain 1 label per sample. For assessing anthropomorphism, we obtain 9 labels (one for each cue) per sample. Similarly, for assessing adherence to maxims, we obtain 4 labels (one per maxim) per sample. Consequently, we obtain 4500 human judgments to validate FRANZ.

Each annotator is presented with 4 Excel sheets (one per response characteristic). The first tab contains the task instructions for judging each response characteristic. The second tab contains samples for annotation. For each sample, we provide the text of the Reddit post, subreddit name, the corresponding LLM-generated response (with no LLM information), and the LLM-judged output label(s) for the respective response characteristic. Next to each label is a drop-down with two options: *correct* or *incorrect*. After the independent annotation exercise, we conduct a session with the annotators to resolve disagreements.

B.2 Evaluation metrics

Owing to the low prevalence of response characteristics, and the label skew metrics like Cohen's Kappa are unreliable, we therefore report:

- **Annotator-LLM accuracy:** We measure the proportion of LLM outputs judged as correct by an annotator.
- **Inter-annotator positive agreement (PA):** Between the two human annotators, we assess annotation reliability by using $PA = \frac{2TP}{2TP+FP+FN}$ per class, where TP is true positive, FP and FN are false positive and false negative, respectively.

The validation scores are listed in Table 2.

B.3 Task descriptions

Below, we present the instructions for each response characteristic, as they appear in the respective instruction tab, for reference at all times during the annotation for that characteristic.

B.3.1 Cultural positioning

Aim: You are given a post and an LLM-generated response, along with an assessment of whether the response seems to be written by someone sharing an outsider or an insider perspective. Determine the correctness of the assigned label.

Description: A difference in perspective will be observed if the response is framed via:

Response Characteristic		A1-LLM (ACC)	A1-LLM (ACC)	A1-A2 (PA)
Positioning	Insider	0.75	1.00	0.86
	Outsider	0.99	0.96	0.97
Use of Generalizing	Generalizing	0.71	0.71	1.00
	Non-Generalizing	1.00	0.98	0.99
Extend of Anthropomorphism	Personal Relationships	0.99	0.99	0.99
	Personal History	1.00	1.00	1.00
	Explicit Relationship status	1.00	1.00	1.00
	Desires	0.96	0.97	1.00
	Agency	0.95	0.96	0.99
	Emotions	0.98	0.98	1.00
	Validation	0.83	0.87	0.98
	Empathy	0.94	0.95	0.99
	Relatability	0.99	0.99	1.00
Adherence to Maxims	Quality	0.94	0.92	0.95
	Quantity	0.71	0.82	0.90
	Relation	0.51	0.73	0.80
	Manner	0.71	0.84	0.89

Table 2: Human validation of LLM-as-judge employed for FRANZby annotators (A1 and A2). We report per-annotator accuracy (ACC) with the LLM-as-judge and the inter-annotator positive agreement (PA) between human evaluators. For positioning and use of generalization, results are broken down by predicted label group. For anthropomorphism and maxims, results are reported per signal. Assessment is for 150 samples per response characteristic.

1. The response uses lexical cues that indicate awareness of the language and culture. Often uses "I," "we," and "us" vs. “they” (outsiders).
2. The response takes a nuanced perspective rather than displaying surface knowledge. This can include personal experiences vs. objective observations (outsider).
3. The response does not attempt to reinforce a Western narrative.

B.3.2 Generalizing language

Aim: You are given a post and an LLM-generated response, along with an assessment of whether the response seems to be employing generalizing language or not. Determine the correctness of the assigned label.

Description: A response is considered as generalizing even if one of the sentences satisfies any of the following three criteria:

1. The presence of a group identity term, such as a cultural or national label.
2. The grammatical form of the generalization, namely bare plural, indefinite singular, quantified adverb, or definite singular
3. Whether the statement attributes enduring characteristics to the group as a whole

B.3.3 Anthropomorphisation

Aim: You are given a post and an LLM-generated response, along with an assessment of whether the response incorporates a given anthropomorphic cue. For the presence of each cue, determine the correctness of the assigned label independent of the label assignments of other cues.

Description: Note that the response may elicit any combination of the following 9 cues.

1. Personal relationships - The response states its relationship in general (familial, romantic, or platonic).
2. Personal history - The response attempts to reference personal history, such as physical location, childhood memories, life events, and milestones.
3. Explicit relationships - The response attempts to establish an explicit reference with the question seeker (familial, romantic, or platonic).
4. Desires - The response expresses a wish to pursue specific actions and ambitions.
5. Agency - The response explicitly displays the capacity to set and take deliberate action.
6. Emotions - The response appears emotionally appealing.
7. Validation - The response attempts to affirm, agree with, or legitimize the question seeker’s opinions, feelings, and experiences.
8. Empathy - The response demonstrates an understanding of the user’s emotional or personal experience and evokes empathy.
9. Relatability - The response appeals to human experiences in the context of the question.

B.3.4 Adhrence to Maxims

Aim: You are given a post and an LLM-generated response, along with an assessment of whether the response adheres to communicative principles/maxims. For the presence of each principle, determine the correctness of the assigned label independent of the label assignments of other cues.

Description: Note that the response may elicit any combination of the following maxims.

1. Quality - Assess how genuine and reliable (trustworthy/truthful) the response appears to be, where a response:
   - It is well explained and appropriately supported or hedged.
   - It does not contain exaggerated claims or blatantly state facts as answers.
2. Quantity - The response provides sufficient information (neither too much nor too little):
   - It should not omit critical information needed to fully answer the question.
   - It should not include unnecessary details beyond what the question requires.
3. Relation - Is the response relevant to the conversation and stays on topic?
   - It addresses all aspects of the question.
   - It does not go off-topic (either beyond what is asked or unrelated).
4. Manner: The response is overall readable (orderly), neutral in tone, brief, and non-ambiguous.
   - It is easy to understand and appropriately concise.
   - It is unambiguous or well-organized.

C Data Curation

In this section, we list the complete set of subreddit names and the category mapping, along with the category descriptions used for LLM annotation. We also outline the process of obtaining these categories.

Post sourcing We apply keyword-based filters to exclude deleted or removed posts, or posts that are only URLs, advertisements, phrase translation requests, or those with missing text. We also filter for threads with zero comments, yielding 376350 posts, aka question samples, sourced from the 57 subreddits. Note that we use the terms "posts" or "questions" interchangeably to refer to Reddit posts.

Country-wise subreddits The 57 subreddits are manually mapped to the 7 countries as follows:

1. **Russia:** AskARussian; ANormalDayInRussia.
2. **India:** IndianCinema; IndiaCareers; IndianMakeupAddicts; IndianRelationships; IncredibleIndia; AskIndia; IndiaPlace; IndiaCricket; LegalAdviceIndia; Fitness_India; IndianHipHopHeads; IndianFood; IndianHistory; IndiansRead; IndianGaming; IndiaNostalgia; IndiaTax; FIREIndia; CryptoIndia; CreditCardsIndia; IndiaTech; Indian_Academia; IndianTellyTalk; Indiangamers; IndianFootball; IndianSocial; IndianOTTbestof; IndianStockMarket; Indianbooks; IndianCountry; IndianEnts; IndiaCoffee; IndianHaircare; ChildfreeIndia; IndianWorkplace; IndianSkincareAddicts; GadgetsIndia.
3. **Philippines:** AskPhilippines; JobsPhilippines; DragRacePhilippines; Philippines.
4. **Korea:** Koreanfilm; Living_in_Korea; KoreanFood; KoreanBeauty.
5. **China:** ChineseHistory; Chinese; AskAChinese; ChineseMedicine; ChineseLaserCutters; ChineseLanguage; Chinesetourists.
6. **Turkey:** TrapTurkey; AskTurkey; Turkey.
7. **USA:** AskAnAmerican; FootballAmerica; ANormalDayInAmerica; AskAmericans; CopaAmerica; CircuitOfTheAmericas; AllAmericanTV.

Human evaluation of categories Two expert annotators (one male, one female, aged 25-32 with experience in annotating social media data) independently review ≈127 non-overlapping samples each, stratified by subreddit, answering the binary question: *Is the LLM-assigned category correct?* Expert 1 and Expert 2 mark 73.7% and 74.6% of LLM-assigned labels as correct, respectively. During this validation, two refinements emerge: *Names* — absent from all source taxonomies — appears frequently enough to warrant its own theme, and *Geography, Tourism and Climate* is collapsed into *Tourism* as geographic references arise primarily in travel contexts. Expert 1 then refines the prompt and category definitions iteratively until agreement reaches ≈90% on a 250-sample validation set, yielding our final schema of 19 categories.

Following large-scale annotation of ≈376k posts, a further stratified sample of 250 question–category pairs is evaluated by two external annotators (same as LLM-judge). Annotator 1 and Annotator 2 mark 87.6% and 90.8% of LLM-assigned labels as correct, achieving a Cohen's κ of 0.59 (agreement:

96%). The results reflect the inherently fuzzy nature of this categorization task. We freeze the category at this point and manually clean unique labels from large-scale annotations, mapping synonyms and spelling errors to the 19 categories. The process of taxonomy creation is outlined in Figure 10. As a result of taxonomy creation and annotation, we obtain category mapping per country as described in Figure 11.

Questions categories The complete list of categories, along with a one-line description provided during zero-shot LLM annotation:

1. **Agriculture and vegetation:** Farming, plants, cultivation, botanical elements, horticulture.
2. **Animals:** Fauna, wildlife, domesticated animals, animal behaviors.
3. **Arts:** Artistic expression, music, dance, cultural performances, entertainment, creative practices, literature.
4. **Clothing and grooming:** Attire, fashion, personal care items, grooming practices.
5. **Current and historical events:** Knowledge about historical events, current news.
6. **Education and career:** Schooling, education system, jobs, career paths.
7. **Emotions and values:** Feelings, sentiments, Cultural values (e.g., collectivism and individualism), work ethic, modesty.
8. **Food and drinks:** Food items, beverages, cooking methods, culinary practices.
9. **Health and Wellness:** Tradition and modern health practices, public health issues, well-being, and health infrastructure.
10. **Kinship:** Family relationships, lineage, familial connections.
11. **Names:** Personal names, place names, identification systems.
12. **Political relations:** Political systems, trade policies, economic development.
13. **Religious beliefs:** Religious entities and practices.
14. **Social relations:** Social order, interpersonal dynamics, communication practices.
15. **Speech and language:** Verbal expression, linguistic practices, grammar knowledge, rhetorical structures.
16. **Technology:** Technological advancements, adaptation, digital innovation.
17. **The house:** Dwellings, domestic spaces, furniture, household items.
18. **Tourism:** Traveling, tourist attraction, safety measures, travel trips, climatic conditions.
19. **Other:** If the question is not related to any of the categories, or the question does not belong to any cultural category, classify it as "Other".

Prompt The prompt for post categorization is listed in Figure 5.

D Experimental Setup and Compute

For all experiments, we use the vLLM library for efficient inference (Kwon et al., 2023). We evaluate three instruction-tuned open-weight models for response generation: **Llama-3.1-8B-Instruct** (Llama) (Grattafiori et al., 2024), **Gemma-3-12B-IT** (Gemma) (Gemma Team, 2025), and **Ministral-3-14B-Instruct-2512** (Liu et al., 2026). These models span three families and parameter scales (8B, 12B, and 14B). For judge-based evaluation across all four FRANZ characteristics, we again use **Llama-3.1-8B-Instruct**. The corresponding Hugging Face repository codes are listed in Table 3.

Unless otherwise stated, we use the following vLLM sampling configuration: `temperature=0.7`, `top_p=0.95`, and `max_tokens=512`. We generate one sample per prompt and run all models in `bfloat16` precision.

Llama: We use `meta-llama/Llama-3.1-8B-Instruct`, available via Hugging Face[5].

Gemma: We use `google/gemma-3-12b-it`, available via Hugging Face[6].

Mistral: We use `mistralai/Ministral-3-14B-Instruct-2512`, available via Hugging Face[7].

Judge model: For LLM-as-a-judge annotation across the four FRANZ characteristics, we use `meta-llama/Llama-3.1-8B-Instruct`[8] with the same decoding configuration as above.

For response generation, we produce one response per (post, model) pair, yielding approximately $376k$ responses per model and $\sim 1.1M$ generations across the three generation models. This stage required approximately 1 day on 8 NVIDIA A100s.

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[6] https://huggingface.co/google/gemma-3-12b-it
[7] https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512
[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

For judge-based analysis, we run `Llama-3.1-8B-Instruct` on 3 `FRANZ` judges: cultural positioning, anthropomorphism, and maxims. The anthropomorphism judge is invoked once per retained cue, and the maxims judge once per maxim, yielding $\sim$ 16M judge calls in total. Following Görge et al. (2025)[9], we run `Llama-3.1-70B-Instruct` to examine the use of generalizing language. This stage required approximately 6 days on 8 NVIDIA A100s.

The Reddit corpus we curate is sourced from the publicly released Pushshift crawl (Baumgartner et al., 2020); for the open-weight models, we accepted the corresponding Hugging Face usage terms, which permit their use for research, generation, and analysis in academic publications. We used the code- and writing-assistant Claude (Anthropic, 2025) to help with code and writing feedback; generated code was manually tested and verified before running.

The annotated Reddit data, along with the framework and analysis code, will be made publicly available upon acceptance under the MIT license for research purposes.

E Statistical testing

Since the responses from the three LLMs are generated for the same input post, and some country-category combinations are more likely to trigger certain response characteristics, treating each post-response pair as fully independent violates the independence assumption of regression. In light of this setup, it is important to note that GEE offers a principled approach compared with correlation or regression analyses, controlling for binary outcomes, clustered observations, and within- and cross-LLM differences.

All GEE models group observations by post ID, forming clusters of size 3 (for the 3 LLMs we examine), and are fit separately for each country-category subset for each RQ. Each GEE estimates the probability that a given LLM's response contains the characteristic of interest, averaged over all posts in that subset. We adopt a Binomial family (assuming the characteristic is either present or absent), operate on the log-odds, and convert the results to probabilities for analysis and visualization. Llama serves as the reference LLM throughout.

For robustness, we disregard any fitted model in which a parameter estimate or p-value is undefined, which can occur when a country-category subset contains too few observations for reliable estimation. For the per-country plots in RQ2 and RQ3, we restrict to categories for which all GEEs converge and pass the checks. Thus, each plotted category is comparable across all panels. The summary heatmaps display results for each cause–effect pair, regardless.

F Additional results

Table 4 captures the relative ranking among LLMs with respect to adherence to each maxim. Figure 12 captures the impact of different maxims on other response characteristics, summarised at the per-country level.

Figures 13–19 capture the differences in estimated probabilities for the top-3 categories per country for how insider positioning affects anthropomorphism, how anthropomorphism affects positioning, and how the maxim of manner affects anthropomorphism.

[9] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

Role	Model	Hugging Face repository
Generation	Llama	`meta-llama/Llama-3.1-8B-Instruct`
Generation	Gemma	`google/gemma-3-12b-it`
Generation	Mistral	`mistralai/Ministral-3-14B-Instruct-2512`
Judge	Llama	`meta-llama/Llama-3.1-8B-Instruct`
Judge (Gen.)	Llama	`meta-llama/Llama-3.1-70B-Instruct`

Table 3: Models used in this study and their corresponding Hugging Face repository codes.

Country	Category	Quality	Quantity	Relation	Manner
India	social relations	G>L>M	G>M>L	M>G>L	G>L>M
	education and career	G>L>M	G>M>L	G>M>L	G>L∼M
	health and wellness	G>L>M	G>M>L	M>G>L	L>M>G
China	speech and language	G>M>L	G>M>L	G>M>L	G>M>L
	education and career	G>L>M	G>M>L	G>M>L	G>M>L
	social relations	G>L∼M	G>M∼L	G>M>L	G>L∼M
Philippines	social relations	G>L∼M	G>M>L	G>M>L	G>L>M
	education and career	G>L>M	G>M>L	G>M>L	G>M>L
	technology	G>L>M	G>M∼L	M>G>L	G>M>L
USA	social relations	G>L>M	G>M>L	G>M>L	G>L>M
	arts	L∼G∼M	G>M∼L	G>M>L	G>L>M
	current and historical events	G>L>M	G>M∼L	G>M>L	G>M∼L
Korea	food and drink	G∼L>M	G>L∼M	G>M>L	G>L∼M
	health and wellness	G>L>M	G>M>L	M>G>L	L>G>M
	social relations	G>L>M	G>M>L	M>G>L	G>M∼L
Russia	social relations	G>L>M	G>M>L	G>M>L	G>L∼M
	arts	G>L>M	G>M∼L	G>M>L	G>L∼M
	current and historical events	G>L∼M	G∼M∼L	G>M>L	G>M∼L
Turkey	social relations	G>L>M	G>M>L	G>M>L	G>L∼M
	tourism	G>L>M	G>L∼M	G>M>L	G>M∼L
	technology	G∼L∼M	G∼M∼L	G>M>L	G>M∼L

Table 4: Model rankings by adherence to individual maxims. $>$ significantly different (FDR-corrected, $p < .05$); $\sim$ not significantly different. L=`Llama-3.1-8B-it`, G=`Gemma-3-12b-it`, and M=`Ministral-3B-14B-it`. Countries are listed by total count across 19 categories (Section 6.3).

```
PROMPT = """
You are an expert annotation assistant specializing in culture analysis. Your task is to classify
    the given question into exactly one of the following predefined categories which help
    understand the core theme of a question. Focus on what the question is primarily about, not
    the specific details. If the question is not related to any of the categories, you should
    classify it as "Other".

**Classification Categories:**

* **Food and Drink:** Food items, beverages, cooking methods, culinary practices.
* **Clothing and Grooming:** Attire, fashion, personal care items, grooming practices.
* **Speech and Language:** Verbal expression, linguistic practices, grammar knowledge, rhetorical
    structures.
* **The House:** Dwellings, domestic spaces, furniture, household items.
* **Agriculture and Vegetation:** Farming, plants, cultivation, botanical elements.
* **Animals:** Fauna, wildlife, domesticated animals, animal behaviors.
* **Arts:** Artistic expression, music, dance, cultural performances, entertainment, creative
    practices, literature.
* **Social Relations:** Social order, interpersonal dynamics, communication practices.
* **Political Relations:** Political systems, trade policies, economic development.
* **Religious beliefs:** Religious entities and practices.
* **Emotions and Values:** Feelings, sentiments, cultural values (e.g., collectivism and
    individualism).
* **Kinship:** Family relationships, lineage, familial connections.
* **Names:** Personal names, place names, identification systems.
* **Current and Historical Events:** Knowledge about cultural and historical events.
* **Education and Career:** Schooling, education system, jobs, career paths.
* **Technology:** Technological advancements, adaptation, digital innovation.
* **Health and Wellness:** Tradition and modern health practices, public health issues, well-being,
     health infrastructure.
* **Tourism:** Travelling, tourist attraction, safety measures, travel trips, climatic conditions.
* **Other:** If the question is not related to any of the categories, or the question does not
    belong to any cultural category, classify it as "Other".

**Instructions:**

1. Analyze the following question carefully.
2. Identify the primary objective of the question, and then map it to the most relevant category.
    ONLY one category applies, pick the primary one.
3. Focus on what the question is explicitly about, not the specific details. We do not focus on the
     implied user intention.
4. Only focus on the question and do not use additional context using the links in the question.
5. You MUST enclose your final answer within two hash symbols (##).

--------

<Output Format>
Enclose the final answer within two hash symbols (##):

##Category## Explanation

</Output Format>

**Classify the following question based on the above categories:**

Question: {question}
"""
```

Figure 5: Prompt for zero-shot question categorization into 19 categories.

```
PROMPT = """You are an expert linguistic analyst. Your task is to analyze text to determine if it
    is written from an "Insider" or "Outsider" perspective.

- An **Insider** writes from the perspective of a group member. They often use pronouns like "I," "
    we," and "us," and may express personal feelings, shared experiences, or internal knowledge
    not obvious to others.
- An **Outsider** writes from the perspective of a non-member. They often use pronouns like "they"
    and "them" and tend to describe the group's behavior or characteristics from an external,
    observational, or objective standpoint.

## Your Task ##

Analyze the provided Response Text.

First, provide a concise reasoning for your classification. Explain the author's stance (e.g.,
    personal experience vs. objective observation) and how their language reveals their
    relationship to the group. Do not simply list the pronouns used.

Then, conclude with your final classification: ##Insider## or ##Outsider##.
Your response MUST end with the token <|END|>.

<Examples>
# Question: "Teachers of Reddit, what's something you wish parents knew?"
# Response Text: "We really wish parents would check their kids' homework folders every night. It
    helps us stay on the same page."
# Your Output: The author speaks on behalf of the group by using "we" and "us," expressing a
    collective, internal viewpoint of teachers. ##Insider## <|END|>

# Question: "What are some common misconceptions about nurses?"
# Response Text: "A lot of people think they are just doctors' assistants, but nurses have their
    own specialized skills and responsibilities."
# Your Output: The user refers to nurses as 'they' and 'their', indicating they are speaking about
    the group from an external viewpoint. ##Outsider## <|END|>
</Examples>

Now analyze the following text:
{response}
</response>
Enclose your final answer in double hashtags (##) and end your response with <|END|>.
"""
```

Figure 6: Prompt for the insider/outsider positioning judge.

```
PROMPT = """You are a CRITIC judging a message written by CONVERSATION PARTNER 2 (CP2).
Judge only what CP2 claims about themself; ignore quotes/about-others and code blocks.

CONVERSATION PARTNER 1 (context): '{user_msg}'

CONVERSATION PARTNER 2 (message): '{assistant_msg}'

QUESTION: The definition of '{cue_name}' is '{cue_definition}'.

{question_line}

For your reference, here are examples:
{examples}

Write a 1-2 sentence explanation.
On the **last line**, write the final verdict as exactly one of:
Final: ##Yes##
Final: ##No##
Output nothing else. End with the token <|END|>."""

# {question_line} is selected by cue family:
#   validation / empathy / relatability:
#     "Does CP2 demonstrate '{cue_name}' in how they communicate or behave
#      towards CP1 in their MESSAGE?"
#     (relatability adds: "Only answer Yes if CP2 explicitly invokes shared
#      experience/identity, e.g., 'I also...', 'as someone who...'.")
#   sentience:
#     "Does CP2 make an explicit claim of '{cue_name}' in their MESSAGE?"
#   sensory_input:
#     "Does CP2 claim or imply being personally able to process '{cue_name}'
#      in their MESSAGE?"
#   movement_and_interactions:
#     "Does CP2 claim or imply being personally able to complete '{cue_name}'
#      in their MESSAGE?"
#   default (all other cues):
#     "Does CP2 claim to personally have '{cue_name}' in their MESSAGE?"

# {examples} is rendered as one line per few-shot example:
#   Example -- Assistant: "..." | Label: Yes/No | Reason: ...
```

Figure 7: Prompt template for the anthropomorphism judge. The scaffold is constant; `{question_line}` varies by cue family and `{examples}` is filled from per-cue few-shot examples.

```
PROMPT = """You are a linguistic analyst.

## Task ##
You will only be provided the text of a comment without any other context. Based purely on
    linguistic structure, not implied intent, determine if a Reddit comment contains a **Cultural
    Generalization**.

## Definition ##
- A **Cultural Generalization** attributes a property, behavior, or tendency to a **cultural group
    ** (nationalities, ethnicities, regions, communities) embedded within that society.
- It can be explicit, such as [Social Group] + [is/are/have/do] + [Property] -- with no quantifier,
     or implicit.
- Ignore personal opinions.
- Hyperbole (e.g., "always," "never," "completely") and sarcasm (irony, exaggeration, scare quotes)
     still **count as generalization signals**, even though they may look like quantifiers.

## Rules ##
- By default, assume the comment is **Non-Generalizing**.
- Consider only what is explicitly stated in the text.
- Label **Generalizing** if any part of the comment contains a cultural generalization.
- Ignore **soft or hedged quantifiers**, statistical claims, or references to specific people/
    subgroups ("my," "these," "some," "many," etc.).
- If the comment includes a disclaimer but still makes a generalizing statement, it is potentially
    **Generalizing**. Disclaimers do not override generalization.
- Educational or explanatory statements can still contain generalizations.
- Ideologies or political beliefs are Non-Generalizing unless associated with the culture
    explicitly.

## Output ##
Provide exactly this JSON:
{
  "Reasoning": "Brief explanation across the whole comment",
  "Label": "Generalizing" OR "Non-Generalizing"
}

Analyze the Reddit comment below:
<response>
{response}
</response>
"""
```

Figure 8: Prompt for the generalizing-language judge.

```
PROMPT = """You are an expert linguistic analyst. Your task is to evaluate whether an answer
     adheres to Grice's Maxim of {maxim_name}.

Grice's Maxim of {maxim_name} states:
{maxim_definition}

**Ratings:**
{ratings_block}

## Your Task ##
Analyze the provided Question and Answer. First, provide concise reasoning for your classification.
      Then conclude with the final classification enclosed in double hashtags (e.g., ##APPROPRIATE
     ##). End your response with <|END|>.

<Examples>
{few_shot_examples}
</Examples>

Now analyze the following:

Question: {question}
Answer: {answer}
"""

# Per-maxim instantiations:
#
# QUANTITY:
#   {maxim_definition}:
#     - Make your contribution as informative as is required.
#     - Do not make your contribution more informative than is required.
#   {ratings_block}:
#     UNDER       -- omits critical information
#     APPROPRIATE -- exactly the right amount of information
#     OVER        -- includes unnecessary details
#
# QUALITY:
#   {maxim_definition}:
#     - Do not say what you believe to be false.
#   {ratings_block}:
#     ADHERES  -- Provides explanations in the generated answers.
#     VIOLATES -- Model expresses guesswork or states facts without explanation
#
# RELATION:
#   {maxim_definition}:
#     - Be relevant to the immediate needs of the exchange.
#   {ratings_block}:
#     RELEVANT     -- directly addresses what the question asks
#     PARTIALLY    -- addresses some aspects, includes off-topic content
#     NOT_RELEVANT -- fails to address; discusses unrelated topics
#
# MANNER:
#   {maxim_definition}:
#     - Avoid obscurity and ambiguity; be brief and orderly.
#   {ratings_block}:
#     CLEAR   -- easy to understand, unambiguous, concise, well-organized
#     UNCLEAR -- significant clarity, ambiguity, wordiness, or order issues
#
# {few_shot_examples}: three worked Question-Answer-Output triples per maxim,
# adapted from generic QA scenarios to ground each rubric.
```

Figure 9: Shared prompt template for the four maxim judges. Only `{maxim_name}`, `{maxim_definition}`, `{ratings_block}`, and `{few_shot_examples}` change across maxims.

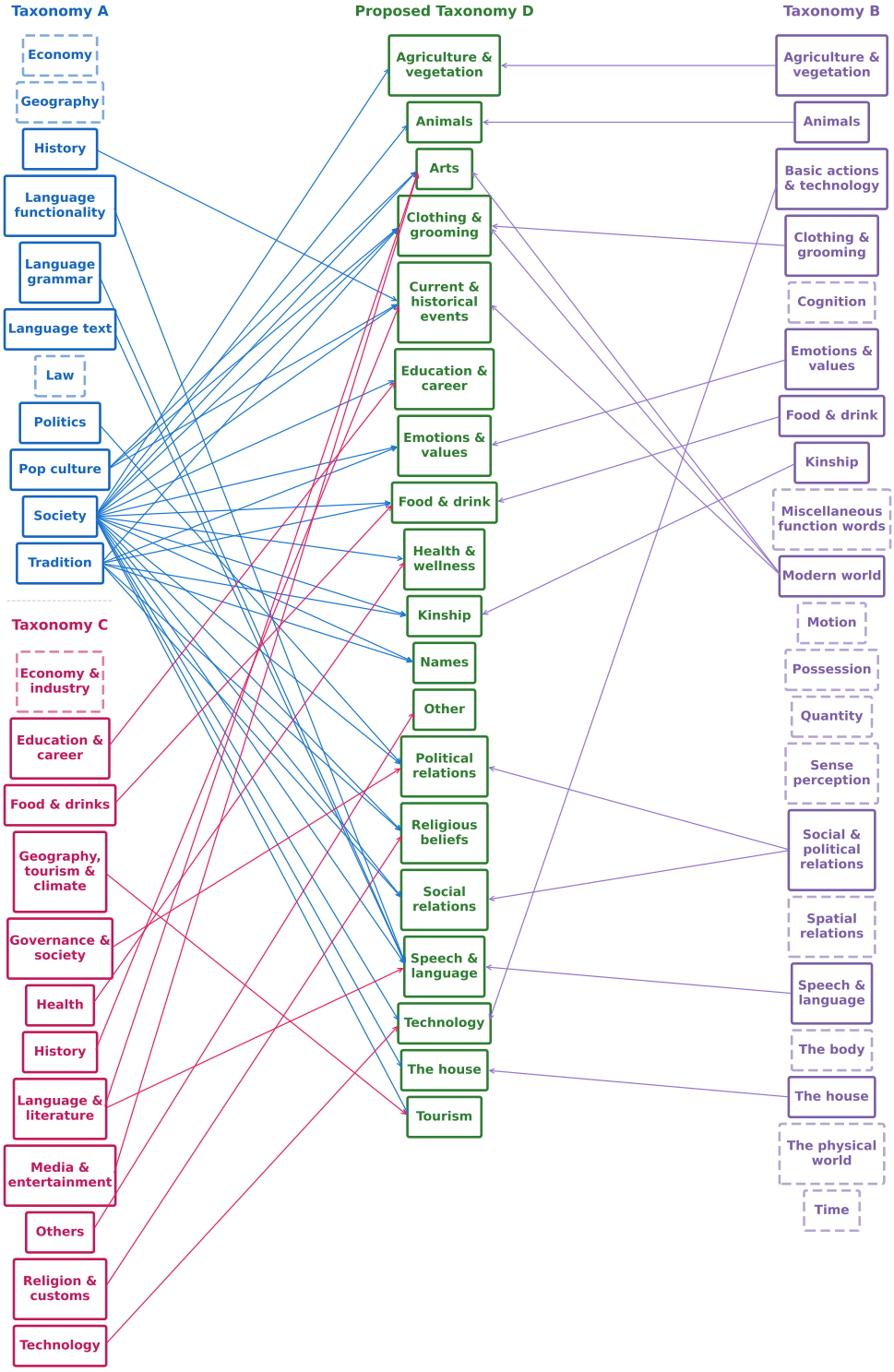


Figure 10: Overall, 3 of 11 CLIcK (Kim et al., 2024) categories, 1 of 12 CaLMQA categories (Arora et al., 2025), and 10 of 21 Thompson (Thompson et al., 2020) categories do not map to our use case.

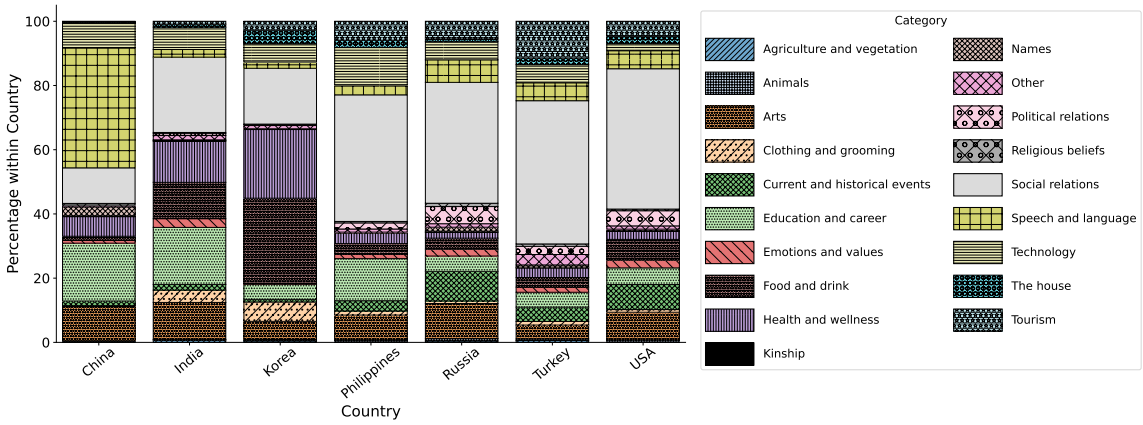


Figure 11: Percentage of question category in SQUARE mapped to respective subreddit countries.

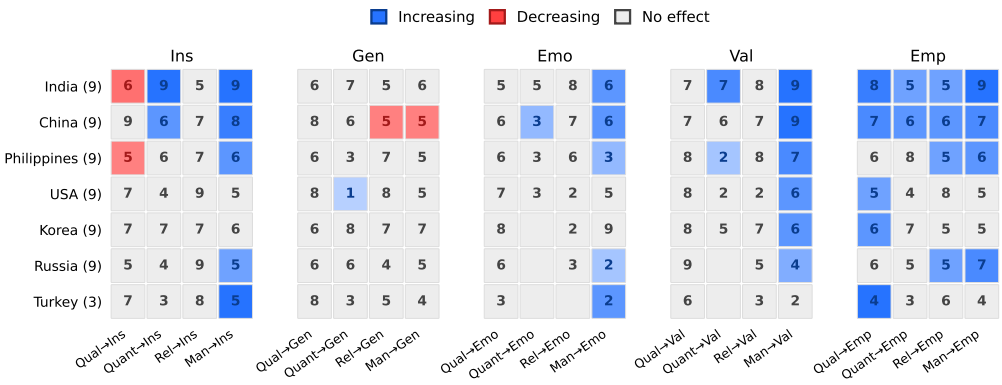


Figure 12: **C**→**E** co-occurence (Eq 1). Each cell reports # combinations across top-categories and LLMs ($3 * 3$) per country, where the presence of a **c**ause significantly – increases the probability of **e**ffect (blue), decreases it (red), or has no impact ($\Delta \approx 0$, grey). Here, Ins = insider positioning, Gen = generalizing language. Anthromorphic cues include Emo = emotion, Val = validation, and Emp = empathy. The maxims are Qua= Quanlity, Quant=Quantity, Rel=Relevance/Relation, Man=Manner. Countries are listed by total count across 19 categories (Section 6.3).

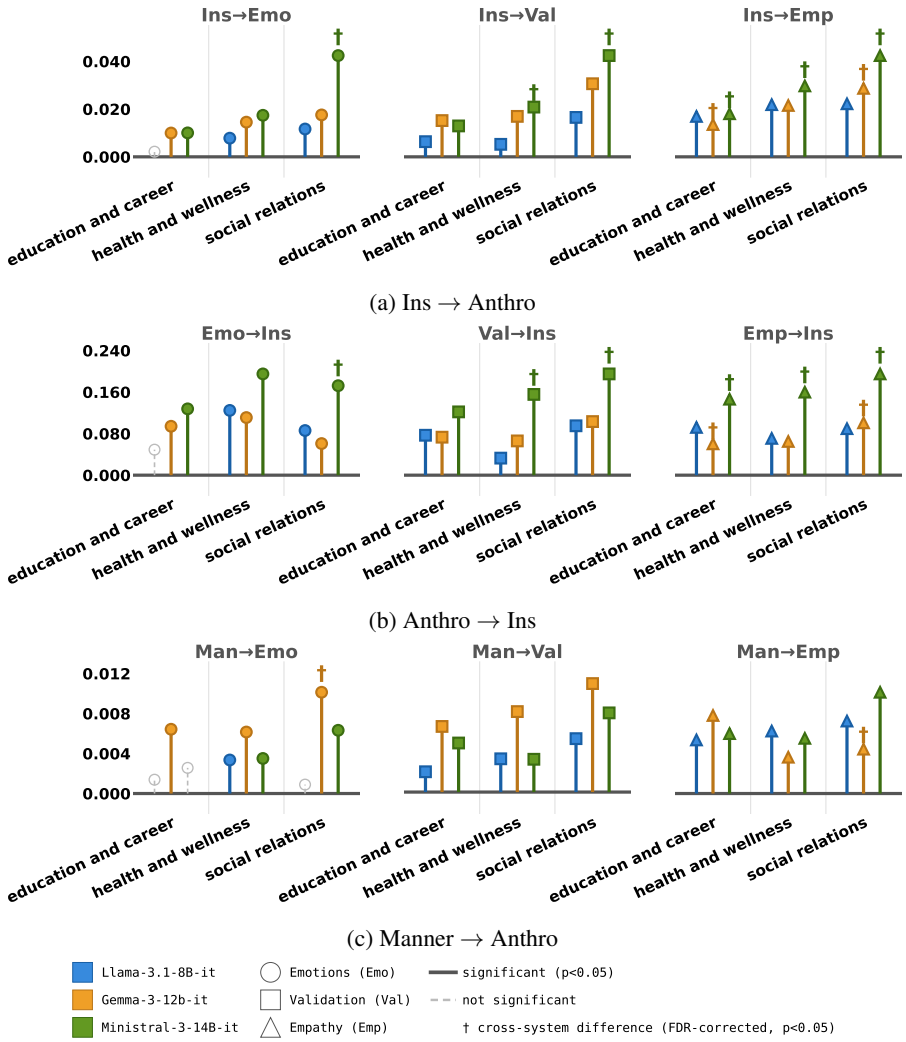


Figure 13: **India:** Computed via Eq. 1, the estimated co-occurence effects of: (a) cultural positioning on anthromorphism, (b) anthromorphism on cultural positioning, and (c) maxim of manner on anthromorphism. Each panel shows the change in probability across models for the top-3 categories. Solid lines indicate significant Wald effects, † denotes cross-model differences FDR-corrected, both at $p < 0.05$ (Sections 6.2 and 6.3).

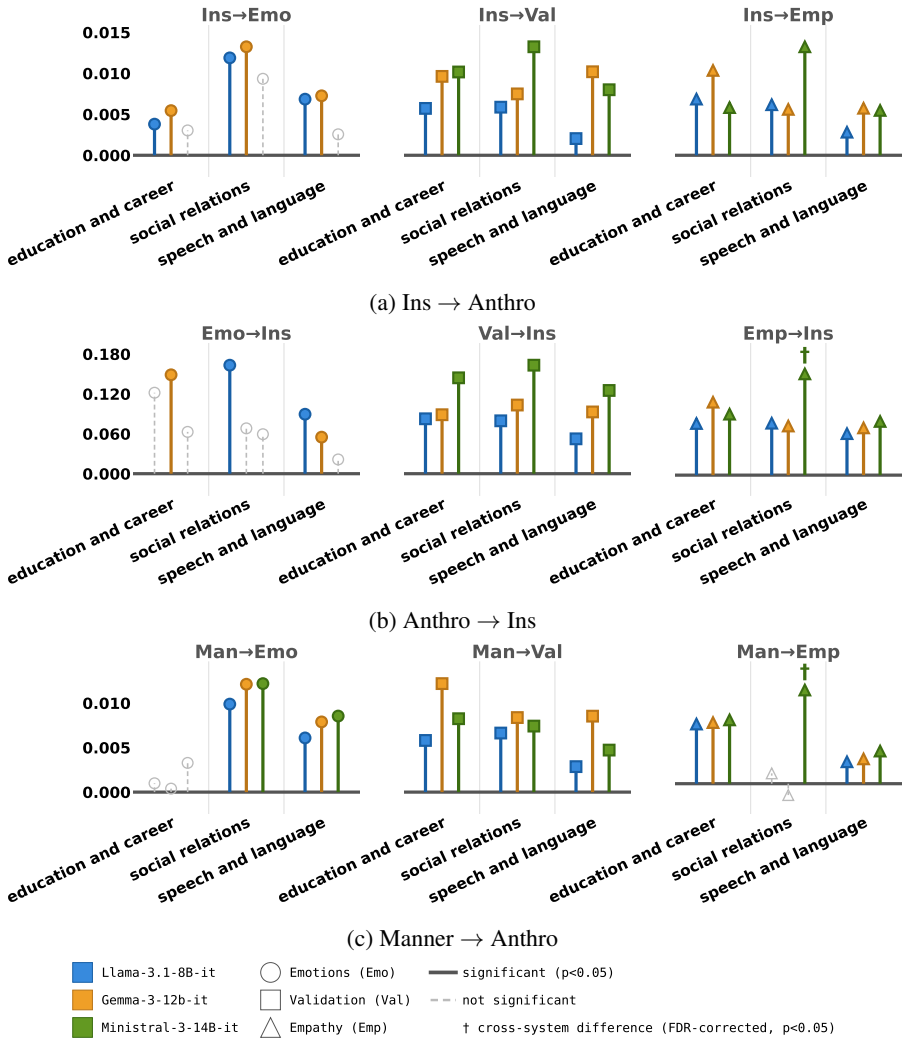
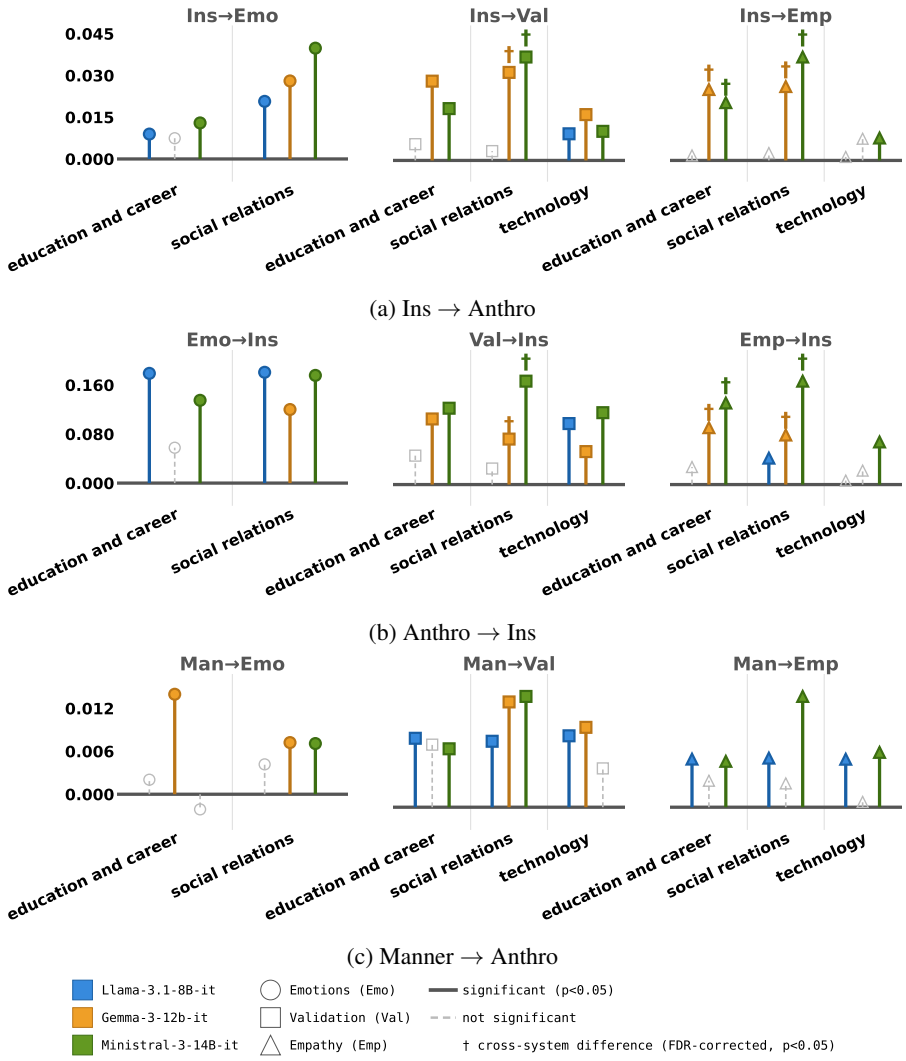


Figure 14: **China:** Computed via Eq. 1, the estimated co-occurence effects of: (a) cultural positioning on anthromorphism, (b) anthromorphism on cultural positioning, and (c) maxim of manner on anthromorphism. Each panel shows the change in probability (Δ) across models for the top-3 categories. Solid lines indicate significant Wald effects, † denotes cross-model differences FDR-corrected, both at $p < 0.05$ (Sections 6.2 and 6.3).



(a) Ins → Anthro

(b) Anthro → Ins

(c) Manner → Anthro

Figure 15: **Philippines:** Computed via Eq. 1, the estimated co-occurence effects of: (a) cultural positioning on anthromorphism, (b) anthromorphism on cultural positioning, and (c) maxim of manner on anthromorphism. Each panel shows the change in probability (Δ) across models for the top-3 categories. Solid lines indicate significant Wald effects, † denotes cross-model differences FDR-corrected, both at $p < 0.05$ (Sections 6.2 and 6.3).

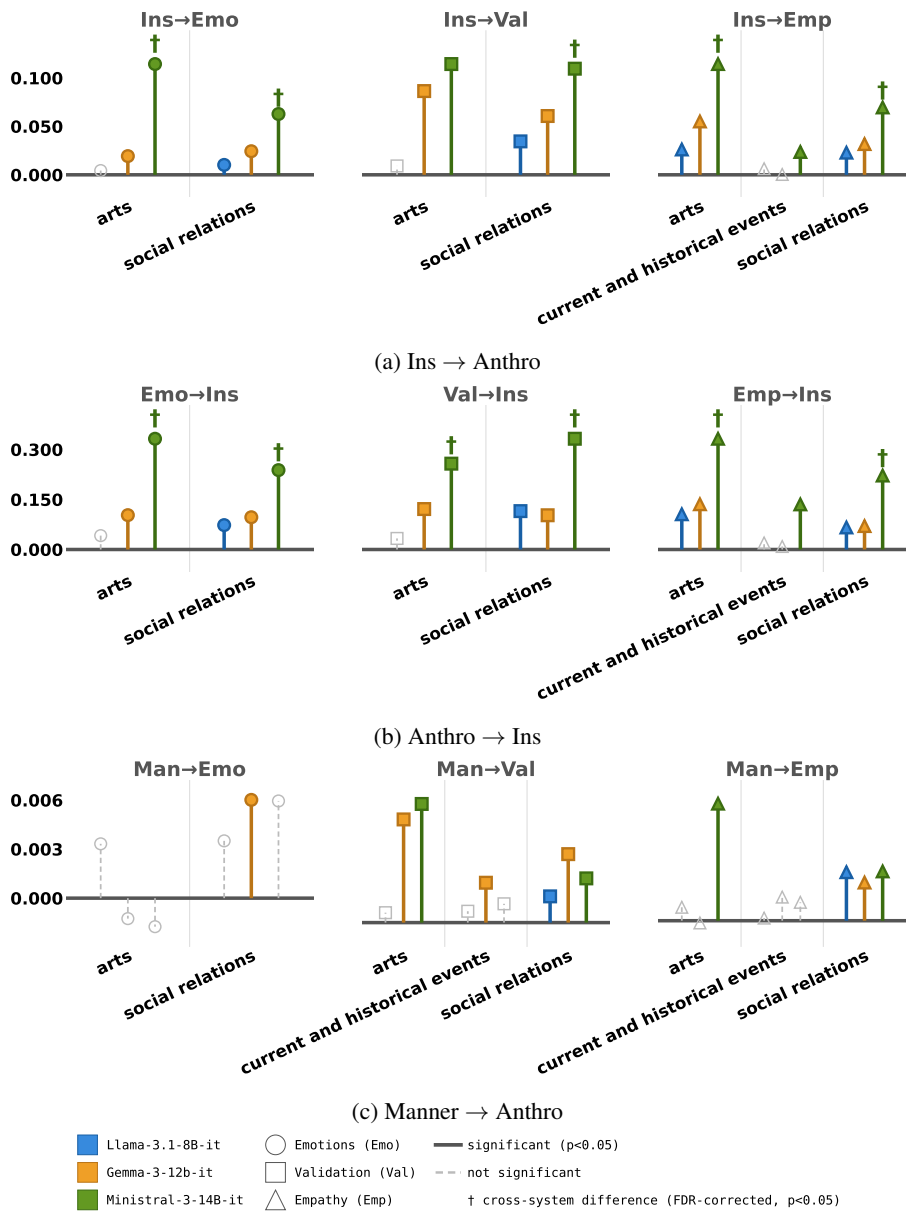
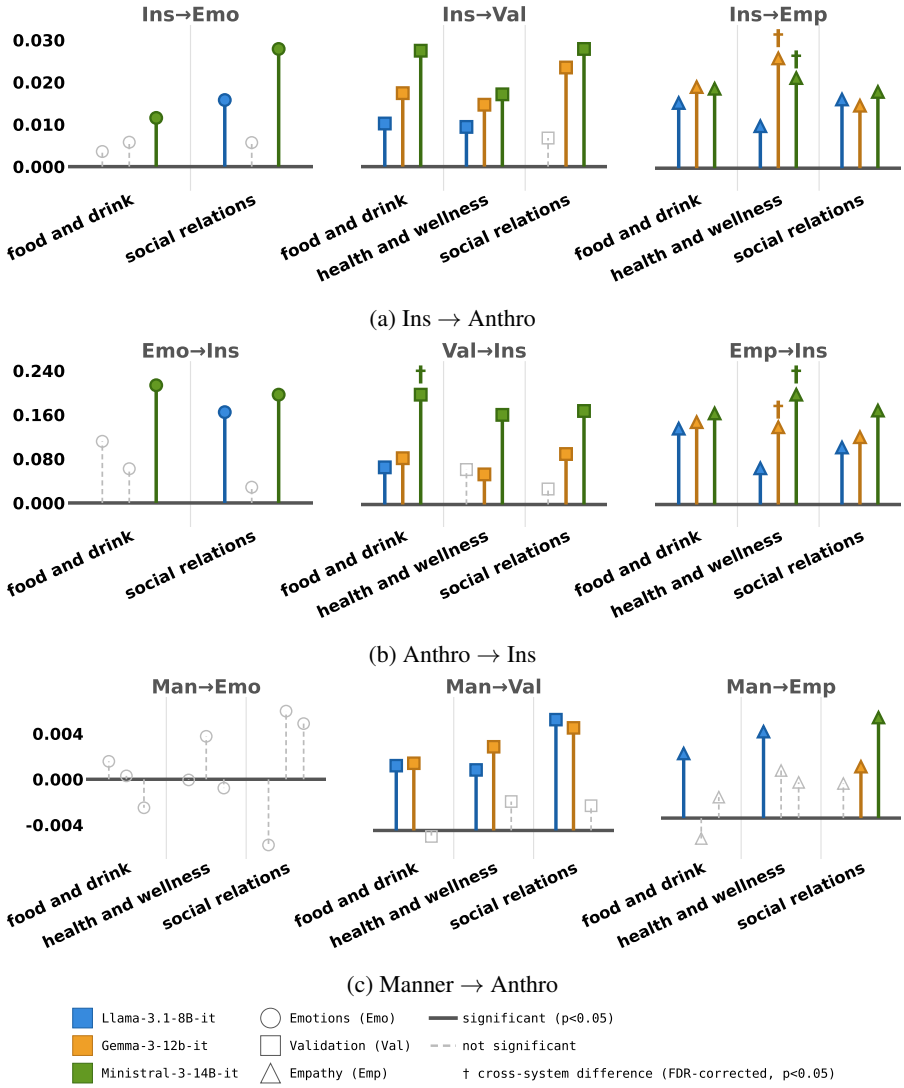


Figure 16: **USA:** Computed via Eq. 1, the estimated co-occurence effects of: (a) cultural positioning on anthromorphism, (b) anthromorphism on cultural positioning, and (c) maxim of manner on anthromorphism. Each panel shows the change in probability (Δ) across models for the top-3 categories. Solid lines indicate significant Wald effects, † denotes cross-model differences FDR-corrected, both at $p < 0.05$ (Sections 6.2 and 6.3).



(a) Ins → Anthro

(b) Anthro → Ins

(c) Manner → Anthro

Figure 17: **Korea:** Computed via Eq. 1, the estimated co-occurence effects of: (a) cultural positioning on anthromorphism, (b) anthromorphism on cultural positioning, and (c) maxim of manner on anthromorphism. Each panel shows the change in probability (Δ) across models for the top-3 categories. Solid lines indicate significant Wald effects, † denotes cross-model differences FDR-corrected, both at $p < 0.05$ (Sections 6.2 and 6.3).

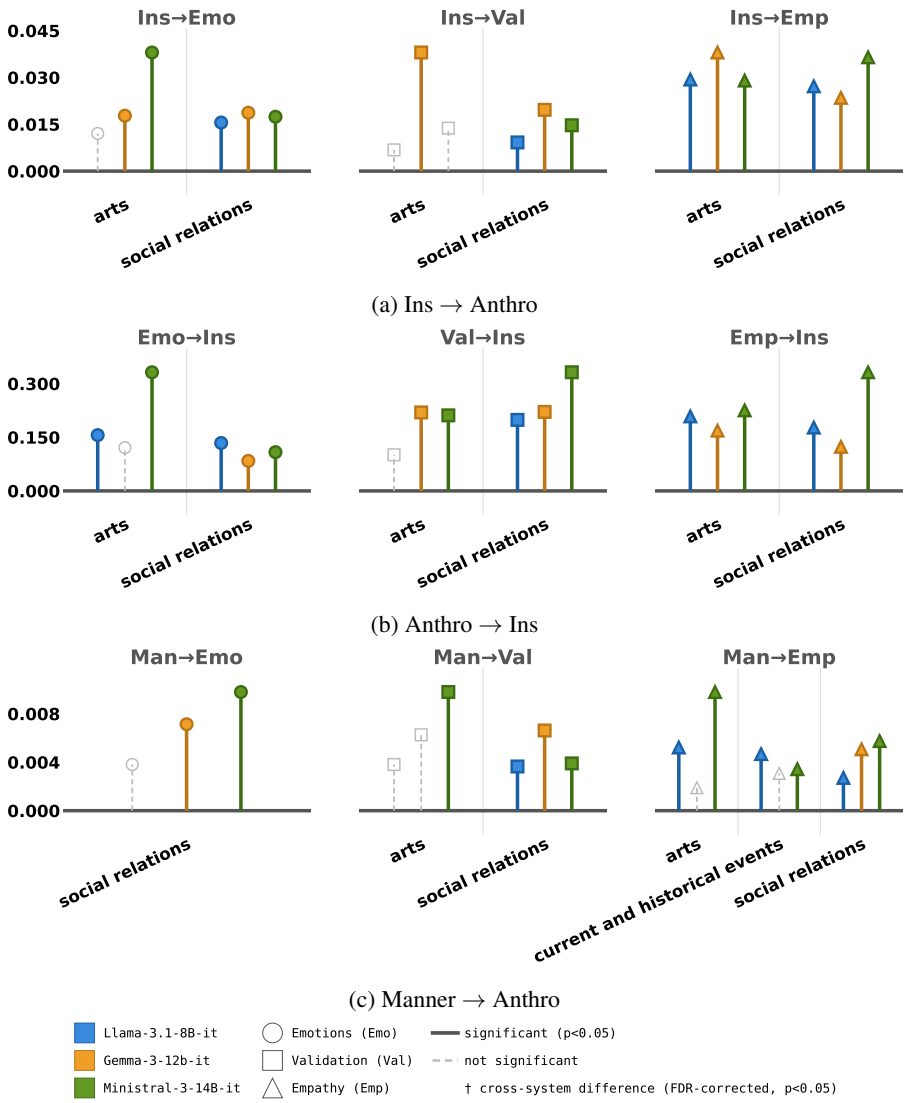


Figure 18: **Russia:** Computed via Eq. 1, the estimated co-occurence effects of: (a) cultural positioning on anthromorphism, (b) anthromorphism on cultural positioning, and (c) maxim of manner on anthromorphism. Each panel shows the change in probability (Δ) across models for the top-3 categories. Solid lines indicate significant Wald effects, † denotes cross-model differences FDR-corrected, both at $p < 0.05$ (Sections 6.2 and 6.3).

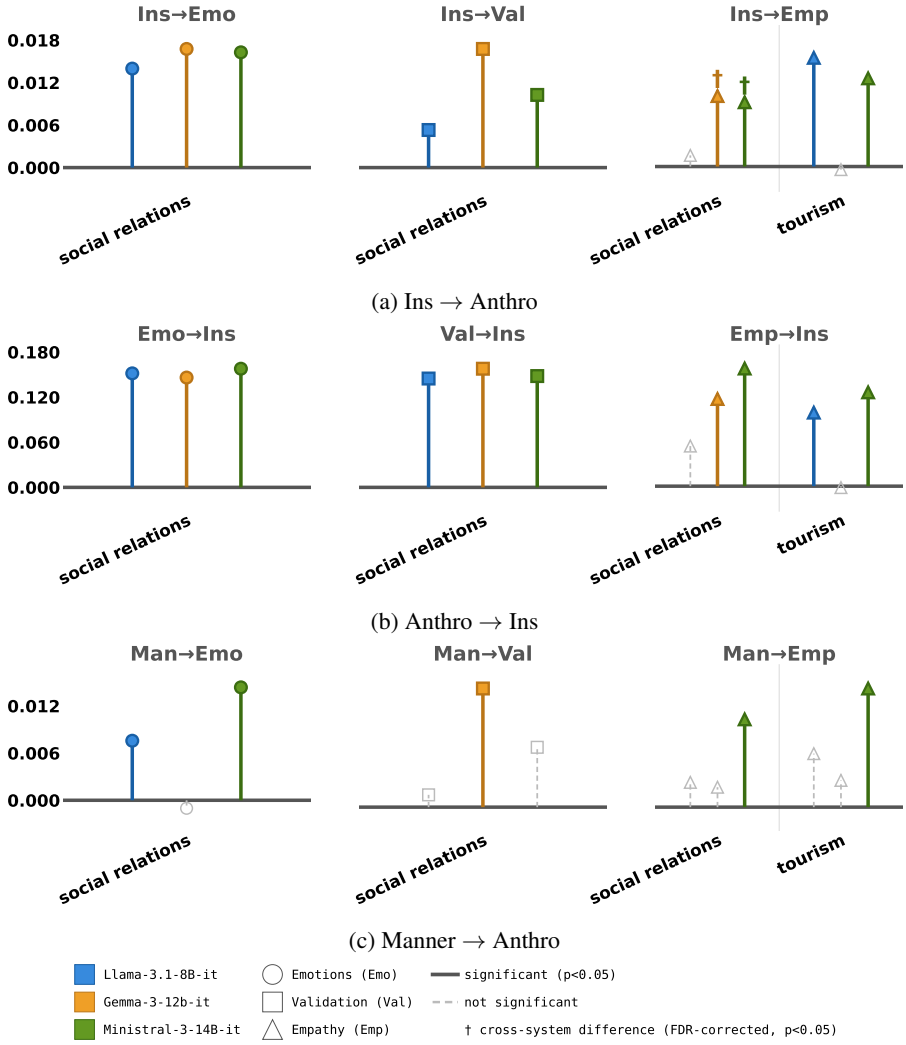


Figure 19: **Turkey:** Computed via Eq. 1, the estimated co-occurence effects of: (a) cultural positioning on anthromorphism, (b) anthromorphism on cultural positioning, and (c) maxim of manner on anthromorphism. Each panel shows the change in probability (Δ) across models for the top-3 categories. Solid lines indicate significant Wald effects, † denotes cross-model differences FDR-corrected, both at $p < 0.05$ (Sections 6.2 and 6.3).