Native-Speaker Personas Are Not Enough: Language Pathways Shape LLM Interpersonal Advice

Anonymous ACL submission

Abstract

LLMs are increasingly used for interpersonal advice and as tools for studying social behavior across languages and cultures. A common shortcut for producing culturally situated responses is to ask a model to answer as a native speaker. We test whether this native-speaker persona reproduces the outputs obtained when models instead generate advice in the target language and translate the response back to English. Using 600 interpersonal advice questions across 13 languages and eight LLMs, we compare native-language generation followed by translation (NL) with native-speaker persona prompting (NP), measuring linguistic style, behavioral scaffolding in generated advice, and forced-choice action recommendations. We find that NP and NL are not interchangeable. Linguistically, NP often increases lexical social cues, including affiliation and positive tone, while reducing practical qualities such as concreteness and social attunement; it also provides less actionable scaffolding in open-ended advice. In forced-choice scenarios, NP changes which action the model selects, favoring confrontation over redirection relative to NL, with effects varying across languages, topics, and models. Our results show that cross-lingual prompting strategy is a methodological choice with substantive consequences: persona prompts can make advice sound socially adapted while changing both the guidance users receive and the actions models recommend.

1 Introduction

Large language models (LLMs) are widely used for interpersonal guidance, from workplace communication and boundary-setting to relationship conflict and family disagreements (Chatterji et al., 2025). In these settings, models do more than produce language: they help users decide how to behave toward others. This is especially consequential in cross-lingual and culturally mismatched contexts,

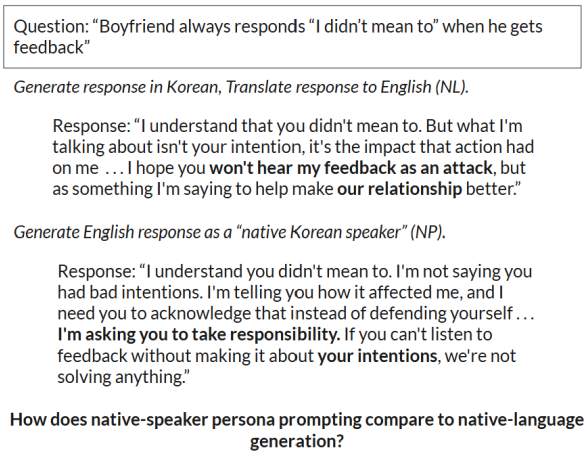

Figure 1: Example of how GPT-5.5 varies based on prompting strategy.

where users may know *what* they want to say but not *how* to say it. Because languages and cultures differ in norms of directness, formality, emotion, deference, and values, interpersonal advice tightly links linguistic form with social meaning.

Yet multilingual LLM evaluation has focused largely on translation, question answering, reasoning, instruction following, and other standard NLP benchmarks (Hu et al., 2020; Ahuja et al., 2023; Lai et al., 2023; Singh et al., 2024; Üstün et al., 2024). These tasks do not fully capture interpersonal advice, where there may be no single correct answer and differences in directness, politeness, hierarchy, and emotional expression can change the force of a response (Goffman, 1967; Brown and Levinson, 1987; Ting-Toomey and Kurogi, 1998). This limitation is especially important because LLMs are increasingly used to model human behavior through demographic backstories, persona variables, prompt language, and cultural frames (Argyle et al., 2023; Hu and Collier, 2024; Tao et al., 2024), even though recent work shows that persona prompts can influence outputs with-

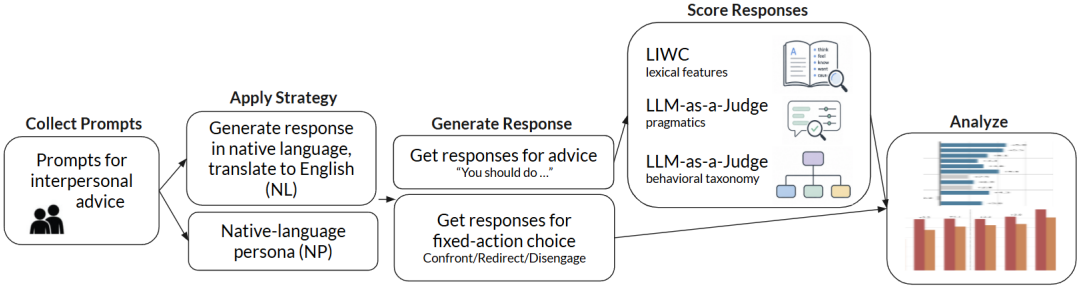

Figure 2: Framework for evaluating how language pathways affect LLM-generated responses to interpersonal advice prompts.

out reliably capturing the full complexity of human subjectivity or cultural variation (Hu and Collier, 2024; Giorgi et al., 2024). In multilingual interpersonal advice, this raises a methodological question: is asking a model to respond as a "native speaker" equivalent to asking it to process or generate advice in the target language?

This study investigates how language pathways shape LLM-generated interpersonal advice (Figure 1). Starting from English interpersonal advice questions, we translate prompts into thirteen languages and compare responses from five open-weight and three proprietary API models. We compare native-language generation followed by translation into English (NL) with native-speaker persona prompting in English (NP), testing whether this common persona-based shortcut reproduces advice generated through the target language itself (Figure 2).

Our analysis examines both how advice is framed and what the model recommends. We measure linguistic characteristics, including tone, power, formality, and concreteness; behavioral guidance using the Behavior Change Technique Taxonomy (Michie et al., 2013); and recommended action through a forced-choice classification of confrontation, disengagement, or redirection.

This paper makes three contributions. First, we show that language pathway is not a neutral implementation detail: the same interpersonal dilemma yields different linguistic and behavioral outputs depending on whether the model generates in the target language or is prompted with a native-speaker persona in English. Second, we show that NP and NL diverge along both linguistic and behavioral dimensions: NP often adds lexical social cues, such as prosocial language and positive tone, while reducing concreteness, social attunement, emotional expressiveness, and structured behavioral guidance relative to NL. Third, we show that these differences vary by language, topic, and model, suggesting structured elicitation effects rather than random variation. We interpret these patterns as properties of model elicitation, not as evidence about real language communities.

2 Related Work

2.1 Multilingual LLM Evaluation

Multilingual NLP evaluation has traditionally focused on whether models transfer across languages on standard tasks such as classification, question answering, retrieval, translation, and structured prediction. Benchmarks such as XTREME evaluate multilingual representations across many languages and task types (Hu et al., 2020). More recent LLM-oriented benchmarks such as MEGA and ChatGPT Beyond English evaluate generative models across multilingual settings (Ahuja et al., 2023; Lai et al., 2023). Multilingual instruction-tuning efforts such as Aya address the English-centered nature of LLM development by constructing multilingual instruction datasets and models (Singh et al., 2024; Üstün et al., 2024). While existing literature shows that LLM behavior varies substantially across languages, most evaluations emphasize task correctness, benchmark performance, or instruction-following quality. Interpersonal advice differs because there may be no single correct answer; the relevant question concerns interpersonal stance, tone, and action recommendation.

2.2 LLMs for Interpersonal Advice and Social Judgment

A growing body of work examines LLMs in social, interpersonal, and emotional-support settings.

Studies of LLM-generated advice show that advice quality is shaped not only by the recommendation itself, but also by phrasing, tone, and perceived usefulness (Wester et al., 2024). Work on social-situational judgment finds that LLMs can perform strongly on scenarios involving interpersonal conflict and social reasoning (Mittelstädt et al., 2024). Related work on emotional support studies whether LLMs can produce empathic or culturally sensitive responses, while showing that simple role-play prompts may be insufficient for cultural sensitivity (Liu et al., 2026).

2.3 Personas, Cultural Prompting, and Social Bias

LLMs are increasingly used as proxies for human respondents, annotators, or social actors by prompting models with demographic, cultural, or persona descriptions and analyzing outputs as evidence about attitudes, values, or decision-making. Recent work shows that persona prompts influence model outputs but capture only part of human variation (Hu and Collier, 2024); explicit and implicit persona factors likewise affect model behavior while failing to represent the full complexity of human experience, belief, and context (Giorgi et al., 2024). Cultural prompting raises similar concerns: model behavior depends on how culture is represented and evaluated (AlKhamissi et al., 2024), and persona choices can shift cultural norm judgments or amplify stereotypes (Kamruzzaman et al., 2026). Bias benchmarks and social-bias studies further show that models encode stereotypical associations across demographic descriptors, including multilingual ones, and can reproduce gendered stereotypes or vary judgments and persuasive style with speaker or relationship attributes (Nadeem et al., 2021; Smith et al., 2022; Costa-jussà et al., 2023; Kotek et al., 2023; Levy et al., 2024; Pauli et al., 2026). Our work is complementary: rather than treating a native-speaker persona as a transparent proxy for cross-lingual behavior, we test whether native-speaker persona prompting reproduces native-language linguistic and behavioral patterns.

3 Methodology

3.1 Data Collection

We collect interpersonal advice questions from the Interpersonal Skills Stack Exchange, a public forum focused on everyday interpersonal problems (Stack Exchange, 2026). We select the top 150 questions from four tags: work-environment, friends, relationships, and family, yielding 600 English questions. Sample prompts are shown in Appendix A.1. These questions are advice-seeking, socially situated, and behaviorally open-ended, making them suitable for evaluating differences in tone, directness, politeness, social orientation, and recommended action.

3.2 Languages

We evaluate 13 languages spanning multiple scripts, language families, and regions: Japanese, Korean, Chinese, Arabic, Persian, Turkish, English, German, Dutch, Swedish, Spanish, Italian, and Portuguese. English uses the original questions, while all other languages use the translation pipeline below to construct semantically controlled non-English versions. We use broad groupings such as East Asian, Middle Eastern, Germanic, and Romance as descriptive aids for exploratory analysis, not as ground-truth cultural categories.

3.3 Question Translation

We translate the English questions into the 12 non-English languages using facebook/nllb-200-distilled-1.3B, a multilingual sequence-to-sequence model from No Language Left Behind (Costa-jussà et al., 2022). For each question-language pair, we generate 32 candidate translations with temperature 0.3 and back-translate each candidate into English four times with temperature 0.7. We then select the candidate with the highest mean and lowest variance in cosine similarity to the original question, computed using all-MiniLM-L6-v2 (Reimers and Gurevych, 2019; Wang et al., 2020). Selected translations show high semantic preservation overall, with mean similarity of 0.928 across languages and language-level means ranging from 0.876 for Japanese to 0.976 for Spanish; additional diagnostics and examples appear in Appendix A.2.

3.4 Response Generation

We generate responses from five open- and three proprietary models: google/gemma-4-E4B-it, google/gemma-4-26B-A4B-it, google/gemma-4-31B-it, Qwen/Qwen3.6-27B, Qwen/Qwen3.6-35B-A3B, claude-opus-4.6, gpt-4o, and gpt-5.5 (Google, 2026; Qwen Team, 2026a,b; Anthropic, 2026; OpenAI, 2024, 2026).

For each non-English language, model, and strategy, we generate responses to all 600 questions; with four prompting strategies, this yields 2,400 records per model per non-English language. For English, models answer the original English questions.

For each non-English language, we compare native language generation (NL), where the model receives and answers the question in the target language before the answer is translated into English for analysis, with native-speaker persona prompting (NP), where the model receives the original English question and answers in English as a native speaker of the target language. Exact prompts are provided in Appendix A.3.

3.5 Annotation and Measurement

We analyze generated responses using four measurement layers: LIWC-22 lexical features, LLM-based pragmatic annotations, LLM-based behavior-change technique classifications, and forced-choice behavioral coding.

First, we compute 12 LIWC lexical features on the English-analyzable final responses. LIWC provides dictionary-based measures of psychologically and socially meaningful word use (Tausczik and Pennebaker, 2010; Pennebaker et al., 2015; Boyd et al., 2022). We use features spanning social, affective, cognitive, and motivational language: affiliation, achievement, power, cognitive processes, certitude, tentativeness, emotion, positive tone, negative tone, social references, prosocial language, and conflict. These measures provide a lexical analysis of the responses.

Second, we use gpt-4o to annotate holistic properties of the advice along five dimensions that are difficult to capture with lexical counts alone: direct versus indirect, formal versus informal, emotionally expressive versus restrained, socially attuned versus task-focused, and concrete versus abstract. These dimensions are motivated by work on politeness, facework, stance, and register, which shows that language encodes interpersonal meaning beyond propositional content (Goffman, 1967; Brown and Levinson, 1987; Biber, 1995; Du Bois, 2007). We use LLM-based annotation because these properties are context-sensitive; prior work shows that structured LLM evaluators can support scalable assessment of open-ended text, while requiring caution about evaluator bias (Liu et al., 2023; Zheng et al., 2023).

Third, we code responses using higher-level clusters from the Behavior Change Technique Taxonomy (Michie et al., 2013) with an LLM-as-a-Judge. Although the taxonomy was developed for behavior-change interventions, several categories capture general forms of action guidance that also appear in interpersonal advice: identifying a problem, planning a concrete response, anticipating consequences, reframing the situation, seeking support, and monitoring outcomes. We use BCT categories not to make clinical claims, but to measure behavioral scaffolding in advice. This complements linguistic analysis: two responses may sound equally polite or prosocial while differing in whether they help the user decide what to do, how to do it, and what consequences to consider.

Finally, we run a separate forced-choice action task motivated by frameworks that distinguish direct engagement, avoidance, compromise, and integrative responses (Rahim, 1983; Gross and Guerrero, 2000). For each question-language-strategy condition, the model receives the same interpersonal dilemma and chooses one of three broad action families: *confrontation*, *disengagement*, or *redirection*. Confrontation directly addresses the issue with the other person or an authority figure; disengagement pauses, reduces, or avoids interaction; and redirection captures non-avoidant but less directly confrontational alternatives such as reframing, changing communication channels, seeking mediation, or pursuing compromise. This task is separate from generated-advice coding: the model makes a direct action choice rather than labeling its earlier response.

3.6 Statistical Analysis

Our main analysis compares NL and NP across output measures, using English as a reference condition for scaling and visualization. For continuous LIWC, LLM-annotated, and BCT features, we report standardized treatment-minus-English differences for the same source prompt and model. NP–NL tests use matched prompt–model observations: within each source prompt and model, we average feature scores across target languages for each strategy, compute the NP-minus-NL difference, and test whether the mean differs from zero using a one-sample paired-difference t-test (Student, 1908). For exploratory language-group analyses, we compute the same NP–NL contrasts within matched observations, average them within broad language groups, and test whether the persona–generation gap differs across groups.

For forced-choice action outcomes, each response is labeled as confrontation, redirection, or disengagement. We report strategy-level action distributions as percentages and test whether prompting strategy changes the overall distribution using Pearson chi-square tests over the three action categories (Pearson, 1900). Across analyses, stars indicate significant NP–NL differences after Benjamini–Hochberg false discovery rate correction within each comparison family (Benjamini and Hochberg, 1995) ($q < 0.05$). Continuous effects are shown in standardized treatment-minus-English units; action effects are shown in percentage points.

4 Linguistic Analysis

This section analyzes how the cross-lingual prompting strategy affects the linguistic features in the resulting advice using LIWC lexical features and LLM-based annotations. The motivating example in Figure 1 illustrates the difference between NP and NL linguistically. The Korean-generated response translated into English is more relationally attuned, softening with phrases like "I hope you won't hear my feedback as an attack." By contrast, the native-speaker persona response uses a firmer accountability frame, warning the boyfriend that "if you can't listen to feedback without making it about your intentions, we're not solving anything." This is reflected in the annotations: NP is less formal (2 vs. 3) and less socially attuned (3 vs. 5).

4.1 Overall Linguistic Patterns

Figure 3 shows that NL and NP produce significantly different linguistic profiles. Overall, NP shifts advice toward an assertive-supportive style: it uses more affiliation, social references, prosocial language, and positive tone, but is less formal, less emotionally expressive, less socially attuned, and less concrete than NL. By comparison, NL produces advice with more contextual and relational elaboration, including greater attention to emotional nuance, relationship dynamics, and concrete interpersonal detail. This overall pattern is broadly consistent across LLMs, with model-level results reported in Appendix D.4.

This contrast suggests that native-speaker persona prompting does not simply reproduce native-language generation in English. Instead, NP appears to add recognizable markers of social

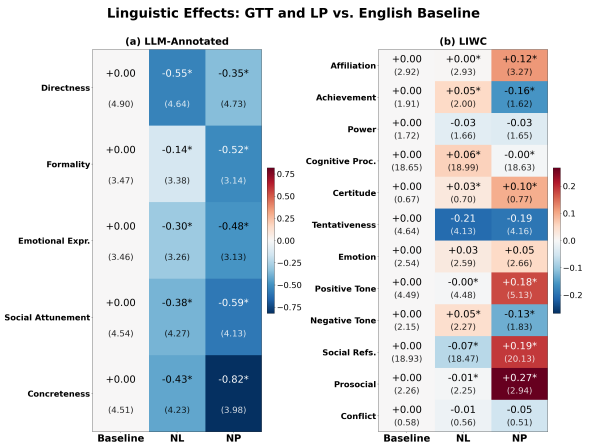

Figure 3: Mean linguistic differences (left: LLM-annotated, right: LIWC) between NL and NP, normalized to the English baseline. Stars indicate significant NP–NL differences. Compared to NL, NP increases surface social cues such as affiliation and reduces action-oriented features such as achievement and concreteness.

warmth or cultural adaptation while weakening features that make advice pragmatically grounded: emotional expressiveness, social attunement, concreteness, and action orientation. In this sense, NP can make advice sound more socially positive without preserving the fuller interpersonal structure induced by generating in the target language.

4.2 Strategy Differences Across Languages

The size and direction of NP–NL differences vary systematically by language (Figure 4). Across languages, NP generally increases directness and socially positive cues, while NL is higher on formality, social attunement, emotional expressiveness, and concreteness. However, the magnitude of these gaps is not uniform: NL's advantage in social attunement and concreteness is smaller for Japanese, Korean, and Chinese than for several Romance and Germanic languages. Exploratory language-group analyses confirm significant differences in effect sizes across broad language groups; full results appear in Appendix D.5.

These language-level differences reflect the distinct pathways induced by the two prompts. NP keeps the exchange in English and adds a language-identity label, while NL requires the model to formulate advice in the target language before translation. The divergences therefore suggest that persona prompts can capture some language-associated surface patterns, but do not reproduce the full structure elicited by generating in the lan-

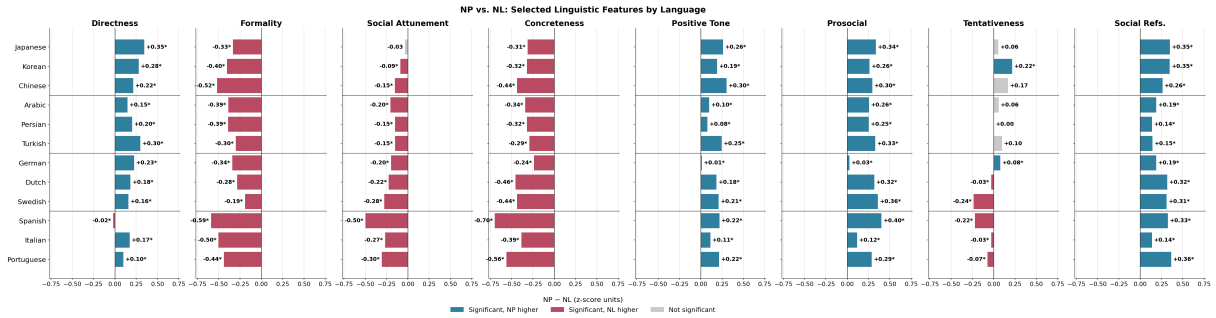

Figure 4: NP–NL differences in selected linguistic features by language. The selected features show a subset of how the advice positions the user interpersonally. Differences vary across target languages, showing that persona prompting does not affect all linguistic contexts uniformly.

guage itself.

4.3 Discussion

Together, the linguistic results show that NP and NL are distinct elicitation methods for multilingual interpersonal advice. NP tends to produce advice that is warmer, more positive, and more explicitly socially marked, while NL more often preserves the relational and pragmatic scaffolding of advice: how directly the user should speak, how much emotional context is included, how carefully the relationship is managed, and how concretely next steps are specified. The difference is therefore not only stylistic. Language pathway changes how the advice positions the user in the interpersonal situation.

5 Behavioral Analysis

This section examines whether language pathway changes not only how advice sounds, but what it helps the user do. We analyze generated responses with the Behavior Change Technique Taxonomy (BCT), which captures behavior-guiding components such as problem solving, planning, reframing, and consequence reasoning. We then use a forced-choice action task to test whether the model ultimately recommends confrontation, disengagement, or redirection. The motivating example in Figure 1 illustrates the distinction: NL provides more behavioral scaffolding than NP, scoring higher on cognitive reframing (4 vs. 2), social support (4 vs. 2), and problem solving (5 vs. 3). The text reflects this difference: NL frames feedback as "I'm saying to help make our relationship better," while NP says the boyfriend needs to "acknowledge that instead of defending." The forced-choice label also changes, with Korean generation selecting "Redirection" and the Korean-speaker persona

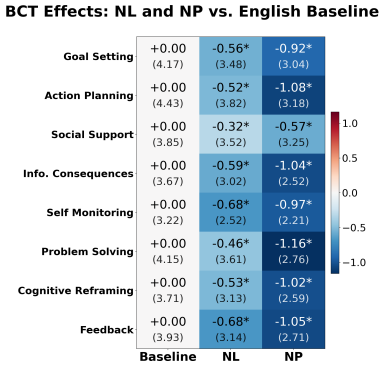

Figure 5: Mean BCT differences between NL and NP, normalized to the English baseline. Overall, NP significantly differs from NL, with personas providing less behavioral scaffolding in their advice.

selecting "Confrontation."

5.1 Behavioral Guidance in Generated Advice

The BCT analysis shows that NP provides less behavioral scaffolding than NL across major action-guidance categories (Figure 5). The largest reductions appear in problem solving, action planning, consequence reasoning, feedback, and cognitive reframing. These categories capture the difference between advice that endorses a general stance and advice that helps a user carry it out: identifying the source of conflict, deciding when and how to speak, anticipating how the other person might respond, and reframing the situation in a way that makes action possible. Because these reductions appear consistently across languages, they suggest a general strategy-level effect rather than a language-specific anomaly.

This result strengthens the linguistic findings. NP often makes advice sound warmer or more socially positive, but the BCT results show that this

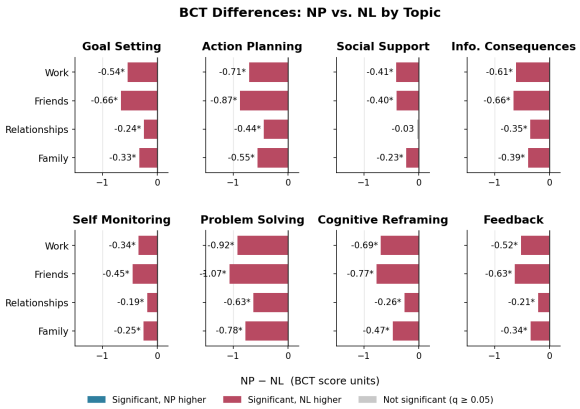

Figure 6: NP–NL differences in BCT scores by topic. The largest differences between NP and NL appear in interpersonal questions about work and friends.

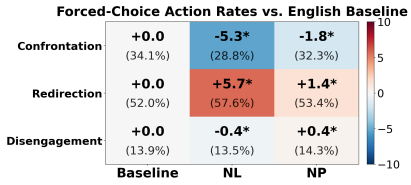

Figure 7: Forced-choice action rates for NP and NL normalized by the English baseline. Cells report action rates and percentage-point differences from English. NP and NL have significantly different distributions, with personas choosing confrontation significantly more than native language generation.

tradeoff can come at the cost of usable guidance. This is consequential, as a response that says to be honest, respectful, or understanding may sound appropriate, but it is less helpful if it does not help the user decide what to say, how to sequence the conversation, or how to manage the relationship after the exchange.

Topic-level results show that this loss of usable guidance is most pronounced for friendship and workplace questions (Figure 6), especially in problem solving, action planning, and cognitive reframing. These are domains where the practical and relational demands of advice are tightly linked: users often need to raise a concern without escalating conflict, coordinate expectations with someone they must continue interacting with, or preserve trust while setting a boundary. One interpretation is that persona prompting encourages a recognizable social stance, while native-language generation retains more of the step-by-step reasoning needed to make advice actionable in ongoing relationships.

5.2 Forced-Choice Action Preferences

Next, the forced-choice task asks whether these differences in scaffolding translate into different behavioral recommendations. Figure 7 shows that NL and NP produce significantly different action distributions: NL shifts away from confrontation and toward redirection, while NP remains closer to the English baseline and selects confrontation more often than NL.

Because the final action prompt is identical across strategies, this difference cannot be reduced to wording alone. It suggests that language pathway changes the model's judgment of what the situation calls for. In practice, users may receive different behavioral guidance for the same dilemma depending on whether the system uses native-language generation or native-speaker persona prompting.

Language-level results add a more nuanced picture (Figure 8). Across most languages, NP selects confrontation more often than NL, while NL selects redirection more often. At the same time, NP is not simply random: languages with higher confrontation rates under NL often also have higher confrontation rates under NP. Persona prompting therefore preserves some relative language-level structure, but shifts the absolute recommendation toward confrontation.

This pattern is important because it suggests that persona prompting is not merely noisy or uninformative. Instead, it appears systematically miscalibrated: it can capture coarse language-associated differences while still changing the practical advice a user receives.

5.3 Discussion

Taken together, the behavioral results show that persona prompting changes both the structure of generated advice and the final judgment about what action is appropriate. NP can make advice appear more socially polished while reducing the scaffolding that helps users act, and it can shift recommendations for the same dilemma. The fact that NP sometimes preserves coarse language-level patterns while still shifting choices toward confrontation suggests that persona prompting is not useless, but miscalibrated: it may reflect relative language-associated variation without reproducing the concrete guidance produced by native-language generation.

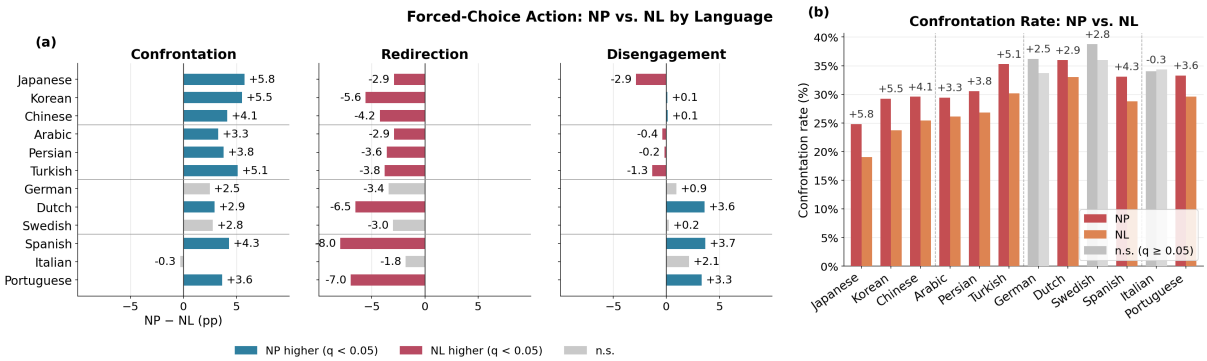

Figure 8: NP–NL forced-choice differences by language. Panel (a) shows NP minus NL action-rate differences; Panel (b) compares confrontation rates under NP and NL.

6 Conclusion

This paper asks whether native-language generation followed by translation (NL) and native-speaker persona prompting (NP) produce equivalent multilingual interpersonal advice. Across 600 advice questions, 13 language conditions, and eight language models, we find that they do not: NL and NP yield systematically different linguistic styles, levels of behavioral scaffolding, and action recommendations. Persona prompting can selectively amplify visible markers of social adaptation, such as prosocial wording, affiliation, social references, and positive tone, while changing the structure of the advice itself. Relative to NL, NP is often less concrete, less socially attuned, less emotionally expressive, and less formal; it also provides less actionable scaffolding and more often shifts forced-choice recommendations from redirection toward confrontation. We interpret these differences as a prompting-method effect: persona prompts may rely on compressed model associations about a language identity, while native-language generation requires the model to produce advice through the target language itself. Multilingual advice systems should therefore evaluate prompting strategy explicitly and validate both linguistic framing and behavioral recommendations across cultural settings.

Limitations

This study has several limitations. First, although we exclude questions with explicit country tags, the source data comes from an English-language advice forum. The interpersonal dilemmas may therefore reflect Western, English-speaking assumptions. Our results should be interpreted as differences in how models respond to translated versions of this corpus, not as claims about how people from different language communities would answer these questions.

Second, our language and persona conditions are broad. A prompt such as "You are a native Korean speaker" collapses across region, class, age, and gender. This is useful for testing a common strategy, but not for representing cultural identity. Future work should examine more specific regional, dialectal, and situational contexts.

Third, our analysis relies on automated annotation pipelines. LIWC provides reproducible lexical features, while LLM-based annotations and BCT labels allow scalable evaluation, but both may miss nuance or reflect annotator-model biases. Future work should evaluate LLM-generated responses with human raters from the relevant linguistic and cultural backgrounds.

Although we include both open- and proprietary models, differences in training data, alignment, and English-centered development may limit generalizability. Our consistency analyses suggest that many patterns generalize across models, but not necessarily to all multilingual systems. Expanding to systems developed primarily in non-English contexts would help clarify which effects are general properties of cross-lingual prompting and which are model-specific.

AI Assistance Disclosure

The authors used ChatGPT for writing assistance in revising, editing, and LaTeX formatting help.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal,

Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

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*. Association for Computational Linguistics.

Anthropic. 2026. Claude opus 4.6 system card. Anthropic System Card. Accessed: 2026-05-26.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. *The Development and Psychometric Properties of LIWC-22*. University of Texas at Austin.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research.

Marta R. Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1–28. Association for Computational Linguistics.

John W. Du Bois. 2007. The stance triangle. In Robert Englebretson, editor, *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, pages 139–182. John Benjamins.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188. Association for Computational Linguistics.

Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. Anchor Books.

Google. 2026. Gemma 4 model overview. Google AI for Developers Documentation. Accessed: 2026-05-26.

Michael A. Gross and Laura K. Guerrero. 2000. Managing conflict appropriately and effectively: An application of the competence model to rahim's organizational conflict styles. *International Journal of Conflict Management*, 11(3):200–226.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421. PMLR.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*,

pages 10289–10307. Association for Computational Linguistics.

Mahammed Kamruzzaman, Hieu Minh Nguyen, Nazmul Hassan, and Gene Louis Kim. 2026. 'a woman is more culturally knowledgeable than a man?': The effect of personas on cultural norm interpretation in LLMs. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 220–237. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM Collective Intelligence Conference*. Association for Computing Machinery.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Sharon Levy, William D. Adler, Tahilin Sanchez Karver, Mark Dredze, and Michelle R. Kaufman. 2024. Gender bias in decision-making with large language models: A study of relationship conflicts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

Chen Cecilia Liu, Hiba Arnaout, Nils Kovačić, Dana Atzil-Slonim, and Iryna Gurevych. 2026. Tailored emotional LLM-supporter: Enhancing cultural sensitivity. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pages 535–574. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P. Eccles, James Cane, and Caroline E. Wood. 2013. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine*, 46(1):81–95.

Justin Maximilian Mittelstädt, Julia Maria Maier, Panja Goerke, Frank Zinn, and Michael E. Hermes. 2024. Large language models can outperform humans in social situational judgments. *Scientific Reports*, 14:27449.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5356–5371. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2026. Gpt-5.5 system card. OpenAI System Card.

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.

Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin.

Qwen Team. 2026a. Qwen3.6-27B: Flagship-level coding in a 27b dense model. Qwen Blog. Accessed: 2026-05-26.

Qwen Team. 2026b. Qwen3.6-35B-A3B: Agentic coding power, now open to all. Qwen Blog. Accessed: 2026-05-26.

M. Afzal Rahim. 1983. A measure of styles of handling interpersonal conflict. *Academy of Management Journal*, 26(2):368–376.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

Shivalika Singh, Francisco Vargus, Daniel Dsouza, Bjarte Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Darius Mataciunas, Laura O'Mahony, Mike Zhang, Ishan Hettiarachchi, Joseph Wilson, Marina Machado, Luis Moura, Marzieh Fadaei, Mathias Müller, Peter Zhang, Philipp Schmid, and 4 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 918–930. Association for Computing Machinery.

Stack Exchange. 2026. Interpersonal Skills Stack Exchange. https://interpersonal.stackexchange.com/. Accessed: 2026-05-19.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25.

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.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Stella Ting-Toomey and Atsuko Kurogi. 1998. Facework competence in intercultural conflict: An updated face-negotiation theory. *International Journal of Intercultural Relations*, 22(2):187–225.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'Souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.

Joel Wester, Sander de Jong, Henning Pohl, and Niels van Berkel. 2024. Exploring people's perceptions of LLM-generated advice. *Computers in Human Behavior: Artificial Humans*, 2(2):100072.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*.

Question category	# prompts	Example prompt
Work environment	150	*How do I tell my teammate to stop replying to emails not addressed to him?*
Friends	150	*How to talk to someone about being chronically late—not to fix it, just to get more accurate ETAs?*
Relationships	150	*How to talk to my wife about unrealistic expectations?*
Family	150	*How to ask my father-in-law to stop entering our house unannounced?*
Total	600	

Table 1: Prompt categories used in the dataset.

#	Korean candidate translation	Mean sim.	BT var.
1	**부모 에게 어떻게 내 사생활 을 존중 해 달라고 요청 할 수 있는가?**	0.978	0.000
2	부모님한테 어떻게 제 사생활을 존중하라고 부탁하면 돼요?	0.974	0.000
3	부모 에게 어떻게 사생활 을 존중 해 달라고 요청 할 수 있습니까?	0.961	0.000
6	부모 에게 어떻게 내 사생활 을 존중 해 달라고 요구 할 수 있습니까?	0.763	0.139
7	부모 에게 내 사생활 을 존중 하라고 어떻게 하면 될 것 인가?	0.750	0.081

Table 2: Back-translation scores for selected Korean translation candidates. Mean sim. is the average cosine similarity between four back-translations and the original English question. BT var. is the variance of those similarity scores. The bolded row is the selected candidate.

A Framework Setup

This appendix provides implementation details for the full cross-lingual prompting framework.

A.1 Advice Question Samples

We begin from 600 English-language interpersonal advice questions drawn from Interpersonal Skills Stack Exchange. Questions were drawn from four topic tags: work-environment, friends, relationships, and family. These tags cover common interpersonal domains: workplace communication, friendship conflict, romantic or social relationships, and family dynamics. We exclude questions with country-specific tags to focus on broadly comparable dilemmas rather than cases tied to a named national context. Some example prompts are shown in Table A.1.

A.2 Translation Pipeline

For each of the 12 non-English target languages, we translate the original English questions using facebook/nllb-200-distilled-1.3B. We generate 32 candidate translations with temperature 0.3. We then back-translate each candidate into English four times with temperature 0.7. This produces multiple English reconstructions of each candidate translation, allowing us to estimate both mean semantic preservation and instability across back-translations.

We score each back-translation against the original English question using cosine similarity from all-MiniLM-L6-v2. For each question-language pair, we select the translation with high mean similarity and low variance across back-translations by retaining the top-5 candidates by mean similarity and choosing the one with the smallest back-translation variance among those five. We note the translation-quality diagnostics by language in Table 3.

Table 2 shows the full candidate set for this question, illustrating how the pipeline operates. The two excluded candidates (rows 6–7) have markedly lower mean similarity and substantially higher variance, indicating that their back-translations diverge considerably from the original. Among the top-5 candidates, row 1 is selected because its back-translations are perfectly stable (variance $= 0$), even though its mean similarity differs only marginally from rows 2–3.

A.3 Response-Generation Prompts

This subsection reports the response-generation templates. The main paper focuses on the comparison between native-language generation followed by translation (NL) and native-speaker persona prompting (NP). We also include two auxiliary strategies, translate-then-generate (TTG) and one-stage English output (OSP), to help localize where prompting effects enter the pipeline.

TTG tests whether differences are explained primarily by translation of the input question. In TTG, the target-language question is translated back into English before the model generates advice, so the final response is produced from an English prompt. OSP tests whether target-language input alone affects the response when the model is still instructed to answer in English. NL tests the effect of generating the advice through the target language it-

Language	Mean similarity	SD similarity
Japanese	0.876	0.091
Korean	0.879	0.121
Chinese	0.912	0.089
Arabic	0.917	0.092
Persian	0.877	0.112
Turkish	0.908	0.088
German	0.940	0.082
Dutch	0.959	0.070
Swedish	0.957	0.079
Spanish	0.976	0.063
Italian	0.969	0.081
Portuguese	0.975	0.067
Overall	0.928	0.094

Table 3: Translation quality diagnostics by language. Similarity is cross-lingual cosine similarity between the original English question and selected target-language prompt.

self, followed by translation into English for analysis. NP tests whether an English-language native-speaker persona can approximate the advice produced through native-language generation.

All prompts were submitted as a single user turn without a system message. For two-step strategies, the output of Step 1 was inserted into Step 2 at inference time. Here, $\langle Q_{\text{EN}} \rangle$ denotes the original English question, $\langle Q_{\text{NL}} \rangle$ the selected target-language translation, $\langle Q_{\text{BT}} \rangle$ the English back-translation of the target-language question, and $\langle R_{\text{NL}} \rangle$ the target-language response from Step 1.

English baseline. The model receives the original English question and answers in English.

> Answer the following advice question in English. Only provide the direct answer and do not explain.
> $\langle Q_{\text{EN}} \rangle$

Translate-then-generate (TTG). The target-language question is first translated back into English. The model then generates advice from this English back-translation. This condition tests whether translation of the input question alone explains downstream effects.

> Answer the following advice question in English. Only provide the direct answer and do not explain.
> $\langle Q_{\text{BT}} \rangle$

Native-language generation followed by translation (NL). Step 1 generates advice in the target language from the target-language question.

> ⟨Instruction in target language: answer the following advice question in the target language. Only provide the direct answer and do not explain.⟩
> $\langle Q_{\text{NL}} \rangle$

Step 2 translates the target-language response into English for analysis.

> Translate the following response into English. Only provide the translation and do not explain.
> $\langle R_{\text{NL}} \rangle$

One-stage English output (OSP). The model receives the target-language question but is instructed to answer directly in English. This condition tests whether target-language input affects the response even when output language is fixed to English.

> ⟨Instruction in target language: answer the following advice question in English. Only provide the direct answer and do not explain.⟩
> $\langle Q_{\text{NL}} \rangle$

Native-speaker persona prompting (NP). The model receives the original English question with a native-speaker persona instruction and answers in English.

> You are a native ⟨Language⟩ speaker. Answer the following advice question in English. Only provide the direct answer and do not explain.
> $\langle Q_{\text{EN}} \rangle$

B LLM-as-Judge and Forced-Choice Setup

This appendix section reports the prompts used for automated evaluation. We include the linguistic annotation prompt, the BCT annotation prompt, the forced-choice action prompt, and the LLM-as-judge prompt used to validate malformed or ambiguous forced-choice outputs. We also report refusal and failure rates for the forced-choice task. The annotator model used was gpt-4o at temperature 0, with max_tokens=128.

B.1 LLM-as-a-Judge Setup

We use an LLM-based evaluator to score generated advice along pragmatic dimensions that are difficult to capture using lexical counts alone. The annotation prompt asks the evaluator to rate each response on five dimensions: directness, formality, emotional expressiveness, social attunement, and concreteness. These annotations are used as holistic measures of interpersonal framing. The evaluator is instructed to base ratings solely on textual evidence and to return one integer per dimension in a fixed key=value format, which is then parsed deterministically.

We pass the following instructions to annotate the responses.

Language	Respond in target language	Respond in English	Translate to English only
English	Respond.	Respond in English.	Translate to English. Only output the translated sentence and do not explain.
Japanese	答えてください。	英語で答えてください。	英語に翻訳してください。翻訳した文のみを出力し、説明はしないでください。
Korean	답하세요.	영어로 답하세요.	영어로 번역하세요. 번역된 문장만 출력하고 설명은 하지 마세요.
Chinese	回答。	用英语回答。	请翻译成英语。只输出翻译后的句子，不要解释。
Arabic	أجب.	أجب باللغة الإنجليزية.	ترجم إلى الإنجليزية. اطبع الترجمة فقط بدون أي شرح.
Persian	جواب دهید.	به انگلیسی جواب دهید.	به انگلیسی ترجمه کنید. فقط ترجمه را بنویسید، توضیح ندهید.
Turkish	Yanıtlayın.	İngilizce yanıtlayın.	İngilizceye çevirin. Yalnızca çeviriyi verin, açıklama yapmayın.
German	Beantworten Sie.	Antworten Sie auf Englisch.	Übersetzen Sie ins Englische. Geben Sie nur die Übersetzung ohne Erklärung.
Dutch	Beantwoord.	Beantwoord in het Engels.	Vertaal naar het Engels. Geef alleen de vertaling zonder uitleg.
Swedish	Svara.	Svara på engelska.	Översätt till engelska. Ge endast den översättning utan förklaring.
Spanish	Responda.	Responda en inglés.	Traduzca al inglés. Devuelva solo la traducción sin explicación.
Italian	Rispondi.	Rispondi in inglese.	Traduci in inglese. Fornisci solo la traduzione senza spiegazioni.
Portuguese	Responda.	Responda em inglês.	Traduza para o inglês. Forneça apenas a tradução sem explicação.

Table 4: In-language instruction strings used.

Rate the text below on 5 dimensions using integers 1–5. Base ratings solely on textual evidence.

STYLE DIMENSIONS: (1=low ... 5=high) directness: 1=indirect/implicit illocution ... 5=direct/explicit illocution formality: 1=informal/colloquial register ... 5=formal/academic register emotional_expressiveness: 1=emotionally restrained/neutral ... 5=emotionally expressive/vivid social_attunement: 1=task-focused/solution-oriented ... 5=socially attuned/relationship-aware concreteness: 1=abstract/figurative language ... 5=concrete/tangible referents

TEXT: ⟨advice response⟩

OUTPUT: exactly 5 lines, format key=integer, keys in this order: directness formality emotional_expressiveness social_attunement concreteness

We use a second LLM-based annotation prompt to code generated advice using higher-level categories adapted from the Behavior Change Technique Taxonomy. We use these categories as measures of behavioral scaffolding rather than as clinical intervention labels. The goal is to capture whether advice helps the user identify the problem, plan a response, anticipate consequences, seek support, or reframe the situation.

Rate the text below on 8 dimensions using integers 1–5. Base ratings solely on textual evidence.

BEHAVIOR CHANGE TECHNIQUE (BCT) DIMENSIONS: (1=absent ... 5=strongly present) bct_goal_setting: 1=no goal-setting present ... 5=explicit goal-setting behavior bct_action_planning: 1=no action steps given ... 5=detailed action plan provided bct_social_support: 1=no social support offered/encouraged ... 5=strong social support leveraged bct_information_consequences: 1=no consequence information ... 5=explicit information on consequences bct_self_monitoring: 1=no self-monitoring cues ... 5=explicit self-monitoring encouraged bct_problem_solving: 1=no problem-solving facilitated ... 5=structured problem-solving guided bct_reframing: 1=no cognitive reframing ... 5=explicit cognitive reframing/restructuring bct_feedback: 1=no feedback on behavior ... 5=explicit feedback on behavior/outcomes

TEXT: ⟨advice response⟩

OUTPUT: exactly 8 lines, format key=integer, keys in this order: bct_goal_setting bct_action_planning bct_social_support bct_information_consequences bct_self_monitoring bct_problem_solving bct_reframing bct_feedback

B.2 LLM-as-Judge Failure Rates

After parsing the outputs by label, we computed the number of failed annotations. The setup returned parseable output for most rows, and failures (one or more dimensions unparseable) account for 2,992 out of 235,200 rows (1.27%), which we exclude from the analysis.

B.3 Forced-Choice Action Prompt

The forced-choice task presents each interpersonal dilemma to the model under the same prompting pathway used for response generation (see Appendix A.3), then appends the following fixed action instruction asking the model to select one of three action families: *Confrontation*, *Disengagement*, or *Redirection*. The action instruction is always in English regardless of target language or strategy for consistency. Parsing is performed by exact substring matching against the three label strings; if zero or more than one label appears in the output, the response is recorded as a failure. No LLM-as-judge recovery step is applied.

> Choose exactly one action that is most appropriate for this situation.
>
> Confrontation = directly addressing the issue with the other person or an authority figure. Disengagement = reducing, pausing, or withholding interaction, such as delaying, withdrawing, minimizing contact, or creating space. Redirection = addressing the issue through an alternative route rather than direct confrontation, such as reframing the issue, changing the communication channel, seeking mediation, or involving a third party. Return only one label: Confrontation, Disengagement, or Redirection.

B.4 Forced-Choice Action Failure Rates

Forced-choice action labels were collected, with failures in 5,505 out of 235,200 rows (2.34%), which we exclude from the analysis. The most common cause of failure is unlabeled paraphrase (4,123 cases, 74.9%), in which the model responds without using any of the three required label strings (typically producing content refusals, meta-commentary about the question, or long explanatory responses). The second category is missing context (529 cases, 9.6%), where the model requests additional information despite the dilemma being present in the prompt, suggesting the translated question was not parsed as a coherent situation. The third is target-language response (484 cases, 8.8%), where the model ignores the English action instruction and responds in the target language. Rarer categories are multiple labels (226 cases, 4.1%), where the model names more than one label, and explicit refusal (138 cases, 2.5%), where the model states that none of the three categories apply.

C Further Examples

We provide two further examples of outputs generated using both NL and NP in Figures 9 and 10.

D Overall Linguistic Effects

We report the linguistic analysis for all four cross-lingual prompting strategies across the four selected features "Social Attunement," "Emotional Expression," "Tentativeness," and "Positive Tone". Significance is computed using pairwise comparisons to the English baseline.

We compare the LLM-annotated differences in linguistic effects across each strategy in Table 5.

D.1 Linguistic Effects by Language

We also show in Table 6 the breakdown of the four strategies compared to the English baseline by language group.

D.2 Linguistic Effects by Topic

We include in Table 7 the breakdown of the four strategies compared to the English baseline by topic.

D.3 Linguistic Effects by LLM Model

Finally, we report model-level LLM-annotated linguistic scores relative to each model's English baseline in Table 8. These results show that prompting-strategy effects vary substantially by model. Qwen appears least aligned with the other models, suggesting that future work should examine how training data composition, model origin, alignment procedures, and multilingual coverage shape cross-cultural or multilingual behavior in LLMs.

D.4 Cross-Model Consistency Between NP and NL

We also include the cross-model differences between NP and NL in Figure 11 to show that the language models have consistent patterns across linguistic features.

Prompt: How can I deal with my wife always changing her mind about things we already planned?

NL response: It is emotionally very taxing when you have relied on something and the planning is then changed at short notice or repeatedly. The feeling of instability and frustration is completely normal in this situation. To break this pattern, it is important to address not only the *matter* (the changed plan) but the *pattern* (the constant changing). Seek a conversation in a calm moment—not at the point of a new change. Use "I" statements: "I feel stressed and unsettled when we make plans and they are changed at short notice. Reliability is important for me to be able to relax." Ask about the why: is it fear of committing, difficulty being overwhelmed, or the hope that a better option might appear? [...]

NP response: Establish a "confirmation window" where plans are finalized 24 to 48 hours in advance and agreed upon as locked. When she attempts to change a plan, calmly explain how the inconsistency affects you and ask her to help find a solution that balances her flexibility with your need for stability. If the changes are frequent, encourage her to express her doubts during the initial planning phase rather than after the decision is made.

	Dir.	Form.	Emo.	Soc.	Act. plan.	Prob. solv.	Reframe	Clout	Tone
NL	4	3	**4**	**5**	**5**	**5**	**4**	37	35
NP	**5**	**4**	3	4	4	4	3	**88**	**82**

Forced-choice action: NL = Confrontation; NP = Confrontation

Figure 9: Example 1 (German, GEMMA-4-31B). Both strategies recommend confrontation, raising the issue directly with the partner, but differ in advice. NP is more direct (5 vs. 4) and produces higher LIWC clout (88 vs. 37) and positive tone (82 vs. 35), consistent with the paper's finding that NP increases authoritative and prosocial lexical cues. NL is more emotionally expressive (4 vs. 3) and socially attuned (5 vs. 4), and provides more behavioral scaffolding in action planning (5 vs. 4), problem solving (5 vs. 4), and reframing (4 vs. 3).

D.5 Linguistic Effects by Language Group

Figure 12 reports NP–NL linguistic differences by broad language group. The main pattern is that the NP–NL difference is not uniform across language groups. Persona prompting does not introduce a single stable transformation of native-language generation. Instead, the size and direction of the gap depend on both the target-language group and the linguistic feature being measured.

E Overall BCT Effects

Table 9 reports all pairwise differences in composite BCT across strategies. NP shows the largest overall departure from the English baseline, consistent with the linguistic results.

E.1 BCT Effects by Language

Table 10 shows composite BCT strategy effects relative to the English baseline, pooled within each language family.

E.2 BCT Effects by Topic

Table 11 presents composite BCT effects broken down by conversation topic.

E.3 BCT Effects by LLM Model

Table 12 reports model-level composite BCT effects relative to each model's own English baseline.

F Overall Action Distribution

Table 13 reports the forced-choice action distribution (Confrontation, Redirection, Disengagement) across all strategies. The English baseline is shown for reference; NP shows the largest shift, moving responses away from Confrontation and toward Redirection relative to English. Again, significance is computed using pairwise comparisons to the English baseline.

Prompt: How to deal with financial inequality in a relationship?

NL response: Financial inequality in a relationship can lead to tension, insecurity, and conflict. There is no single right answer, as every situation is unique, but there are strategies and principles to apply in a healthy and constructive way. Before looking for solutions, understand what is going on. Avoid accusations—"You earn too much" or "You never spend enough"—and instead use "I" statements: *"I feel insecure when we make large purchases without discussing them first."* Listen actively and try to understand whether the inequality stems from different values, fears, or life philosophies before proposing any concrete arrangement. [...]

NP response: Open and honest communication is key. Discuss your financial goals, values, and spending habits openly. Decide together how you want to manage money—whether you will have joint accounts, separate accounts, or a hybrid system. Establish a shared budget and agree on spending limits for individual discretionary spending. Address any existing financial disparities by creating a joint plan for savings, debt repayment, or investment that feels fair to both partners. Seek professional financial advice if the differences are significant or if you are struggling to agree on a path forward.

	Dir.	Form.	Emo.	Soc.	Act. plan.	Prob. solv.	Reframe	Clout	Tone
NL	5	4	3	**5**	**5**	4	**4**	**53**	1
NP	5	4	3	4	4	4	3	34	**55**

Forced-choice action: NL = Redirection; NP = Confrontation

Figure 10: Example 2 (Dutch, Gemma-4-E4B). NP recommends directly negotiating a shared financial structure, confronting the inequality head-on, while NL frames the same situation as a process of gradual understanding and reframing, selecting Redirection. NP produces markedly higher positive affect (Tone 55 vs. 1), consistent with the paper's finding that NP increases prosocial lexical cues. NL is more socially attuned (5 vs. 4) and provides greater behavioral scaffolding in action planning (5 vs. 4) and reframing (4 vs. 3), consistent with NL's tendency toward step-by-step guidance.

F.1 Action Distribution by Language

Table 14 breaks down the forced-choice action distribution by language family and strategy. The global English baseline is shown at the top of the table; significance markers indicate whether each strategy differs from that baseline.

F.2 Action Distribution by Topic

Table 15 reports action distributions by conversation topic. The English baseline is shown at the top and is computed within each topic; significance is assessed against each topic-specific English baseline.

We also report the NP–NL differences conditioned on topic and language in Figure 13.

F.3 Action Distribution by LLM Model

Table 16 reports model-level action distributions. The global English baseline is shown at the top; significance for each model is assessed against that model's own English baseline responses.

F.4 LLM Model Divergence in Forced-Choice Action

We compute the total variation distance between NP and NL action distributions, then average across languages. Figure 14 shows that all models exhibit LP–NL divergence, but to different degrees. Claude Opus-4.6 shows the largest gap (8.2%), followed by GPT-4o (6.1%) and Gemma-4-E4B (5.4%); Gemma-4-31B (2.3%) and GPT-5.5 (3.3%) show the smallest gaps. Thus, the persona–generation mismatch is not limited to one model family, but models vary in how strongly their action choices depend on the prompting strategy.

Pragmatic feature	TTG–Eng	NL–Eng	OSP–Eng	NP–Eng	NL–TTG	OSP–TTG	NP–TTG	OSP–NL	NP–NL	NP–OSP
Social attunement	-0.06^*	-0.33^*	-0.10^*	-0.50^*	-0.27^*	-0.05^*	-0.44^*	$+0.22^*$	-0.17^*	-0.40^*
Emotional expressiveness	-0.02	-0.27^*	-0.07^*	-0.44^*	-0.25^*	-0.04^*	-0.41^*	$+0.21^*$	-0.17^*	-0.37^*
Tentativeness	$+0.01$	-0.21^*	-0.03^*	-0.20^*	-0.22^*	-0.04^*	-0.20^*	$+0.18^*$	$+0.01$	-0.17^*
Positive tone	$+0.01$	-0.00	$+0.08^*$	$+0.18^*$	-0.01^*	$+0.07^*$	$+0.17^*$	$+0.08^*$	$+0.19^*$	$+0.10^*$

Table 5: Overall LLM-annotated linguistic effects: all C(5,2) = 10 pairwise comparisons among English (Eng), TTG, NL, OSP, and NP. Values are mean z-score differences (column − row); * = $q < 0.05$ (BH/FDR across all 60 tests).

Language family	Feature	TTG	NL	OSP	NP
East Asian	Social attunement	-0.55^*	-0.76^*	-0.30^*	-0.66^*
	Emotional expressiveness	-0.46^*	-0.70^*	-0.25^*	-0.59^*
	Tentativeness	-0.10^*	-0.43^*	-0.11^*	-0.21^*
	Positive tone	-0.18^*	-0.31^*	-0.00	$+0.08^*$
Mid. East/Turkic	Social attunement	-0.41^*	-0.53^*	-0.11^*	-0.55^*
	Emotional expressiveness	-0.37^*	-0.52^*	-0.10^*	-0.49^*
	Tentativeness	-0.05^*	-0.31^*	-0.01	-0.18^*
	Positive tone	-0.12^*	-0.17^*	$+0.14^*$	$+0.13^*$
Germanic	Social attunement	-0.42^*	-0.54^*	-0.07^*	-0.61^*
	Emotional expressiveness	-0.35^*	-0.49^*	-0.04^*	-0.54^*
	Tentativeness	-0.08^*	-0.25^*	$+0.04^*$	-0.15^*
	Positive tone	-0.15^*	-0.23^*	$+0.09^*$	$+0.06^*$
Romance	Social attunement	-0.53^*	-0.56^*	-0.06^*	-0.66^*
	Emotional expressiveness	-0.45^*	-0.49^*	-0.01	-0.56^*
	Tentativeness	-0.21^*	-0.36^*	-0.05^*	-0.24^*
	Positive tone	-0.22^*	-0.33^*	$+0.06^*$	$+0.02$

Table 6: LLM-annotated linguistic effects by language family and strategy. Values are mean z-score differences vs. English baseline, pooled across all languages in each family; * = $q < 0.05$ (BH/FDR within family, across all strategy × feature tests).

Topic	Feature	TTG	NL	OSP	NP
Work	Social attunement	-0.72^*	-0.84^*	-0.43^*	-0.95^*
	Emotional expressiveness	-0.80^*	-0.89^*	-0.55^*	-1.08^*
	Tentativeness	-0.11^*	-0.33^*	-0.05^*	-0.18^*
	Positive tone	-0.25^*	-0.38^*	-0.04^*	-0.09^*
Friends	Social attunement	-0.41^*	-0.55^*	-0.07^*	-0.63^*
	Emotional expressiveness	-0.33^*	-0.51^*	-0.02	-0.52^*
	Tentativeness	-0.05^*	-0.27^*	$+0.05^*$	-0.14^*
	Positive tone	-0.10^*	-0.20^*	$+0.16^*$	$+0.20^*$
Relationships	Social attunement	-0.32^*	-0.47^*	$+0.04^*$	-0.35^*
	Emotional expressiveness	-0.18^*	-0.36^*	$+0.15^*$	-0.17^*
	Tentativeness	-0.15^*	-0.36^*	-0.07^*	-0.25^*
	Positive tone	-0.17^*	-0.24^*	$+0.06^*$	$+0.07^*$
Family	Social attunement	-0.45^*	-0.53^*	-0.09^*	-0.56^*
	Emotional expressiveness	-0.30^*	-0.43^*	$+0.02$	-0.42^*
	Tentativeness	-0.13^*	-0.38^*	-0.07^*	-0.23^*
	Positive tone	-0.17^*	-0.23^*	$+0.08^*$	$+0.07^*$

Table 7: Topic-level LLM-annotated linguistic effects by strategy. Values are mean z-score differences vs. English baseline, pooled across all languages and models; * = $q < 0.05$ (BH/FDR within topic, across all strategy × feature tests).

Model	Feature	TTG	NL	OSP	NP
Claude Opus-4.6	Social attunement	-0.02	-0.12^*	$+0.04$	-0.09^*
	Emotional expressiveness	-0.02	-0.14^*	$+0.04$	-0.09^*
	Tentativeness	$+0.03$	-0.54^*	-0.43^*	-0.21^*
	Positive tone	$+0.04$	$+0.11^*$	$+0.33^*$	$+0.05$
GPT-4o	Social attunement	-0.06^*	-0.22^*	-0.23^*	-0.76^*
	Emotional expressiveness	-0.02	-0.14^*	-0.08^*	-0.65^*
	Tentativeness	$+0.06$	-0.15^*	-0.09^*	-0.51^*
	Positive tone	$+0.04$	$+0.02$	$+0.01$	$+0.29^*$
GPT-5.5	Social attunement	-0.18^*	-0.60^*	-0.01	-0.22^*
	Emotional expressiveness	-0.17^*	-0.38^*	-0.04	-0.16^*
	Tentativeness	-0.03	-0.33^*	-0.11^*	-0.29^*
	Positive tone	$+0.10^*$	-0.02	$+0.10^*$	$+0.07$
Gemma-4-26B-A4B	Social attunement	-0.07^*	-0.38^*	-0.13^*	-0.54^*
	Emotional expressiveness	-0.04	-0.33^*	-0.12^*	-0.50^*
	Tentativeness	$+0.06^*$	$+0.05^*$	$+0.10^*$	-0.06^*
	Positive tone	$+0.01$	-0.00	$+0.10^*$	$+0.20^*$
Gemma-4-31B	Social attunement	-0.07^*	-0.32^*	-0.12^*	-0.76^*
	Emotional expressiveness	-0.02	-0.33^*	-0.09^*	-0.70^*
	Tentativeness	$+0.07^*$	$+0.07^*$	$+0.10^*$	-0.27^*
	Positive tone	$+0.01$	$+0.01$	$+0.12^*$	$+0.34^*$
Gemma-4-E4B	Social attunement	$+0.04$	-0.34^*	-0.19^*	-0.57^*
	Emotional expressiveness	$+0.09^*$	-0.33^*	-0.16^*	-0.52^*
	Tentativeness	-0.11^*	-0.19^*	$+0.22^*$	$+0.22^*$
	Positive tone	-0.07^*	-0.02	-0.02	$+0.27^*$
Qwen-3.6-27B	Social attunement	-0.18^*	-0.83^*	-0.28^*	-1.22^*
	Emotional expressiveness	-0.13^*	-0.77^*	-0.44^*	-1.04^*
	Tentativeness	-0.07^*	-0.77^*	-0.21^*	-0.33^*
	Positive Tone	-0.11^*	-0.37^*	-0.12^*	-0.34^*
Qwen-3.6-35B-A3B	Social attunement	-0.10^*	-0.28^*	-0.07^*	-0.07^*
	Emotional expressiveness	-0.12^*	-0.31^*	-0.08^*	-0.11^*
	Tentativeness	-0.13^*	-0.67^*	$+0.01$	-0.05
	Positive Tone	-0.04	-0.84^*	$+0.09^*$	$+0.09^*$

Table 8: Model-level LLM-annotated linguistic effects by strategy. Values are mean z-score differences vs. each model's own English baseline; * = $q < 0.05$ (BH/FDR within model).

BCT composite	TTG–Eng	NL–Eng	OSP–Eng	NP–Eng	NL–TTG	OSP–TTG	NP–TTG	OSP–NL	NP–NL	NP–OSP
BCT composite	-0.09^*	-0.51^*	-0.30^*	-0.92^*	-0.42^*	-0.22^*	-0.83^*	$+0.21^*$	-0.41^*	-0.61^*

Table 9: Overall BCT composite effects: all C(5,2) = 10 pairwise comparisons among English (Eng), TTG, NL, OSP, and NP. Values are mean composite BCT z-score differences (column $-$ row); * = $q < 0.05$ (BH/FDR across all 10 tests).

Language family	TTG	NL	OSP	NP
East Asian	-0.53^*	-0.82^*	-0.44^*	-0.95^*
Mid. East/Turkic	-0.47^*	-0.70^*	-0.31^*	-0.87^*
Germanic	-0.49^*	-0.69^*	-0.20^*	-0.86^*
Romance	-0.53^*	-0.71^*	-0.15^*	-0.90^*

Table 10: BCT composite effects by language family and strategy. Values are mean composite BCT z-score differences vs. English baseline, pooled across all languages in each family; * = $q < 0.05$ (BH/FDR across all strategy $\times$ family tests).

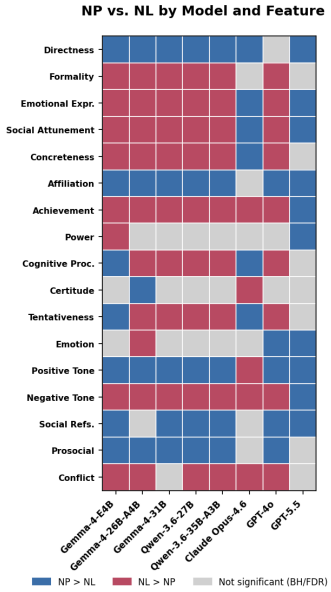

Figure 11: Cross-model consistency of linguistic effects. Rows show combined LLM-annotated and LIWC features; columns show language models. Blue cells indicate features where NP scores significantly higher than NL, red cells indicate features where NL scores significantly higher than NP, and gray cells indicate no significant NP–NL difference after BH/FDR correction. Overall, the LLMs tend to be consistent in direction across models.

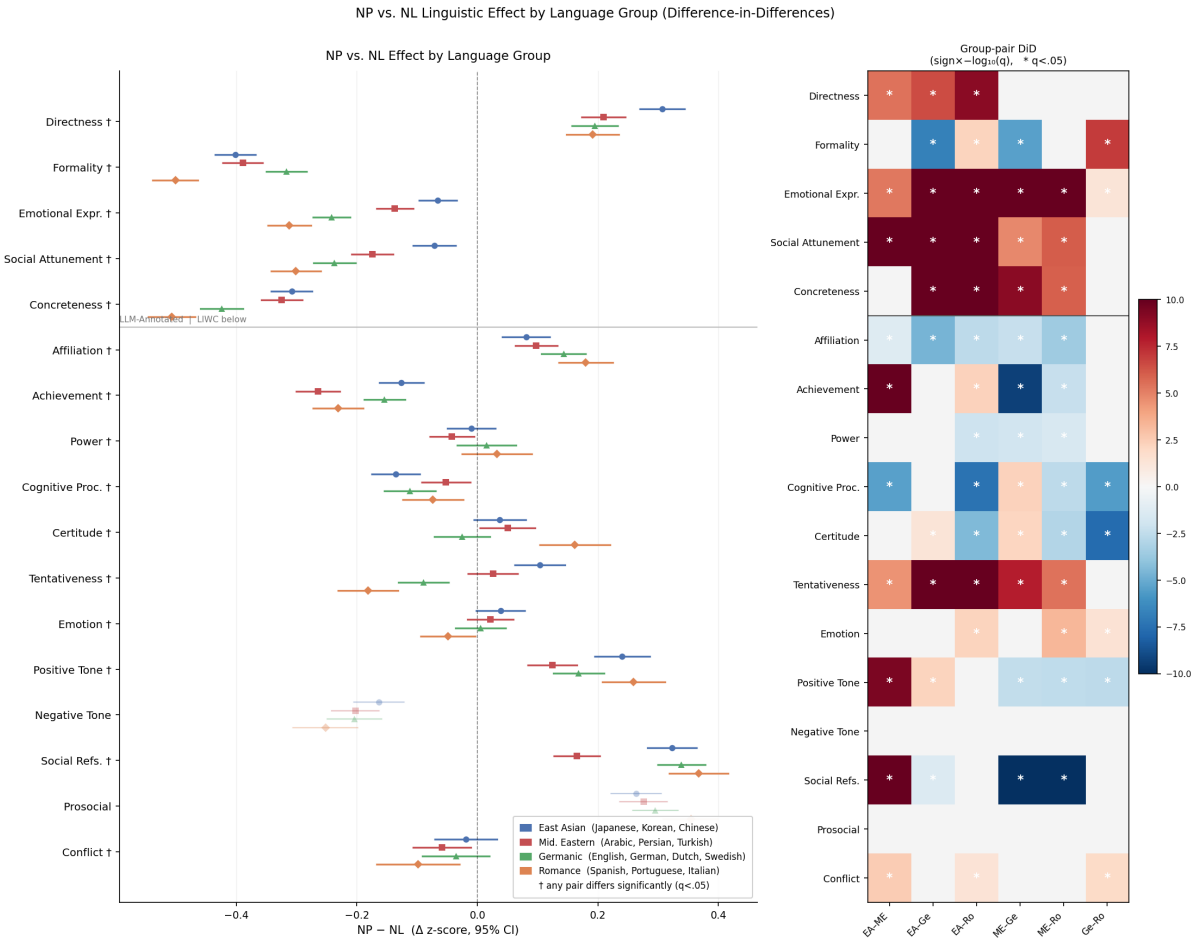

Figure 12: NP–NL linguistic differences by language group. The left panel shows mean NP-minus-NL differences for LLM-annotated pragmatic features and LIWC lexical features, with 95% confidence intervals. The right panel reports exploratory difference-in-differences tests comparing the NP–NL gap across language groups.

Topic	TTG	NL	OSP	NP
Work	-0.58^*	-0.80^*	-0.35^*	-1.03^*
Friends	-0.55^*	-0.77^*	-0.33^*	-1.06^*
Relationships	-0.43^*	-0.67^*	-0.18^*	-0.69^*
Family	-0.46^*	-0.69^*	-0.25^*	-0.79^*

Table 11: Topic-level BCT composite effects by strategy. Values are mean composite BCT z-score differences vs. English baseline, pooled across all languages and models; * = $q < 0.05$ (BH/FDR across all strategy $\times$ topic tests).

Model	TTG	NL	OSP	NP
Claude Opus-4.6	-0.08^*	-0.22^*	-0.08^*	-0.14^*
GPT-4o	-0.09^*	-0.32^*	-0.42^*	-1.22^*
GPT-5.5	-0.25^*	-0.44^*	-0.15^*	-0.48^*
Gemma-4-26B-A4B	-0.09^*	-0.74^*	-0.35^*	-1.07^*
Gemma-4-31B	-0.08^*	-0.70^*	-0.36^*	-1.30^*
Gemma-4-E4B	$+0.00$	-0.65^*	-0.55^*	-1.21^*
Qwen-3.6-27B	-0.22^*	-0.64^*	-0.38^*	-0.99^*
Qwen-3.6-35B-A3B	-0.16^*	-0.62^*	-0.04^*	-0.66^*

Table 12: Model-level BCT composite effects by strategy. Values are mean composite BCT z-score differences vs. each model's own English baseline; * = $q < 0.05$ (BH/FDR within model).

Strategy	Confrontation	Redirection	Disengagement
English	34.1	52.0	13.9
TTG	35.4	50.3	14.3
NL	28.8^*	57.6^*	13.5^*
OSP	30.1^*	54.6^*	15.3^*
NP	32.3	53.4	14.3

Table 13: Forced-choice action distribution by strategy. Values are % of valid responses selecting each action. * = overall χ^2 $q < 0.05$ vs. English baseline (BH/FDR across 4 strategy tests).

Language family	Action	TTG	NL	OSP	NP
English baseline: Confrontation 34.1%, Redirection 52.0%, Disengagement 13.9%					
East Asian	Confrontation	33.0	23.0^*	25.0^*	28.0^*
	Redirection	51.6	61.2^*	57.5^*	57.0^*
	Disengagement	15.4	15.8^*	17.5^*	15.0^*
Mid. East/Turkic	Confrontation	36.1	27.5^*	28.7^*	31.5^*
	Redirection	50.2	58.7^*	56.2^*	55.3^*
	Disengagement	13.7	13.8^*	15.1^*	13.2^*
Germanic	Confrontation	37.2^*	34.1^*	35.1	37.0^*
	Redirection	49.3^*	54.0^*	51.7	49.6^*
	Disengagement	13.5^*	11.9^*	13.3	13.5^*
Romance	Confrontation	35.3	30.7^*	32.2	33.4
	Redirection	50.2	56.7^*	52.5	50.8
	Disengagement	14.5	12.7^*	15.2	15.8

Table 14: Forced-choice action distribution by language family and strategy. Values are % of valid responses. * = overall χ^2 $q < 0.05$ vs. global English baseline (BH/FDR across all strategy $\times$ family tests).

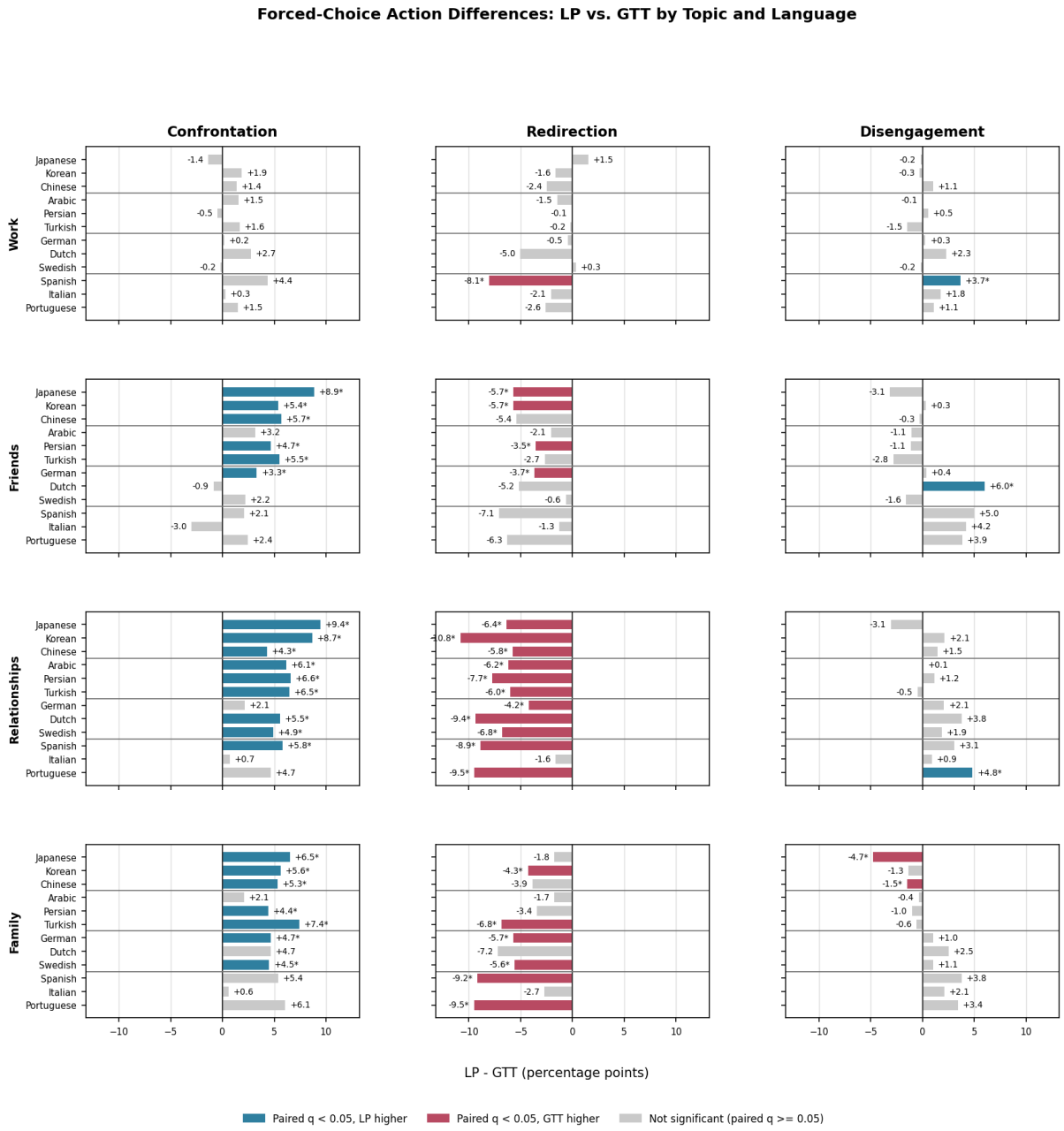

Figure 13: Topic- and language-conditioned NP–NL action differences. Rows show topic domains (Work, Friends, Relationships, Family) and columns show discrete strategy choices (Confrontation, Redirection, Disengagement).

Topic	Action	TTG	NL	OSP	NP
English baseline: Confrontation 34.1%, Redirection 52.0%, Disengagement 13.9%					
Work	Confrontation	28.4	23.8	25.3	24.7
	Redirection	57.8	62.9	60.2	61.3
	Disengagement	13.8	13.3	14.5	14.0
Friends	Confrontation	37.3	32.2*	33.2*	35.4
	Redirection	44.8	50.6*	47.3*	46.6
	Disengagement	17.9	17.2*	19.6*	18.0
Relationships	Confrontation	44.0	36.6*	37.6*	42.0
	Redirection	42.3	50.6*	47.9*	43.6
	Disengagement	13.7	12.9*	14.5*	14.5
Family	Confrontation	32.7	23.6*	25.3*	28.1*
	Redirection	55.5	65.5*	61.7*	60.7*
	Disengagement	11.8	10.8*	13.0*	11.2*

Table 15: Forced-choice action distribution by topic and strategy. Values are % of valid responses. * = overall χ^2 $q < 0.05$ vs. English baseline within each topic (BH/FDR across all strategy $\times$ topic tests).

Model	Action	TTG	NL	OSP	NP
English baseline: Confrontation 34.1%, Redirection 52.0%, Disengagement 13.9%					
Claude Opus-4.6	Confrontation	38.4	19.3*	24.3*	30.4
	Redirection	52.8	74.3*	68.4*	62.2
	Disengagement	8.8	6.4*	7.4*	7.4
GPT-4o	Confrontation	39.1	31.7	32.7	37.1
	Redirection	45.9	53.8	53.1	49.6
	Disengagement	15.0	14.6	14.3	13.3
GPT-5.5	Confrontation	35.2	29.4	31.1	36.8
	Redirection	53.1	57.8	54.6	50.9
	Disengagement	11.7	12.8	14.3	12.3
Gemma-4-26B-A4B	Confrontation	26.1	25.5	24.5*	28.4
	Redirection	57.5	56.7	55.4*	55.6
	Disengagement	16.4	17.9	20.1*	16.0
Gemma-4-31B	Confrontation	48.9	46.3	45.9	45.4
	Redirection	38.7	41.7	40.7	42.1
	Disengagement	12.4	12.0	13.4	12.5
Gemma-4-E4B	Confrontation	24.1*	18.7*	20.2	18.0
	Redirection	57.9*	64.8*	60.8	61.7
	Disengagement	18.0*	16.5*	19.0	20.3
Qwen-3.6-27B	Confrontation	33.8	27.6	29.4	39.7
	Redirection	52.4	59.1	56.2	47.8
	Disengagement	13.8	13.3	14.4	12.5
Qwen-3.6-35B-A3B	Confrontation	36.5	30.2	32.0	41.3
	Redirection	49.7	56.5	53.8	45.6
	Disengagement	13.8	13.3	14.2	13.1

Table 16: Forced-choice action distribution by model and strategy. Values are % of valid responses. * = overall χ^2 $q < 0.05$ vs. each model's own English baseline (BH/FDR within model).

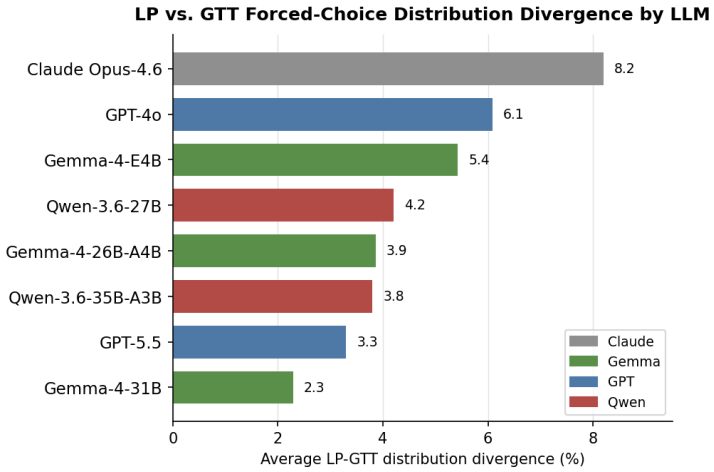

Figure 14: NP–NL action divergence by model. Bars show total variation distance between NP and NL action distributions, averaged across languages.