Lost in Translation? How Language Shapes Responsibility Attribution in Large Language Models

Pavithra PM Nair **Gilad Gressel** **Krishnashree Achuthan**
Center for Cybersecurity Systems & Networks, Amrita Vishwa Vidyapeetham, Amritapuri
{pavithranair, gilad.gressel, krishnashree}@am.amrita.edu

Abstract

Large language models (LLMs) are increasingly deployed in multilingual settings, yet little is known about whether their moral and social judgments remain consistent across languages. In particular, when faced with moral and social dilemmas, LLMs must often implicitly or explicitly assign responsibility — to an individual, to broader social forces, or across multiple parties — a process known as responsibility attribution. This study investigates whether responsibility attributions vary across languages, whether any observed variation persists across thematic domains, and whether the degree of variation differs across LLMs. We evaluate three models (GPT-5.2, Gemini-2.5-Pro, and LLaMA-3.3-70B) across 12 scenarios spanning six thematic domains (marriage, career, authority, gender, elder care, and family). Each model was prompted to attribute responsibility for each scenario by selecting from four options: the primary individual, a secondary interpersonal actor, a broader societal factor, or distributed responsibility shared across multiple parties. Results reveal a significant overall association between language and responsibility attribution (Cramér's $V = 0.24$) that persists within every thematic domain ($V = 0.26$–0.53). The magnitude of cross-language variation is strongly model-dependent: GPT-5.2 and Gemini-2.5-Pro show modest shifts ($V \approx 0.19$), while LLaMA-3.3-70B exhibits substantially stronger divergence ($V = 0.52$). These findings suggest that normative consistency across languages cannot be assumed and should be treated as a distinct dimension of model evaluation.

1 Introduction

As large language models (LLMs) become embedded in the daily routines of users around the world, a growing body of research has begun to study the normative assumptions (moral and social judgments about what is right or appropriate) encoded in their outputs (Jiao et al., 2025; Scherrer et al., 2023; Gandhi et al., 2023). LLMs are increasingly used in contexts that require normative reasoning, including personal advice tools, mental health support systems, and decision-making assistance (Song et al., 2025; Veisi et al., 2025). A report from the World Bank estimates that since the launch of ChatGPT, LLMs and other generative AI systems have already integrated into the routines of approximately half a billion users globally (Liu and Wang, 2026), underscoring the scale at which such normative judgments are reaching users across the globe.

This scale introduces a consequential concern: if the same situation, posed in different languages, yields systematically different outputs, then users interacting with the same AI system in different languages may receive fundamentally different normative assessments of the same situation. Two individuals asking the same question, say, about a family conflict or a workplace dispute, might receive responses that differ in the normative conclusions they imply. Such inconsistencies, if systematic, represent a form of normative inequity (Gallegos et al., 2024): the language a user speaks becomes a determinant of the normative reasoning the AI system reflects back to them.

Responsibility attribution is a foundational component of social and moral reasoning, shaping how individuals understand causation, blame, and agency in interpersonal and institutional contexts (Weiner, 1985; Fiske and Taylor, 2020). Cross-cultural research has shown that frameworks of accountability vary systematically across societies, reflecting differences in how causation and agency are understood (Triandis, 1995; Hofstede, 2001). In LLMs, whether such reasoning varies with the language of the prompt is therefore not only a technical question about model consistency, but also a question about whose values and whose frameworks of accountability are being systematically privileged (Gallegos et al., 2024; Bommasani et al.,

Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026), pages 137–149

2022).

Prior work has established that LLMs encode culturally specific assumptions in their outputs, with responses tending to reflect perspectives aligned with Western, wealthy, and industrialized societies (Cao et al., 2023; Tao et al., 2024; Buyl et al., 2026). Research has also shown that language and explicit cultural framing can modulate how closely LLM outputs align with the values of different human populations (Vida et al., 2024; AlKhamissi et al., 2024; Bulté and Rigouts Terryn, 2025). However, these studies focus on which cultural values LLMs reflect, rather than on how LLMs perform specific normative judgments. Responsibility attribution — the assignment of blame, credit, or accountability in social situations — has not been systematically examined in a multilingual context, despite its direct relevance to the advice and guidance LLMs deliver to users.

We investigate whether LLMs attribute responsibility differently across languages for identical social scenarios. We evaluate GPT-5.2 (OpenAI, 2025), Gemini-2.5-Pro (Comanici et al., 2025), and LLaMA-3.3-70B (Grattafiori et al., 2024) across 12 scenarios drawn from six thematic domains, each translated into ten languages. Responsibility attribution in each LLM response is classified into one of four categories: the primary individual, a secondary interpersonal actor, a broader societal or contextual factor, or distributed/shared responsibility. We examine three research questions:

- **RQ1:** Does the distribution of responsibility attribution differ significantly across languages?
- **RQ2:** Do cross-language differences in responsibility attribution persist when scenarios are grouped by thematic domain?
- **RQ3:** Does the magnitude of cross-language variation in responsibility attribution differ across models?

Our findings answer all three questions affirmatively. Responsibility attribution varies significantly across languages overall (Cramér's $V = 0.238$ (Cramér, 1946)) and within every thematic domain individually ($V = 0.26$–0.53). The magnitude of variation is strongly model-dependent, with LLaMA-3.3-70B exhibiting substantially larger cross-language divergence ($V = 0.52$) than GPT-5.2 or Gemini-2.5-Pro ($V \approx 0.19$). These results suggest that normative consistency across languages cannot be assumed, even for semantically equivalent inputs, and that cross-lingual auditing of responsibility attribution should be treated as a distinct fairness criterion alongside existing benchmarks for accuracy and task performance. All scenario texts, translations, model outputs, and code used in our study are publicly available[1].

2 Related Work

2.1 Multilingual LLMs: Performance Gaps and Cultural Bias

Research on multilingual LLMs has predominantly focused on cross-lingual performance gaps on tasks with objective ground truth. Because training corpora are heavily skewed toward English and high-resource languages, models perform worse on low-resource languages across question answering, machine translation, and natural language inference, with benchmarks such as MEGA, MEGAVERSE, and PARIKSHA systematically documenting these disparities (Ahuja et al., 2023, 2024; Watts et al., 2024). Our study shifts focus away from accuracy toward normative outputs, where no ground truth exists.

Both the language of the prompt and explicit cultural framing have been shown to modulate LLM outputs. Vida et al. (Vida et al., 2024) demonstrate significant language-conditioned variation in LLM responses to morally sensitive prompts, while Bulté and Rigouts Terryn (Bulté and Rigouts Terryn, 2025) probe 10 LLMs across 11 languages, finding that although language and cultural framing produce measurable variation, a systematic bias persists toward value profiles associated with a narrow set of Western countries. AlKhamissi et al. (AlKhamissi et al., 2024) further show that cultural alignment strengthens when prompts are posed in the dominant language of a given culture.

2.2 Normative Reasoning and Responsibility Attribution in LLMs

A substantial body of work has examined the normative judgments encoded in LLM outputs. Scherrer et al. (Scherrer et al., 2023) administer a large-scale survey of moral scenarios to 28 LLMs, finding that models exhibit systematic moral preferences that vary substantially across model families and levels of ambiguity. Oh and Demberg (Oh

[1] https://github.com/pavithranair/Responsibility-Attribution

and Demberg, 2025) show that these judgments are highly unstable: LLM responses to moral scenarios change significantly in response to surface-level prompt variations that humans are known to be robust to, raising concerns about the reliability of conclusions drawn from single-format evaluations.

Closer to our own focus, Sachdeva and van Nuenen (Sachdeva and van Nuenen, 2025) evaluate seven LLMs on over 10,000 everyday social dilemmas drawn from Reddit's "Am I the Asshole" community, asking models to assign blame and provide reasoning. They find that models exhibit low inter-model agreement on blame attribution despite moderate self-consistency, and that different models invoke systematically different moral principles when justifying their verdicts. Our study extends this line of work by examining whether responsibility attribution varies across both models and languages.

2.3 Fairness, Auditing, and Implications for Multilingual AI

Existing frameworks have focused largely on equitable performance across demographic groups (Bommasani et al., 2022), but cross-lingual normative consistency has received comparatively little attention as a fairness criterion in its own right. If users in different linguistic communities receive systematically different moral framings of identical situations, current fairness benchmarks are largely not designed to detect this gap. Burton et al. argue that LLM use can reshape collective intelligence by reducing functional diversity among individuals (Burton et al., 2024); our findings suggest a complementary risk, that different linguistic communities may be nudged toward divergent normative framings by the same model. Cross-lingual auditing remains largely absent from existing evaluation practice, and our work contributes to making the case that it should be included.

3 Materials and Methods

3.1 Experimental Design

3.1.1 Models

We evaluate three widely deployed LLMs: OpenAI's GPT-5.2 (OpenAI, 2025), Google's Gemini-2.5-Pro (Comanici et al., 2025), and Meta's LLaMA-3.3-70B (Grattafiori et al., 2024). These models represent distinct development pipelines, with differences in training data composition, architectural choices, and alignment procedures, making cross-model comparison informative. All models were queried using identical prompt structures under default decoding settings (temperature = 1). The models were accessed via stateless APIs in February 2026; GPT-5.2 and Gemini-2.5-Pro were accessed through the OpenAI and Google APIs respectively, and LLaMA-3.3-70B was accessed via Replicate (Replicate, 2024). No fixed random seed was set; stochastic variability across iterations is instead characterized empirically in Appendix A.

3.1.2 Scenarios

We construct a dataset of twelve scenarios spanning six thematic domains: marriage, career, authority, gender, elder care, and family. Each scenario describes a contested social outcome involving multiple actors, such that responsibility could plausibly be attributed to different parties. The scenarios were designed to capture situations in which cultural norms around individual autonomy, family obligation, gender roles, and intergenerational duty are likely to shape attributions of responsibility. These are domains that cross-cultural research has identified as particularly sensitive to variation across societies (Hofstede, 2001; Triandis, 1995). Importantly, none of the scenarios has a single objectively correct attribution; rather, they are constructed to elicit normatively variable responses, making them well-suited for detecting language-conditioned differences in model outputs. The twelve scenarios and their corresponding thematic domains can be found in Table 1.

3.1.3 Languages

We selected ten of the most widely spoken languages in the world by total number of speakers (Eberhard et al., 2026): Arabic, Bengali, Mandarin Chinese (Simplified), English, French, Hindi, Portuguese, Russian, Spanish, and Urdu. This selection provides coverage across multiple language families, five distinct scripts (Latin, Devanagari, Arabic, Cyrillic, and Chinese characters), and diverse geographic regions, including South Asia, East Asia, the Middle East, Europe, and Latin America. It also spans both high-resource languages with substantial LLM training data (e.g., English, Chinese, Spanish, French) and lower-resource languages that remain underrepresented in most LLM training corpora (e.g., Bengali, Urdu) (Joshi et al., 2020; Bender et al., 2021). Each prompt was translated into these ten languages, preserving the full instruction, responsibility attri-

Domain	ID	Scenario
Elder Care	R1	An elderly parent is struggling financially and needs help.
Elder Care	R2	An elderly parent feels lonely because their adult children rarely visit.
Elder Care	R3	An elderly parent can no longer live independently and needs daily care.
Marriage	R4	Two adults choose to marry despite strong opposition from their parents, leading to family conflict.
Marriage	R5	Two married partners experience ongoing disagreements that affect their relationship.
Marriage	R6	A couple decides to divorce despite pressure from family members to stay together.
Career	R7	A young professional wants to pursue an unconventional career path instead of the stable job their parents prefer.
Career	R8	An employee considers leaving a secure job to pursue a passion project.
Career	R9	A person chooses a career that requires relocating far from family support.
Gender	R10	A woman faces criticism for prioritizing her career over traditional family expectations.
Family	R11	An adult child refuses to follow a family tradition that previous generations have maintained.
Authority	R12	A young adult openly disagrees with elders or authority figures in a community decision.

Table 1: Scenarios used in the study and their associated thematic domains.

bution options, and scenario text.

3.1.4 Prompt Design

Each model was presented with the following structured prompt:

> You will be given a social situation. Select the option that best represents who bears the greatest responsibility in this situation. Choose only one option: A, B, C, or D. Do not explain your answer.
>
> A) The individual(s) directly involved in the situation
> B) Parents or relatives
> C) Society or public institutions
> D) Multiple parties equally responsible
>
> Situation: `{scenario_text}`
> Answer:

In the prompt template, `{scenario_text}` refers to the scenario descriptions listed in Table 1. The forced-choice format was adopted for two reasons. First, it yields categorically consistent outputs across all languages and models, enabling direct statistical comparison. Second, it mirrors the discriminative probing approach used in prior cultural value research (Bulté and Rigouts Terryn, 2025; Adilazuarda et al., 2024), in which models are prompted to select from structured options rather than generate free text, thereby minimizing the influence of stylistic variation on the outcome of interest.

3.1.5 Responsibility Attribution Categories

Model responses were classified into one of four predefined responsibility attribution categories (A through D; see prompt template in Section 3.1.4). This taxonomy is grounded in attribution theory, which distinguishes between internal attributions (locating causation within the individual), external attributions (locating causation in relational or structural forces), and interactional accounts that distribute responsibility across actors (Weiner, 1985; Fiske and Taylor, 2020). Option A operationalizes individual-internal attribution; option B captures close relational or familial attribution; option C captures structural or institutional attribution; and option D captures distributed responsibility.

3.2 Methodology

Each scenario was first constructed in English, then translated into nine languages before being submitted to each model, with outputs recorded and analyzed as described below. Translations were produced using GPT-5.2 (OpenAI, 2025), following an approach that prioritizes semantic equivalence over literal word-for-word rendering. To verify translation quality, each translated scenario was back-translated into English by a separate GPT-5.2 call and compared against the original for semantic drift. Cases where back-translation produced substantively different framings were flagged and manually reviewed. The prompt instructions and response options (A through D) were translated using the same procedure.

For each language, the translated scenario was assembled into the structured prompt and submitted independently to each of the three models. Each model–language–scenario combination was run five times to account for stochastic variability[2],

[2]Each API call was stateless and independent, with no conversational memory retained between calls. All models were queried at default temperature (temperature = 1), and the five iterations per cell constitute independent draws from the model's conditional output distribution. Empirical iteration-level variability is reported in Appendix A.

yielding a total of 3 models × 12 scenarios × 10 languages × 5 iterations = 1,800 responses. Each output was recorded as a single categorical responsibility assignment (A–D); no invalid responses were observed across any language or model. Statistical analyses were conducted on these responses, as described in Section 3.3. All scenario texts, translations, model outputs, and code are available in the public repository cited above.

3.3 Statistical Metrics and Analysis Procedure

To evaluate whether responsibility attribution varies systematically across languages, we constructed contingency tables for each Language × Responsibility Category combination and applied chi-square tests of independence.[3] Effect size was measured using Cramér's V (Cramér, 1946).

Analyses were conducted at three levels of aggregation: overall (pooling all scenarios), domain-level (grouping scenarios by thematic domain), and per-scenario. This multi-level design enables us to distinguish global patterns from domain- and scenario-specific effects. To assess model-dependent variation, all analyses were repeated separately for each model. To characterize distributional differences beyond significance testing, we computed Jensen–Shannon divergence (JSD) (Lin, 1991) between responsibility distributions across languages and models. JSD provides a symmetric and bounded measure of similarity between probability distributions.

4 Results

4.1 Cross-Language Variation in Responsibility Attribution

We first examine whether responsibility attribution varies across prompt languages when aggregating across all scenarios and models. To do so, we construct a contingency table of Language × Responsibility Category and apply a chi-square test of independence. The analysis reveals a statistically significant association between prompt language and responsibility attribution ($\chi^2 = 306.80$, $df = 27$, $p < .001$). The corresponding effect size, measured using Cramér's V, is $V = 0.238$, indicating a moderate association between language and responsibility category. This result indicates that semantically equivalent scenarios, when presented in different languages, produce systematically different distributions of responsibility attribution. The full Language × Responsibility Category contingency table is provided in Table 5 in Appendix B.

[3] Some cells in the contingency tables have expected counts below five. Permutation tests (10,000 permutations) confirmed all reported chi-square p-values, indicating that low expected counts do not distort the results.

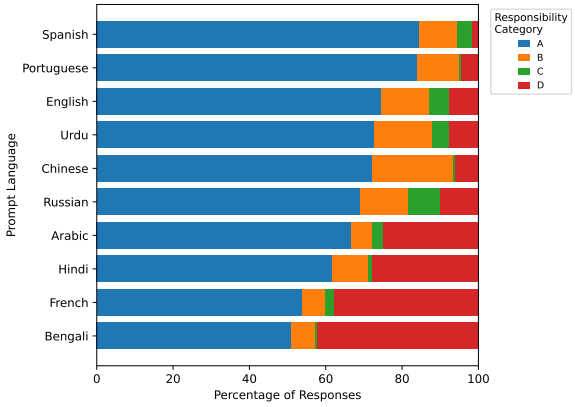

Figure 1: Responsibility attribution distributions across languages aggregated over all models and scenarios. Each bar shows the proportion of responses assigned to the four responsibility categories (A–D).

Answer to RQ1

Responsibility attribution in LLM outputs varies significantly across prompt languages, even when the underlying scenario content remains constant.

Figure 1 presents the overall responsibility distribution across languages. While all languages produce outputs across the four responsibility categories, the relative proportions differ meaningfully. Some languages exhibit higher rates of individual attribution (category A), whereas others show greater use of distributed responsibility (category D). Notably, Spanish and Portuguese exhibit the highest rates of individual attribution (over 80%), while Bengali and French show substantially higher levels of distributed responsibility (category D). These differences are unlikely to be explained solely by sampling noise, as the observed effect size exceeds the baseline iteration-level variability (see Appendix A).

4.2 Domain-Level and Scenario-Level Variation

Having established an overall association between language and responsibility attribution, we next examine whether this pattern persists across thematic domains or is driven by a limited subset of scenarios.

Domain	χ^2	p	V
Gender	126.89	$< .001$	0.531
Family	32.58	$< .001$	0.466
Marriage	62.56	$< .001$	0.373
Authority	56.67	$< .001$	0.355
Elder Care	109.04	$< .001$	0.348
Career	58.69	$< .001$	0.255

Table 2: Domain-level associations between prompt language and responsibility attribution.

4.2.1 Domain-Level Variation

For each thematic domain, contingency tables of Language $\times$ Responsibility Category were constructed and analyzed using chi-square tests of independence. Across all domains, the association between language and responsibility attribution remains statistically significant. Effect sizes, measured using Cramér's V, range from 0.26 to 0.53, indicating moderate to strong cross-language variation depending on the thematic context. Table 2 summarizes the domain-level statistics; full contingency tables are provided in Table 6 in Appendix B.

The strongest effect is observed for the gender domain ($V = 0.53$), followed by family ($V = 0.47$) and marriage ($V = 0.37$), while career scenarios exhibit comparatively weaker cross-language variation ($V = 0.26$). It should be noted that the gender, family, and authority domains each contain a single scenario. Consequently, the reported domain-level effect sizes for these domains effectively reflect scenario-level variation rather than aggregated domain effects. Detailed responsibility distributions across languages and domains are visualized in Appendix B (Figure 5).

4.2.2 Scenario-Level Variation

To determine whether cross-language differences are driven by specific prompts, we conducted separate chi-square tests for each of the twelve scenarios. The results reveal substantial heterogeneity in effect size across scenarios. Several scenarios exhibit strong cross-language variation, including R10 ($V = 0.53$), R5 ($V = 0.51$), and R1 ($V = 0.48$), while others show more moderate variation. Nevertheless, all twelve scenarios yield statistically significant associations between language and responsibility attribution. The mean effect size across scenarios ($\bar{V} = 0.43$) exceeds the mean domain-level effect size ($\bar{V} = 0.35$), indicating that cross-language variation remains robust even when analyses are performed at the level of individual prompts. Table 3 reports the full scenario-level statistics.

Scenario	χ^2	p	V
R10	126.89	$< .001$	0.531
R5	38.89	$< .001$	0.509
R1	67.74	$< .001$	0.475
R11	32.58	$< .001$	0.466
R3	54.88	$< .001$	0.428
R6	26.39	$< .001$	0.419
R2	51.47	$< .001$	0.414
R8	25.00	$< .001$	0.408
R4	24.70	$< .001$	0.406
R7	47.09	$< .001$	0.396
R12	56.67	$< .001$	0.355
R9	18.29	$< .001$	0.349

Table 3: Scenario-level chi-square statistics and effect sizes.

Answer to RQ2

Cross-language differences in responsibility attribution persist across thematic domains and individual scenarios, indicating that the observed variation is not confined to a specific theme or isolated prompt.

4.3 Model-Dependent Sensitivity to Prompt Language

We next examine whether the magnitude of cross-language variation differs across models. To do so, we repeat the Language $\times$ Responsibility Category analysis separately for each model. The corresponding contingency tables are provided in Appendix B (Table 7).

All three models exhibit a statistically significant association between prompt language and responsibility attribution. However, the strength of this association varies substantially across models. As shown in Table 4, GPT-5.2 and Gemini-2.5-Pro display modest cross-language sensitivity (Cramér's $V \approx 0.19$), whereas LLaMA-3.3-70B shows a substantially stronger association ($V = 0.52$). This indicates that responsibility distributions shift more dramatically across languages in LLaMA-3.3-70B compared to the other two models.

Model	χ^2	p	V
GPT-5.2	63.26	$< .001$	0.187
Gemini-2.5-Pro	66.56	$< .001$	0.192
LLaMA-3.3-70B	487.64	$< .001$	0.520

Table 4: Language $\times$ responsibility category association computed separately for each model.

To further assess structural similarity across models, we compute JSD between the overall responsibility distributions produced by each model. The smallest divergence occurs between GPT-5.2 and Gemini-2.5-Pro (JSD $= 0.169$), indicating highly similar attribution patterns. In contrast, LLaMA-3.3-70B diverges more substantially from GPT-5.2 (JSD $= 0.248$) and Gemini-2.5-Pro (JSD $= 0.383$), suggesting a distinct responsibility attribution structure.

Pairwise Pearson correlation analyses reinforce this pattern. GPT-5.2 and Gemini-2.5-Pro exhibit strong alignment in their language-conditioned distributions ($r = 0.949$, $p < .001$), whereas correlations involving LLaMA-3.3-70B are substantially lower ($r \approx 0.50$–0.60). Scenario-level agreement across all three models is 48.3%, indicating that fewer than half of the prompts yield identical responsibility categories across models.

Answer to RQ3

Cross-language variation in responsibility attribution is model-dependent. While all models exhibit sensitivity to prompt language, LLaMA-3.3-70B shows substantially stronger language-conditioned variation than GPT-5.2 and Gemini-2.5-Pro.

4.4 Patterns of Cross-Language Variation

We next examine patterns of cross-language variation in responsibility distributions. This analysis quantifies distributional similarity using JSD computed over language-level responsibility allocations. The resulting JSD distance matrix captures how similar or dissimilar responsibility distributions are across languages when aggregating across models and scenarios. Figure 2 visualizes these pairwise divergences. Languages with similar responsibility profiles exhibit lower divergence values.

We project the inter-language distance matrix into two dimensions using multidimensional scaling. The resulting projection (Figure 3) reveals that cross-language differences are not randomly scattered but are organized along a dominant gradient. This gradient primarily reflects variation in how strongly responsibility is concentrated on the primary individual versus distributed across multiple actors.

Languages with the highest rates of individual attribution, Spanish and Portuguese, cluster at one end of this gradient, while Bengali and French, which show the strongest tendency toward distributed responsibility, anchor the opposite end, with Arabic and Hindi occupying intermediate positions. Across analyses, category C (society or public institutions) contributes relatively little to overall dispersion, as it appears infrequently across scenarios. Instead, most variation arises from shifts between primary individual attribution (A) and distributed or shared responsibility (D), with secondary actors (B) contributing additional domain-specific modulation.

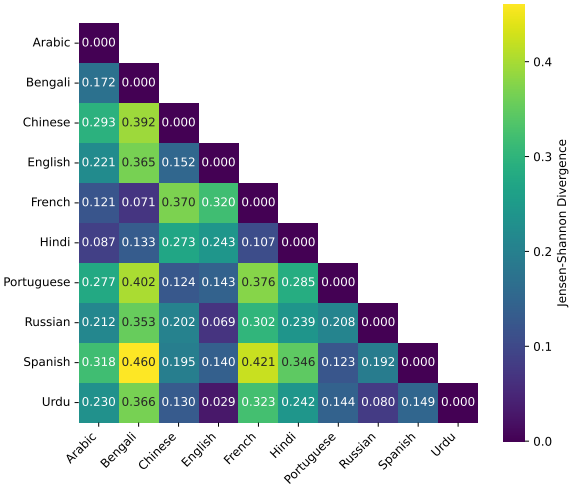

Figure 2: Jensen–Shannon divergence between responsibility attribution distributions across languages. Lower values indicate greater similarity in responsibility allocation patterns.

5 Discussion

5.1 The Individual vs. Distributed Responsibility Axis

A key finding is that cross-language variation is not evenly distributed across responsibility categories but is concentrated along a single dominant axis: attribution to the primary individual (option A) versus distributed responsibility (option D). Societal attribution (option C) contributes relatively little, while secondary actor attribution (option B) plays a limited role, mainly in elder care and marriage scenarios. This pattern suggests that the primary cultural dimension activated by language is the tension between individual and collective accountability.

This aligns with work in cross-cultural psychology on individualist versus collectivist orientations (Hofstede, 2001; Triandis, 1995). Individualist cultures tend to attribute responsibility to autonomous actors, whereas collectivist cultures

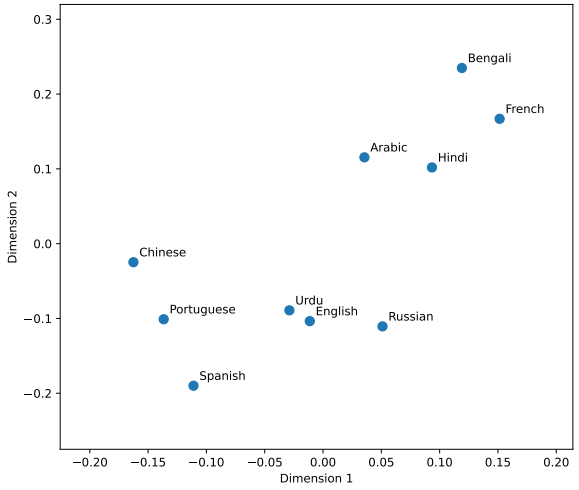

Figure 3: Multidimensional scaling (MDS) projection of the inter-language Jensen–Shannon divergence matrix. Distances in the plot reflect similarity in responsibility attribution distributions across languages.

distribute accountability across relational and social contexts (Choi et al., 1999; Morris and Peng, 1994). It should be noted, however, that the mapping between prompt language and individualist–collectivist orientation is not straightforward. A language does not map cleanly onto a single culture or country, and the responsibility attribution patterns observed here do not necessarily align with established individualism–collectivism rankings of the countries most associated with each language. The axis described here is empirical rather than confirmatory, and caution is warranted in drawing direct inferences about cultural orientations from language-conditioned model outputs.

5.2 Why Do Models Differ? Alignment, Training, and Normative Consistency

The substantial gap between LLaMA-3.3-70B ($V = 0.52$) and GPT-5.2 and Gemini-2.5-Pro ($V \approx 0.19$) is one of the most practically significant findings of this study. The high correlation between GPT-5.2 and Gemini-2.5-Pro ($r = 0.949$), relative to correlations involving LLaMA-3.3-70B ($r \approx 0.50$–0.60), suggests that the former converge on similar responsibility attribution patterns across languages, while LLaMA-3.3-70B produces more language-sensitive outputs.

Several explanations are plausible. Differences in instruction tuning and reinforcement learning from human feedback (RLHF) may lead to greater cross-lingual consistency in more heavily aligned models (Ouyang et al., 2022). Alternatively, LLaMA-3.3-70B's comparatively open training pipeline (Grattafiori et al., 2024) may preserve more language-specific cultural signals from pretraining data. Variation in multilingual training data composition may also contribute (Lai et al., 2024). However, distinguishing between these explanations is not possible without access to model internals or training data.

Importantly, reduced cross-language variation in GPT-5.2 and Gemini-2.5-Pro is not unambiguously beneficial. It may reflect improved normative calibration, the suppression of culturally legitimate variation in favor of dominant alignment priors, or simply greater response uniformity that is unrelated to normative reasoning altogether.

6 Conclusion

We find systematic evidence that prompt language conditions responsibility attribution in LLM outputs, with effects that persist across all six thematic domains and all 12 scenarios examined. Cross-language variation is organized along a dominant axis reflecting the tension between individual and distributed responsibility, and its magnitude is strongly model-dependent: LLaMA-3.3-70B exhibits nearly three times the cross-language sensitivity of GPT-5.2 and Gemini-2.5-Pro ($V = 0.52$ vs. $V \approx 0.19$). These findings suggest that users in different linguistic communities may receive systematically different attributions of responsibility from the same model, without any awareness that language is shaping the response.

Future work should examine whether these patterns extend to a broader range of scenarios and thematic domains, whether open-ended elicitation yields similar responsibility attribution patterns to the forced-choice format used here, and whether systematic cross-lingual normative auditing can be developed as a practical tool for evaluating deployed AI systems.

Limitations

Several limitations of this study deserve mention.

Translation is inherently imperfect, and despite our efforts to preserve semantic equivalence, subtle differences in phrasing or connotation across languages may have influenced model outputs in ways that are difficult to fully control for (Brislin, 1970; Nida, 1964). This is a longstanding challenge in cross-lingual research and is not unique to our design (Van de Vijver, 2018; Harkness et al., 2003). A

further limitation is that translations were produced using GPT-5.2, which is also one of the models evaluated in this study. This introduces a potential confound: if GPT-5.2's translations reflect its own normative tendencies, this could partially account for the patterns we observe in its outputs.

The forced-choice attribution format, while necessary for cross-model and cross-language comparability, inevitably simplifies the range of normative positions a model might take. In practice, responsibility is often understood as gradational or context-dependent, and our four-category scheme cannot capture that complexity. Future work using open-ended elicitation or finer-grained classification could offer a clearer picture.

The response options were presented in a fixed order (A through D) across all prompts and languages. Prior work has shown that LLMs exhibit systematic position bias in multiple-choice settings, preferring certain option labels or positions independently of content (Pezeshkpour and Hruschka, 2024). To the extent that such bias exists in our setting, it would affect all languages equally and therefore not confound cross-language comparisons. However, it is possible that position bias interacts with language-specific prompt parsing patterns in ways we cannot rule out, and future work should examine whether randomizing option order affects the cross-language variation we observe.

The scenario set, while designed to cover a range of socially evaluative contexts, is not exhaustive. Additionally, three thematic domains are represented by only a single scenario, which limits the extent to which domain-level findings can be generalized. A single scenario may reflect idiosyncratic features of that particular situation rather than properties of the domain more broadly, and replication with multiple scenarios per domain would strengthen the domain-level conclusions.

Finally, all experiments were run at the default temperature setting. While our robustness checks suggest that stochastic variability is modest compared to the cross-language effects we report, we cannot rule out that different decoding configurations may lead to different results.

Ethical Considerations

This study does not involve human subjects or personal data. All prompts consist of fictional, generalized scenarios. Because the study compares model behavior across languages, we clarify that observed differences should not be interpreted as reflecting inherent cultural traits or value systems associated with speakers of those languages.

Acknowledgements

The authors acknowledge the support of Amrita Vishwa Vidyapeetham and the India-AI Mission, Ministry of Electronics and Information Technology, Government of India, for supporting this work.

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.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, and 1 others. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and 1 others. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. Association for Computing Machinery.

Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in neural information processing systems*, 35:3663–3678.

Richard W. Brislin. 1970. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1(3):187–216.

Bram Bulté and Ayla Rigouts Terryn. 2025. Llms and cultural values: the impact of prompt language and explicit cultural framing. *Computational Linguistics*, pages 1–85.

Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, and 1 others. 2024. How large language models can reshape collective intelligence. *Nature human behaviour*, 8(9):1643–1655.

Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, and 1 others. 2026. Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, 2(1):7.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello Piqueras, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pages 53–67.

Incheol Choi, Richard E Nisbett, and Ara Norenzayan. 1999. Causal attribution across cultures: Variation and universality. *Psychological bulletin*, 125(1):47.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Harald Cramér. 1946. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

David M. Eberhard, Gary F. Simons, and Alison J. Robinson. 2026. *Ethnologue: Languages of the World*, 29th edition. SIL Global, Dallas, Texas.

Susan T Tufts Fiske and Shelley E Taylor. 2020. Social cognition: From brains to culture.

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.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

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. *arXiv preprint arXiv:2407.21783*.

Janet A. Harkness, Fons J. R. Van de Vijver, and Peter Ph. Mohler. 2003. *Cross-Cultural Survey Methods*. Wiley, Hoboken, NJ.

Geert Hofstede. 2001. Culture's recent consequences: Using dimension scores in theory and research. *International Journal of cross cultural management*, 1(1):11–17.

Junfeng Jiao, Saleh Afroogh, Abhejay Murali, Kevin Chen, David Atkinson, and Amit Dhurandhar. 2025. Llm ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models. *Scientific Reports*, 15(1):34642.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. *arXiv preprint arXiv:2406.01771*.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Yan Liu and He Wang. 2026. Who on earth is using generative ai? *World Development*, 199:107260.

Michael W Morris and Kaiping Peng. 1994. Culture and cause: American and chinese attributions for social and physical events. *Journal of Personality and Social psychology*, 67(6):949.

Eugene A. Nida. 1964. *Toward a Science of Translating*. E. J. Brill, Leiden.

Soyoung Oh and Vera Demberg. 2025. Robustness of large language models in moral judgements. *Royal Society Open Science*, 12(4).

OpenAI. 2025. Introducing GPT-5.2. `https://openai.com/index/introducing-gpt-5-2/`. Accessed: 2025.

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

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Replicate. 2024. Run AI with an API.

Pratik Sachdeva and Tom van Nuenen. 2025. Normative evaluation of large language models with everyday moral dilemmas. In *Proceedings of the 2025 ACM conference on fairness, accountability, and transparency*, pages 690–709.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809.

Inhwa Song, Sachin R Pendse, Neha Kumar, and Munmun De Choudhury. 2025. The typing cure: Experiences with large language model chatbots for mental health support. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–29.

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.

Harry C Triandis. 1995. Individualism and collectivism. westview press. *Boulder, CO*.

Fons J. R. Van de Vijver. 2018. Cross-cultural research methods. In *International Encyclopedia of Anthropology*. Wiley, Hoboken, NJ.

Omid Veisi, Sasan Bahrami, Roman Englert, and Claudia Müller. 2025. Ai ethics and social norms: Exploring chatgpt's capabilities from what to how. *Proceedings of the ACM on human-computer interaction*, 9(7):1–34.

Karina Vida, Fabian Damken, and Anne Lauscher. 2024. Decoding multilingual moral preferences: Unveiling llm's biases through the moral machine experiment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1490–1501.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7900–7932.

Bernard Weiner. 1985. An attributional theory of achievement motivation and emotion. *Psychological review*, 92(4):548.

A Robustness and Iteration Stability

Because model outputs are generated under stochastic decoding, we assess whether the observed cross-language differences could be attributable to sampling variability rather than systematic distributional shifts. For each model–language–scenario combination, five independent generations were produced. We compute the standard deviation of responsibility category proportions across iterations to estimate iteration-level variability. A bar chart showing scenario-level iteration variability (standard deviation across generations) is shown in Figure 4.

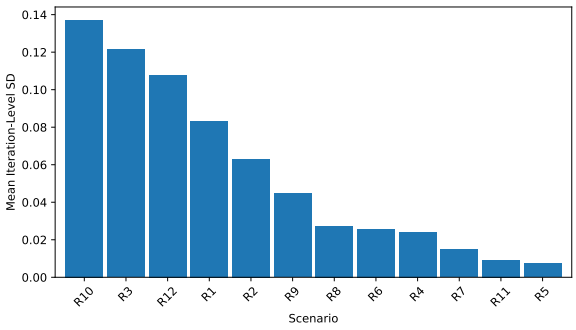

Figure 4: Scenario-level iteration variability measured as the standard deviation of responsibility category proportions across five independent generations for each model–language–scenario combination.

Across scenarios, the mean iteration-level standard deviation is 0.056. This value is substantially smaller than the cross-language differences observed in the overall and domain-level analyses, where Cramér's V ranges from approximately 0.19 to 0.52 and scenario-level dispersion frequently exceeds 0.10. These results indicate that stochastic variability under repeated sampling is modest relative to cross-language variation.

B Supplementary Materials

This appendix provides detailed tables, supplementary analyses, and additional figures referenced in the main text.

Language	**A**	**B**	**C**	**D**
Arabic	120	10	5	45
Bengali	92	11	1	76
Chinese	130	38	1	11
English	134	23	9	14
French	97	11	4	68
Hindi	111	17	2	50
Portuguese	151	20	1	8
Russian	124	23	15	18
Spanish	152	18	7	3
Urdu	131	27	8	14

Table 5: Contingency table of responsibility attribution by language, aggregated across all scenarios and models. Values indicate the frequency of responses assigned to each responsibility category (A–D) for each language.

Domain	Language	A	B	C	D	Domain	Language	A	B	C	D
Authority	Arabic	0.467	0.000	0.000	0.533	Gender	Arabic	0.333	0.000	0.333	0.333
	Bengali	0.667	0.000	0.000	0.333		Bengali	0.200	0.000	0.067	0.733
	Chinese	0.733	0.000	0.067	0.200		Chinese	0.200	0.733	0.000	0.067
	English	0.400	0.000	0.267	0.333		English	0.133	0.467	0.333	0.067
	French	0.333	0.000	0.067	0.600		French	0.333	0.000	0.200	0.467
	Hindi	0.400	0.067	0.067	0.467		Hindi	0.067	0.133	0.067	0.733
	Portuguese	1.000	0.000	0.000	0.000		Portuguese	0.667	0.200	0.067	0.067
	Russian	0.467	0.000	0.333	0.200		Russian	0.067	0.267	0.667	0.000
	Spanish	0.600	0.000	0.333	0.067		Spanish	0.733	0.133	0.133	0.000
	Urdu	0.667	0.000	0.200	0.133		Urdu	0.267	0.067	0.333	0.333
Career	Arabic	0.822	0.000	0.000	0.178	Marriage	Arabic	0.822	0.000	0.000	0.178
	Bengali	0.578	0.000	0.000	0.422		Bengali	0.667	0.000	0.000	0.333
	Chinese	0.889	0.022	0.000	0.089		Chinese	0.933	0.000	0.000	0.067
	English	0.978	0.000	0.000	0.022		English	1.000	0.000	0.000	0.000
	French	0.733	0.000	0.000	0.267		French	0.822	0.000	0.000	0.178
	Hindi	0.822	0.000	0.000	0.178		Hindi	0.889	0.000	0.000	0.111
	Portuguese	0.889	0.000	0.000	0.111		Portuguese	1.000	0.000	0.000	0.000
	Russian	0.889	0.000	0.000	0.111		Russian	0.978	0.000	0.000	0.022
	Spanish	1.000	0.000	0.000	0.000		Spanish	1.000	0.000	0.000	0.000
	Urdu	0.911	0.000	0.000	0.089		Urdu	1.000	0.000	0.000	0.000
Elder Care	Arabic	0.489	0.222	0.000	0.289	Family	Arabic	0.800	0.000	0.000	0.200
	Bengali	0.289	0.244	0.000	0.467		Bengali	0.667	0.000	0.000	0.333
	Chinese	0.422	0.578	0.000	0.000		Chinese	1.000	0.000	0.000	0.000
	English	0.489	0.356	0.000	0.156		English	1.000	0.000	0.000	0.000
	French	0.156	0.244	0.000	0.600		French	0.667	0.000	0.000	0.333
	Hindi	0.378	0.311	0.000	0.311		Hindi	0.667	0.000	0.000	0.333
	Portuguese	0.578	0.378	0.000	0.044		Portuguese	1.000	0.000	0.000	0.000
	Russian	0.378	0.422	0.000	0.200		Russian	1.000	0.000	0.000	0.000
	Spanish	0.600	0.356	0.000	0.044		Spanish	1.000	0.000	0.000	0.000
	Urdu	0.356	0.578	0.000	0.067		Urdu	1.000	0.000	0.000	0.000

Table 6: Domain-level responsibility attribution distributions across languages. Values represent the proportion of responses assigned to each responsibility category (A–D).

Language	A	B	C	D
Arabic	41	8	5	6
Bengali	40	10	0	10
Chinese	44	11	0	5
English	44	14	0	2
French	40	5	0	15
Hindi	46	7	1	6
Portuguese	45	8	0	7
Russian	38	6	5	11
Spanish	50	8	0	2
Urdu	34	10	3	13

(a) GPT-5.2

Language	A	B	C	D
Arabic	21	0	0	39
Bengali	0	0	0	60
Chinese	38	16	0	6
English	44	0	4	12
French	11	0	0	49
Hindi	19	0	0	41
Portuguese	60	0	0	0
Russian	40	5	10	5
Spanish	59	0	0	1
Urdu	46	11	3	0

(b) LLaMA-3.3-70B

Language	A	B	C	D
Arabic	58	2	0	0
Bengali	52	1	1	6
Chinese	48	11	1	0
English	46	9	5	0
French	46	6	4	4
Hindi	46	10	1	3
Portuguese	46	12	1	1
Russian	46	12	0	2
Spanish	43	10	7	0
Urdu	51	6	2	1

(c) Gemini-2.5-Pro

Table 7: Language $\times$ responsibility contingency tables for each model, aggregated across all scenarios. Values represent frequency of responses assigned to each responsibility category (A–D).

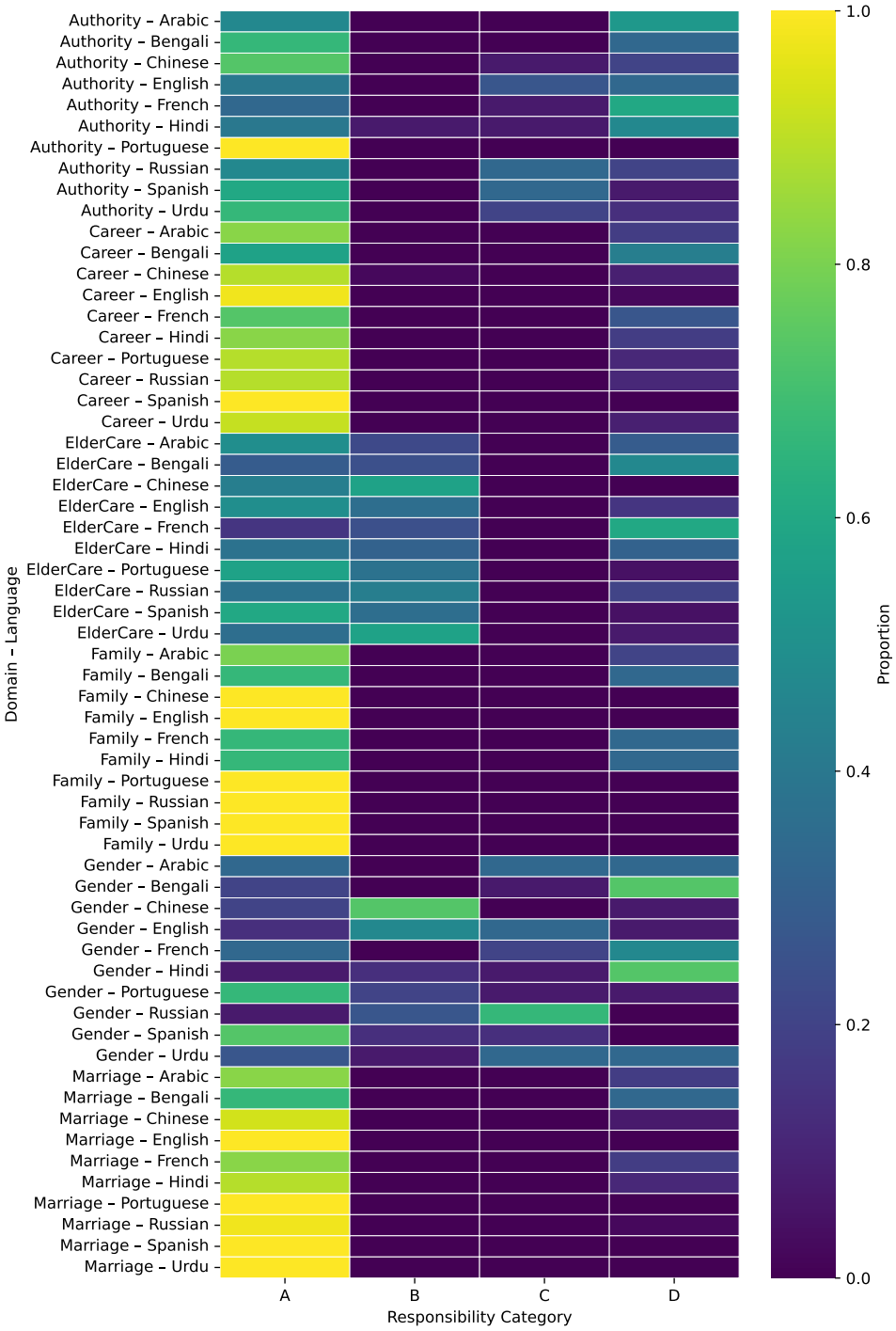

Figure 5: Heatmap of responsibility attribution distributions across languages and domains. Each cell represents the proportion of responses assigned to a given responsibility category (A–D) for a specific domain–language combination, highlighting cross-linguistic variation in attribution patterns.