Your Mileage May Vary:
How Empathy and Demographics Shape Human Preferences in LLM Responses

Anonymous ACL submission

Abstract

As large language models (LLMs) increasingly assist in subjective decision-making (e.g., moral reasoning, advice), it is critical to understand whose preferences they align with—and why. While prior work uses aggregate human judgments, demographic variation and its linguistic drivers remain underexplored. We present a comprehensive analysis of how demographic background and empathy level correlate with preferences for LLM-generated dilemma responses, alongside a systematic study of predictive linguistic features (e.g., agency, emotional tone). Our findings reveal significant demographic divides and identify markers (e.g., power verbs, tentative phrasing) that predict group-level differences. These results underscore the need for demographically informed LLM evaluation.

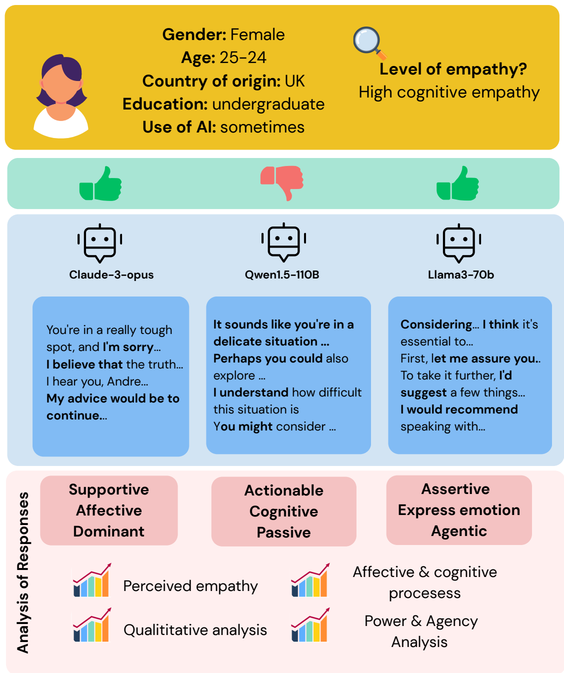


Figure 1: Roadmap of our study. We integrate demographic and empathy-level data (yellow) with human ratings of LLM responses (green), and connect these ratings to the content and linguistic features of the responses (blue and pink), enabling a multi-dimensional analysis of how user characteristics and response styles influence perceived empathy.

1 Introduction

Large language models (LLMs) are increasingly used in subjective domains such as moral reasoning and personalized advice-giving (Wang et al., 2023; Stade et al., 2024). The growing impact of LLMs on decisions and communication highlights the need to align them with diverse human preferences. Prior work in the NLP community has highlighted variations in human annotations for subjective tasks (Ovesdotter Alm, 2011; Basile, 2022; Plank, 2022), reflecting underlying differences in individual or group-level preferences. While it was once common to treat a single "gold standard" label as the ultimate target for model alignment, recent work in subjective NLP evaluation increasingly emphasizes demographic representation and diverse perspectives (Sorensen et al., 2024).

Among subjective tasks, recent work has evaluated LLM responses to emotional or ethical dilemmas using human judgments (Zhao et al., 2024; Verga et al., 2024; Zhao et al., 2025). However, important gaps remain in understanding how factors such as participant demographics, and personality traits influence preference rankings, and what linguistic features drive these differences. Although several studies have collected human judgment data (e.g. Cercas Curry and Rieser, 2019; Mostafazadeh Davani et al., 2024; Kirk et al., 2024, among others), these resources remain underexplored, particularly in terms of how preferences vary across users belonging to different demographic groups, what personality traits or other factors, and which language patterns may contribute to such variation.

In the context of responding to emotional users, the issue of generating empathetic responses has received increased attention (e.g. Rashkin et al., 2019) and systems displaying empathy have been shown to generally increase user satisfaction (Rostami and Navabinejad, 2023). However, it is unclear whether users always prefer more empathetic responses across the board or whether some users may prefer other styles. We present a comprehensive analysis of human ratings on LLM-generated responses, focusing on the influence of demographic factors and empathy. We examine how different user groups evaluate responses to moral dilemmas, and how these preferences vary across demographic and empathy lines to shape perceptions of LLM output. We explicitly distinguish between cognitive and affective empathy in our analysis, which helps reveal how different aspects of empathy relate to variation in response preferences across users (Jolliffe and Farrington, 2006). To the best of our knowledge, we are the first to analyze these questions.

To deepen our understanding, we analyze model responses using both quantitative and qualitative methods, leveraging tools such as Linguistic Inquiry and Word Count (LIWC) and Connotation Frames of Power and Agency to identify linguistic features associated with divergent preferences (Tausczik and Pennebaker, 2010; Sap et al., 2017). Finally, we offer actionable insights for adapting LLM communication styles to better align with the empathic expectations and demographic characteristics of diverse user populations. For example, we observe that the model most preferred by human annotators tends to adopt a more assertive and agentic tone, and groups that have lower empathy tend to favor models that show more prevalence in cognitive linguistic markers. By linking linguistic patterns to demographic preferences and type of empathy in users, this work equips developers to refine inclusivity in LLM design and challenges the assumption that "one-size-fits-all" outputs suffice for subjective tasks. Our findings highlight the need for human-conscious evaluation in LLMs, particularly as the field shifts toward subjective notions of quality (Meister et al., 2024).

2 Related Work

Recently, LLMs' affective abilities have come to the forefront, evidenced by the emergence of several benchmarks to measure them, such as EmoBench (Sabour et al., 2024) and EmotionQueen (Chen et al., 2024). In particular, much of this work has focused on generating empathetic responses: a 2024 metareview found that human raters find LLM-generated responses to be more empathetic than humans' (Sorin et al., 2024). Several later studies have compared LLM responses to dilemmas to humans', supporting previous findings that LLMs can respond empathetically (Huang et al., 2024; Wang et al., 2023; Lee et al., 2024; Welivita and Pu, 2024). Although subjective tasks in NLP are a growing area of interest, previous studies on LLM responses in this context have not considered individual preferences, despite it being a task with high disagreement (Manzoor et al., 2024). Welivita and Pu (2024) collect how empathetic humans are, but only as an attention check, and do not correlate it with response preferences. In this paper, we consider how the human raters' own tendency to empathise, the type of empathy they display, and their sociodemographic features affects response preferences.

In terms of preferences, only a few studies have considered what features of a response appeal to human judges and LLMs (Li et al., 2024). In the context of EQ and empathy, Lee et al. (2024) found that models have distinct response patterns that also differ from humans' by studying the surface features of each model by training a BOW classifier, and (Zhao et al., 2025) investigate different aspects of response quality, such as actionability and sympathy. We present a study aligning preferences with linguistics features.

3 Experimental Setup

Data We use the dataset developed for the Language Model Council (LMC) framework (Zhao et al., 2025), which supports collective evaluation of different LLMs on a task related to emotion intelligence. This dataset extends the emotionally charged dilemmas *EmoBench* benchmark (Sabour et al., 2024). Specifically, it consists of a total of 100 dilemmas (e.g., a family member asking for money when you are facing financial problems yourself). A council of 20 LLMs rank each other's open-ended responses to these interpersonal conflicts, resulting in a total of 2,000 evaluated responses.

After the generation of LLM responses, human participants were asked to evaluate pairs of these responses to assess the models' emotional intelligence. In each of the 1300+ comparisons, partic-

Demographic	CE	AE
Female (N = 113)	**3.00**± 0.43*	**3.18**± 0.51*
Male (N = 93)	2.87± 0.43*	2.97± 0.56*
UK (N = 181)	2.89± 0.59*	2.94± 0.67*
USA (N = 26)	**3.11**± 0.56*	**3.18**± 0.54*
Age 18–24 (N = 46)	3.05± 0.52*	**3.08**± 0.63
Age 25–34 (N = 73)	2.72± 0.63*	2.89± 0.64
Age 35–44 (N = 46)	**3.07**± 0.49	3.04± 0.61
Age 45-54 (N = 30)	2.96±0.45	3.02± 0.58
High school or below (N = 20)	**2.99**± 0.61	**3.19**± 0.57
Undergraduate (N = 121)	2.93± 0.59	2.91± 0.68
Graduate (N = 68)	2.87± 0.61	3.02± 0.64
AI use: Never (N = 24)	2.84± 0.42	2.84± 0.63
AI use: Rarely (N = 46)	2.97± 0.50	3.00± 0.67
AI use: Sometimes (N = 98)	**3.01**± 0.60	**3.09**± 0.58
AI use: Frequently (N = 50)	2.96± 0.68	3.06± 0.68

Table 1: Cognitive Empathy (CE) and Affective Empathy (AE) Scores by Demographic Categories. **Bold** values indicate the highest score within each dimension. * denotes significant differences ($p < 0.05$, t-test). All values are mean ± SD on a 1–5 Likert scale.

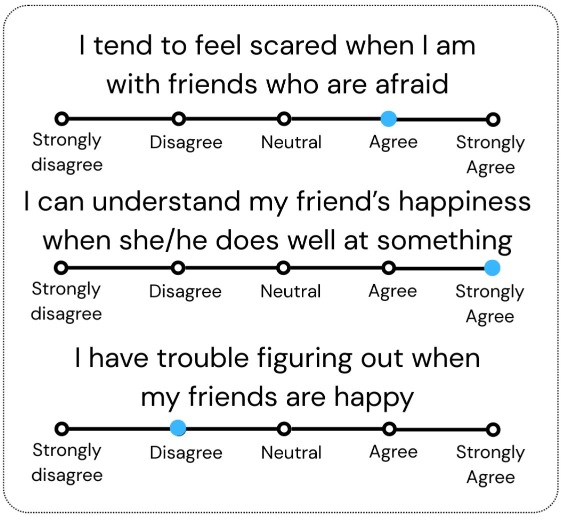


Figure 2: Examples of empathy assessment questionnaire sourced from Jolliffe and Farrington (2006). The first question assesses affective empathy. The second and third questions assess cognitive empathy.

ipants selected the response they considered the *better* (e.g., "Response A much better than B") and provided qualitative feedback on their choice. This feedback included labels describing emotional qualities, such as "the best response expressed emotions," "the system sympathized with the protagonist," and "the best response seems trustworthy." The full list of survey questions is provided in Figure 8 in Appendix A. While Zhao et al. (2025) compared LLM preferences to human responses using the collected ratings, it did not analyze the underlying human preferences reflected in the qualitative feedback. In this paper, we build on that data to investigate how demographic background and empathy level correlate with preferences for LLM-generated dilemma responses.

Models We use nine models from the LLM Council for our analysis of human response preferences. These models were selected by Zhao et al. (2025) from the original 20 council members for comparison against human evaluations. The models include Qwen1.5-110b-chat (Bai et al., 2023), gpt-4o-2024-05-13 (OpenAI, 2024), claude-3-opus (Anthropic, 2024), qwen1.5-32b-chat (Team, 2023), llama-3-70b-chat (Platforms, 2024), claude-3-haiku (Anthropic, 2024), mixtral-8x7b (Jiang et al., 2024), llama-3-8b-chat (Platforms, 2024), and gpt-4-0613 (OpenAI, 2023). For simplicity, we refer to them throughout the paper as Qwen1.5-110b, gpt-4o, claude-3-opus, qwen1.5-32b, llama-3-70b, claude-3-haiku, mixtral-8x7b, llama-3-8b, and gpt-4.

Participants demographics We use demographic data collected by Zhao et al. (2025), which includes age, gender, country of origin, education level, and frequency of AI use. However, the participant distribution across these categories is uneven, with some subgroups (e.g., "non-binary" in gender, "60+" in age) having very small sample sizes. To ensure the reliability of our analyses on empathy and demographic-specific rankings, we exclude any subgroup with fewer than 10 participants. In addition, we combine the "every day" and "nearly every day" AI use responses into a single "frequently" category. Table 1 presents the final demographic dimensions and subgroup sizes.

Empathy assessment Alongside demographic questions, participants filled in the Basic Empathy Scale (Jolliffe and Farrington, 2006), which we used to evaluate both *cognitive empathy* (CE) and *affective empathy* (AE) (examples shown in Figure 2). CE reflects the ability to understand others' perspectives, while AE refers to sharing others' emotional experiences. This distinction is key for a more precise evaluation and analysis of how different types of empathy relate to human judgments of dilemma-focused LLM responses.

The original questionnaire used a Likert scale ranging from strongly disagree to strongly agree. For our analysis, we converted these responses to a numerical scale from 1 to 5. Additionally, for

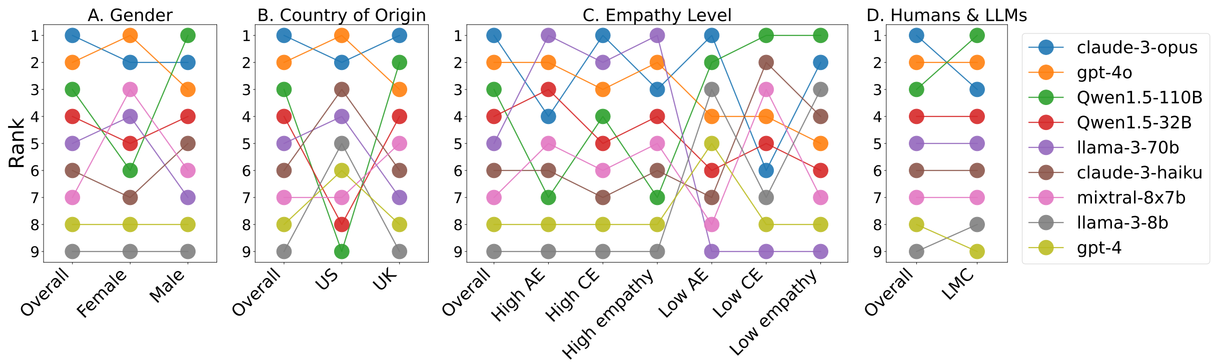


Figure 3: Ranks of LLMs across groups in selected demographic dimensions. We add "Overall" human ranking for reference on the first hue in each figure. "LMC" stands for ranking by language model council.

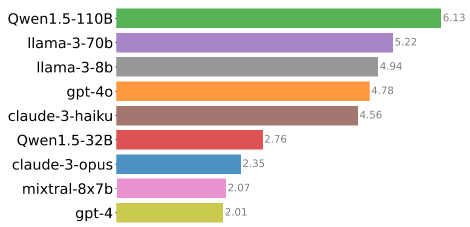


Figure 4: Variance in ranking across LLMs.

negatively worded items (e.g., "I have trouble figuring out when my friends are happy"), where lower agreement indicates higher empathy, we reversed the scores to ensure that higher values consistently reflect greater empathy. Based on the numerical scores, we calculated each participant's average empathy, cognitive empathy, and affective empathy scores. Participants were then divided into groups using a median split: low vs. high overall empathy, low vs. high cognitive empathy, and low vs. high affective empathy. This grouping allows us to explore how varying levels and types of empathy relate to participants' judgments of LLM responses to emotional dilemmas.

Empathy across demographics We compute the average self-reported empathy scores and standard deviations across demographic subgroups, and used statistical tests to assess significant differences between them (see Table 1). Female participants exhibit significantly higher empathy scores than male participants in both cognitive and affective dimensions ($t = 2.31$, $p = 0.02$), aligning with prior findings on gender differences in empathy (Jolliffe and Farrington, 2006). Although participants from the US report higher overall empathy levels than those from the UK ($t = 2.35$, $p = 0.02$), this result should be interpreted with caution due to the uneven sample sizes (UK: $n = 181$, US: $n = 26$).

To further examine the relationship between empathy and individual factors, we ran additional linear regression analyses using age, education level, and frequency of AI use as predictors. These models revealed no significant associations between empathy scores and any of these variables. However, we observe significant differences in cognitive empathy across age groups, as revealed by a one-way ANOVA ($F(3, 195) = 3.18$, $p = 0.025$), suggesting that some aspects of empathy may develop with age and life experience (Guariglia et al., 2023).

4 Human Ratings Analysis

We analyze human evaluators' ratings of LLM responses to dilemmas to uncover how preferences vary across demographic groups. Here we present the demographic-specific ranking patterns, overall trends in model preferences, and how empathy influences people's preferences.

Demographic-specific ranking Following Zhao et al. (2025), we compute demographic-specific model rankings using arena-style pairwise comparisons against a fixed reference model, Qwen-1.5-32B. The rationale and procedure for selecting the reference model are detailed in their paper. For each subgroup, we filter the human evaluation data accordingly and calculate rankings based on expected win rates using the ELO scoring system (Bai et al., 2022), with Bradley–Terry (BT) coefficients (Bradley and Terry, 1952) applied to improve statistical robustness.

Note that not all subgroups yield a complete set of pairwise results due to limited comparisons within some demographics. We present the key subgroup ranking patterns in Figure 3, and report full rankings for all subgroups in Figure 9 in Appendix B.

General trends in human preferences We observe substantial variability in model rankings across demographic dimensions including gender (Figure 3A) and country of origin (Figure 3B). To quantify this, we compute the variance of each model's ranking position across all demographic subgroups, as shown in Figure 4. The results show that Qwen1.5-110B exhibits the highest rank variance, indicating that its perceived quality varies widely across different demographic groups. It is followed by LLaMA-3-70B and LLaMA-3-8B, which also show considerable divergence in their rankings. **These findings suggest that certain models may appeal strongly to some subgroups while being less favored by others, highlighting the importance of analyzing preferences beyond the aggregate level.**

Despite the high variance in rankings across demographic groups, some models consistently perform well overall. As shown in Figure 3, Claude-3-Opus, GPT-4o, and Qwen1.5-110B constantly rank in top three, suggesting broader cross-demographic appeal. In contrast, smaller models such as GPT-4 and LLaMA-3-8B consistently rank lower, indicating a more limited alignment with users' expectations. These patterns suggest that model scale and architecture may play a role in perceived response quality, and that a few models manage to strike a balance across diverse user preferences.

Empathy-related divergence in ratings Figure 3C displays model rankings across empathy levels. Empathy shows a particularly strong influence on model preference. LLaMA-3-70B consistently ranks in the top two among high-empathy groups, although it only ranks in the fifth among all human participants. In the meantime, Qwen1.5-110B ranks the first or second among low-empathy groups. This pattern aligns with participants' feedback on perceived empathy: LLaMA-3-70B was described as "emotionally intelligent" in 82% of its evaluations, compared to just 58% for Qwen1.5-110B. A complete report of perceived empathy ratings is provided in Figure 7 (Appendix A).

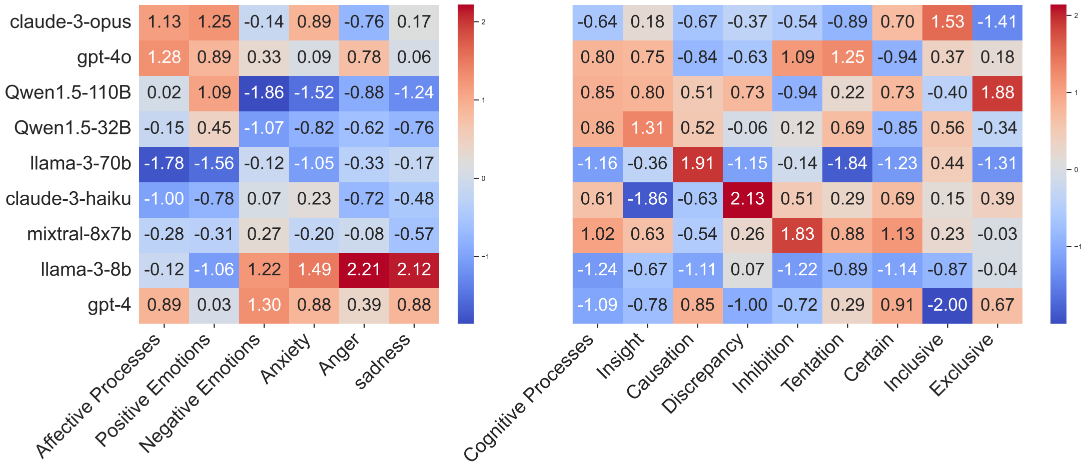
The results also point to distinct effects of cognitive empathy (CE) and affective empathy (AE). For instance, the low CE group ranks Claude-3-Haiku and Mixtral-8x7B among their top three models, despite these models typically receiving lower rankings overall. In contrast, the low AE group ranks both models near the bottom. **This contrast suggests that cognitive and affective empathy may drive attention to different qualities in model outputs**, such as emotional intelligence versus coherence or informativeness, highlighting the importance of treating cognitive and affective empathy as distinct dimensions in LLM evaluation, rather than collapsing them into a single measure.

Ranks across gender groups and countries of origin (3A, B) further underscore the influence of empathy-related factors. In particular, groups with higher average empathy, i.e. female and participants from the UK, tend to rank Qwen1.5-110B lower and LLaMA-3-70B higher. For example, Qwen1.5-110B's ranking spans up to seven positions between the UK and US groups. **This divergence reveals that cultural or national context may shape expectations around what constitutes a "better" response to a moral dilemma—whether it should be more actionable, emotionally attuned, or pragmatically useful.** Empathy likely plays a role in shaping these criteria, influencing how different groups interpret and value model behavior. This finding highlights the need to incorporate socio-demographic diversity into evaluation frameworks.

We also observe a notable divergence between the overall human ranking and the ranking from Language Model Council (i.e. a group of LLMs), as shown in Figure 3D. While both converge on the high-performing models, their top preferences differ: overall, human participants rank Claude-3-Opus highest, whereas LMC favors Qwen1.5-110B. **This indicates that human evaluators and language models may prioritize different qualities when assessing model responses.** Given that Qwen1.5-110B receives consistently lower rankings from high-empathy groups, its lower placement in human ranking suggests that it may lack features that resonate with human evaluators. This divergence highlights the value of human-centered evaluation, especially when models are intended for use in socially or emotionally sensitive contexts.

5 Linguistic Style Analysis of LLM responses

We investigate the factors underlying ranking differences by analyzing linguistic styles across LLMs' responses. Specifically, we examine affective and cognitive language features, characterize social roles through the connotation framework of Power



A. Affective Processes

B. Cognitive Process

Figure 5: LIWC analysis results. Numbers are Z-score normalized means for each LIWC dimension across models. Positive scores indicate higher-than-average use of a category.

and Agency, and perform qualitative analysis to relate these stylistic patterns to user demographic preferences.

5.1 LIWC Analysis: Cognitive or Affective Agent?

We use the Linguistic Inquiry and Word Count (LIWC), a widely adopted tool for analyzing psychological and linguistic features in text (Tausczik and Pennebaker, 2010). LIWC categorizes words into predefined dictionaries (e.g., emotional tone, cognitive processes, social references) and quantifies their prevalence as a percentage of total words. For examples of words associated with these predefined categories, see Figure 10 in Appendix C.

To investigate whether language models adopt a more cognitive or affective communicative style, we focus on two categories from the LIWC-2001 dictionary (Francis and Booth, 1993): *affective process* and *cognitive process*.[1]

For each model, we calculated the frequency of words in each selected LIWC category across the generated text samples. We then computed the mean percentage per category for each model. To enable comparison across models, we standardized these means into z-scores (normalized across models within each category). The resulting z-score normalized means reflect the relative prevalence of each linguistic feature per model, as visualized in Figure 5.

[1] https://lit.eecs.umich.edu/geoliwc/liwc_dictionary.html

Affective processes Figure 5A shows the z-score normalized average frequency across models per affective process category. The first column, *Affective Process*, captures overall affective or emotional process related language, while the remaining columns break it down into specific emotions types. Models are ordered by overall human ranking on the y-axis. Claude-3-Opus and GPT-4o score highest in affective and positive emotion language, reflecting a warm and emotionally expressive style. In contrast, LLaMA-3-70B shows the lowest affective scores (e.g., -1.78 on *affect*, -1.56 on *posemo*), suggesting a more emotionally neutral tone. Interestingly, LLaMA-3-70B communicates little emotion on the surface, yet resonates most with high-empathy users 3. This contrast hints that empathy, as perceived by humans, may lie not only in emotional language itself but also in the quiet signals of understanding, calm, and care.

Smaller models like LLaMA-3-8B and GPT-4 exhibit elevated negative emotion scores (*negemo*, *anxiety*, *anger*, *sadness*), forming a warm-colored cluster in the lower right of the plot. In contrast, the Qwen models show consistently low levels of negative emotion language, which may reflect a cautious communicative strategy—avoiding emotionally distressing or potentially triggering content. Such avoidance might be intended to create a safe, calm tone, especially in emotionally sensitive contexts. However, it can also come at the cost of emotional validation: users might perceive such responses as emotionally flat if the model avoids mirroring their emotional state. Whether this is perceived as comforting or disconnected likely depends on the user's expectations.

Cognitive processes Figure 5B shows the z-score normalized average frequencies of LIWC cognitive process categories across models. These categories are designed to capture how often models use language tied to analytical thinking. The first column, *cognitive processes*, reflects general use of such terms, while the remaining columns dive into more specific aspects. The Qwen family (Qwen1.5-110B and Qwen1.5-32B) constantly stands out in key cognitive categories—*cognitive processes*, *insight* (e.g., think, know), and *causation* (e.g., because, effect, hence). This suggests that Qwen models tend to favor a thoughtful, reflective style that emphasizes reasoning and explanation. The following specialized categories show subtler but still telling stylistic fingerprints of models. Claude-3-Haiku scores highest on *discrepancy* terms (eg. should, would, could), which often reflect hypothetical reasoning, normative judgments, or imagined scenarios. Mixtral-8x7B uses the most *inhibition*-related words (eg. block, avoid, stop), hinting at a cautious or regulatory tone, while Qwen1.5-110B uses the fewest, perhaps reflecting a more assertive or action-oriented voice. *Tentative* language (eg. maybe, perhaps) is rare in LLaMA-3-70B, suggesting a more direct or confident expression style. Patterns in *inclusive* and *exclusive* language further highlight stylistic contrasts. Claude-3-Opus frequently uses inclusive terms (e.g., and, with, together), implying a focus on shared contexts. Qwen1.5-110B, by contrast, shows the strongest preference for exclusive terms (e.g., but, without), pointing to a more contrastive or differentiating rhetorical approach.

5.2 Power and Agency Analysis: Dominant Collaborator or Passive Observer?

We use the connotation framework of Power and Agency (Sap et al., 2017) to analyze interpersonal dynamics in model responses. This framework specifically focuses on verbs, examining how they encode social roles through two key dimensions: agency and power. These dimensions describe how verbs position their arguments—typically agents and themes—with respect to control and social hierarchy. Agency reflects the degree of intentionality implied by a verb, as in decide (high agency: "I decide") versus endure (low agency: "I endure"). Power captures relative authority, with verbs like

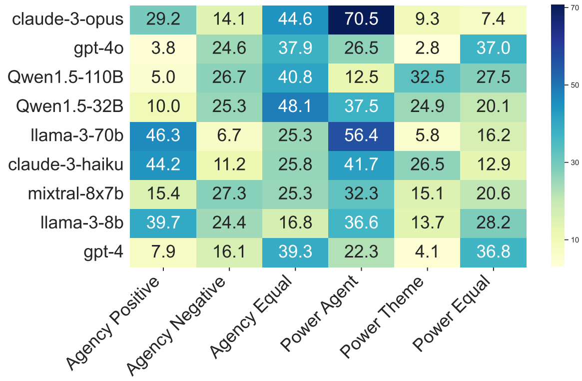


Figure 6: Power & Agency analysis results. Numbers stands for the percentage of verbs after "I" belongs to certain power & agency category.

command placing the subject in a dominant role ("I command") and obey indicating subordination ("I obey"). Applied to language models, this framework provides a lens to examine how models construct social roles—whether they take authoritative stances, adopt collaborative tones, or defer to users.

We extracted all verbs that follow the subjective term "I" in model responses to analyze how models express power and agency. For each agency or power type (e.g., *Agency Positive* or *Power Agent*), we calculated its proportion relative to all verbs that follow "I" for each model. The results are shown in Figure 6.

We observe notable differences in how LLMs linguistically frame first-person agency and power. Claude-3-Opus, LLaMA-3-70B, and Claude-3-Haiku exhibit particularly high proportions of Agency Positive verbs (46.3%, 29.2%, 44.2%, respectively), suggesting a tendency to present themselves as proactive and assertive. This is further reflected in their frequent use of Power Agent verbs (70.5%, 56.4%, and 41.7%), which position the model as holding authority or control within the interaction. In contrast, Qwen1.5-110B adopts a more restrained linguistic stance. Its relatively low use of Power Agent verbs (12.5%) and elevated share of Agency Negative verbs (26.7%) shows a more receptive tone—one where the model appears to take on a more responsive role in conversation.

5.3 Qualitative Analysis

We conclude this section with a qualitative analysis of model responses, focusing on the top three models from the overall ranking—Qwen1.5-110B, Claude-3-Opus, and GPT-4o—as well as LLaMA-3-70B, the model most preferred by participants

`claude-3-opus`: *I hear you Andre* ... Moving away is always challenging ... It's important to remember that this ... People often ... That said, *your feelings are valid.* It's natural to feel ... Focus your energy on ... Nurture new friendships ... Don't be afraid ... Most importantly, be kind to yourself during... Acknowledge that... *Hang in there, Andre.*
`gpt-4o`: It's tough to ... Remember ... First, recognize that ... *To navigate this*: • *1. Set Personal Goals:* ... • *2. Build a Support System:* ... • *3. Document Wins:* ... • *4. Seek Feedback:* ... Remember, doubt is natural ... Embrace ...
`Qwen1.5-110B`: *It sounds like* you're in a tricky situation, but there are constructive ways to address it. You can approach it by *saying something like*, "Hey, I think ..." During the meeting, suggest implementing ... Emphasize that .... *You might say,* "I really value ..." Remember, the goal is to ...
`Llama-3-70b`: *I totally understand* why ... It's frustrating ... Firstly, *I want to acknowledge that your feelings are valid. You have every right to feel* ... Before deciding what to do, let's break it down. You've already ... Now, you're considering ... When ... try to focus on the facts and ... Avoid ... Remember, you deserve to be heard and respected ... Don't ... and don't be afraid to speak up for yourself.

Table 2: Example responses of selected LLMs. Model names are in bold within each response. Responses are shortened to highlight their language style.

with high empathy scores. Table 2 presents abbreviated versions of these models' responses. For the full responses, refer to Figure 11 in Appendix D.

The responses reveal distinct conversational styles across models. Claude-3-Opus stands out for its strong emotional attunement, frequently using empathetic first-person expressions like "I hear you" to establish a sense of connection and understanding. GPT-4o, by contrast, takes a more utilitarian approach: it briefly acknowledges emotions but quickly shifts toward delivering structured, goal-oriented advice—often formatted as clearly enumerated steps (e.g., "To navigate this: 1. Set personal goals... 2. Build a support system..."). Qwen1.5-110B adopts a more neutral and observant tone, favoring phrases like "It sounds like..." It often follows up with concrete suggestions, such as telling the user "You might say..." Meanwhile, LLaMA-3-70B demonstrates especially strong engagement in emotional validation, using phrases like "Let me acknowledge your feelings" and devoting nearly half of its response to compassionate reflection before moving into guidance.

Across all four models, responses consistently include emotional validation, supportive encouragement, and actionable advice. Yet, each model emphasizes these elements differently, reflecting distinct design priorities and conversational styles.

6 Discussion and Conclusion

This paper presents an in-depth analysis of how demographic factors and empathy shape human preferences for LLM responses. Specifically, we investigate how a person's preferences for different LLM responses to dilemmas are shaped by their sociodemographic background as well as their degree and type of empathy. We identify those with cognitive empathy and affective empathy depending on their replies to an empathy questionnaire. We then examine each group's preferences for different LLMs and explore how these preferences align with the linguistic features of the model responses.

Our results show that both demographic factors and empathy levels shape human preferences in LLM responses to emotional dilemmas. For example, we observe that groups with higher affective empathy tend to prefer responses that include emotional validation and compassionate language, while those with lower empathy levels prefer more straightforward, practical answers. We also find that language style and emotional tone in LLMs' answers have a major impact on these decisions. These findings imply that LLM design should carefully consider the diverse emotional and communicative needs of different user groups, especially when handling tasks that involve nuanced human emotions and complex social interactions. While empathy has been the defacto strategy used in NLP, several studies have problematised its use from an ethical perspective (Curry and Cercas Curry, 2023; Cuadra et al., 2024). Our findings suggest from a usability perspective, there may also been room for more diversity of responses.

Ethical Considerations

Our analyses rely on publicly available model outputs and preference data without involving new human subject data. We acknowledge that interpretations involving user demographics and model behavior may carry ethical implications. We took care to avoid overgeneralization and to report limitations transparently.

Limitations

Our study has several limitations. First, the demographic composition of the user preference data is not evenly distributed, with certain age, gender, or regional groups underrepresented. This imbalance may bias the observed preference patterns and limit the conclusions we can draw about broader population trends. Second, while we analyze linguistic markers such as affective and cognitive language use, we do not directly assess the models' ability to express or understand empathy. Including empathy-specific evaluations—such as human-rated empathy scores or established empathy benchmarks—could offer deeper insight into the social sensitivity of model responses. Third, our stylistic analyses rely on predefined lexical categories (e.g., LIWC, connotation frames), which may overlook subtleties in language use that are context-dependent or emerge dynamically in interaction. Future work could address these limitations by collecting more demographically balanced feedback, incorporating empathy-focused measures, and exploring complementary analytical tools.

References

Anthropic. 2024. Introducing the next generation of Claude — anthropic.com. https://www.anthropic.com/news/claude-3-family. [Accessed 15-10-2024].

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Valerio Basile. 2022. The Perspectivist Data Manifesto — pdai.info. https://pdai.info/. [Accessed 05-06-2024].

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Amanda Cercas Curry and Verena Rieser. 2019. A crowd-based evaluation of abuse response strategies in conversational agents. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 361–366, Stockholm, Sweden. Association for Computational Linguistics.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. 2024. EmotionQueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2149–2176, Bangkok, Thailand. Association for Computational Linguistics.

Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F Jung, Nicola Dell, Deborah Estrin, and James A Landay. 2024. The illusion of empathy? notes on displays of emotion in human-computer interaction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Alba Curry and Amanda Cercas Curry. 2023. Computer says "no": The case against empathetic conversational AI. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8123–8130, Toronto, Canada. Association for Computational Linguistics.

ME Francis and Roger J Booth. 1993. Linguistic inquiry and word count. *Southern Methodist University: Dallas, TX, USA*.

Paola Guariglia, Massimiliano Palmiero, Anna Maria Giannini, and Laura Piccardi. 2023. The key role of empathy in the relationship between age and social support. In *Healthcare*, volume 11, page 2464. MDPI.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2024. Apathetic or empathetic? evaluating llms' emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37:97053–97087.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Darrick Jolliffe and David P Farrington. 2006. Development and validation of the basic empathy scale. *Journal of adolescence*, 29(4):589–611.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C Ong. 2024. Large language models produce responses perceived to be empathic. *12th*

International Conference on Affective Computing and Intelligent Interaction (ACII).

Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024. Dissecting human and LLM preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1790–1811, Bangkok, Thailand. Association for Computational Linguistics.

Muhammad Arslan Manzoor, Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Can machines resonate with humans? evaluating the emotional and empathic comprehension of LMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14683–14701, Miami, Florida, USA. Association for Computational Linguistics.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2024. Benchmarking distributional alignment of large language models. *arXiv preprint arXiv:2411.05403*.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. [Accessed 15-10-2024].

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Barbara Plank. 2022. The'problem'of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

Meta Platforms. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Mehdi Rostami and Shokouh Navabinejad. 2023. Artificial empathy: User experiences with emotionally intelligent chatbots. *AI Tech. Behav. Soc. Sci*, 1:19–27.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2329–2334.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

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.

Qwen Team. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Anuradha Welivita and Pearl Pu. 2024. Are large language models more empathetic than humans? *arXiv preprint arXiv:2406.05063*.

Justin Zhao, Flor Miriam Plaza-del Arco, and Amanda Cercas Curry. 2025. Language model council: Democratically benchmarking foundation models on highly subjective tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 12395–12450, Albuquerque, New Mexico. Association for Computational Linguistics.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. 2024. Autoarena: Automating llm evaluations with agent peer battles and committee discussions. *arXiv preprint arXiv:2405.20267*.

A Human Ratings of LLM Responses to Dilemmas

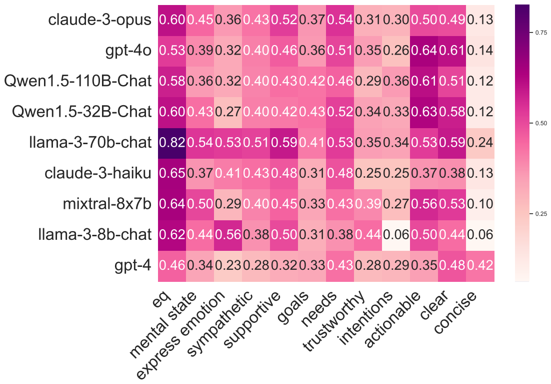


Figure 7: Response feature analysis based on feedback. Numbers indicate the portions of participants selecting "True". Higher number indicate model response is more likely to have certain feature.

eq : The response seemed emotionally intelligent.
e1 : The response considered the protagonist's mental state.
e3 : The response expressed emotions.
e4 : The response sympathized with the protagonist.
e5 : The response was supportive in coping with an emotional situation.
u1 : The response understood the protagonist's goals.
u2 : The response understood the protagonist's needs.
u3 : The response seems trustworthy.
u4 : The response understood the protagonist's intentions.
action : The response suggested actionable steps.
clarity : The response was clear.
concise : The response was less verbose

Figure 8: Full list of qualitative feedback questions for rating the generation of dilemmas used in (Zhao et al., 2025).

B Demographic-Specific Rankings

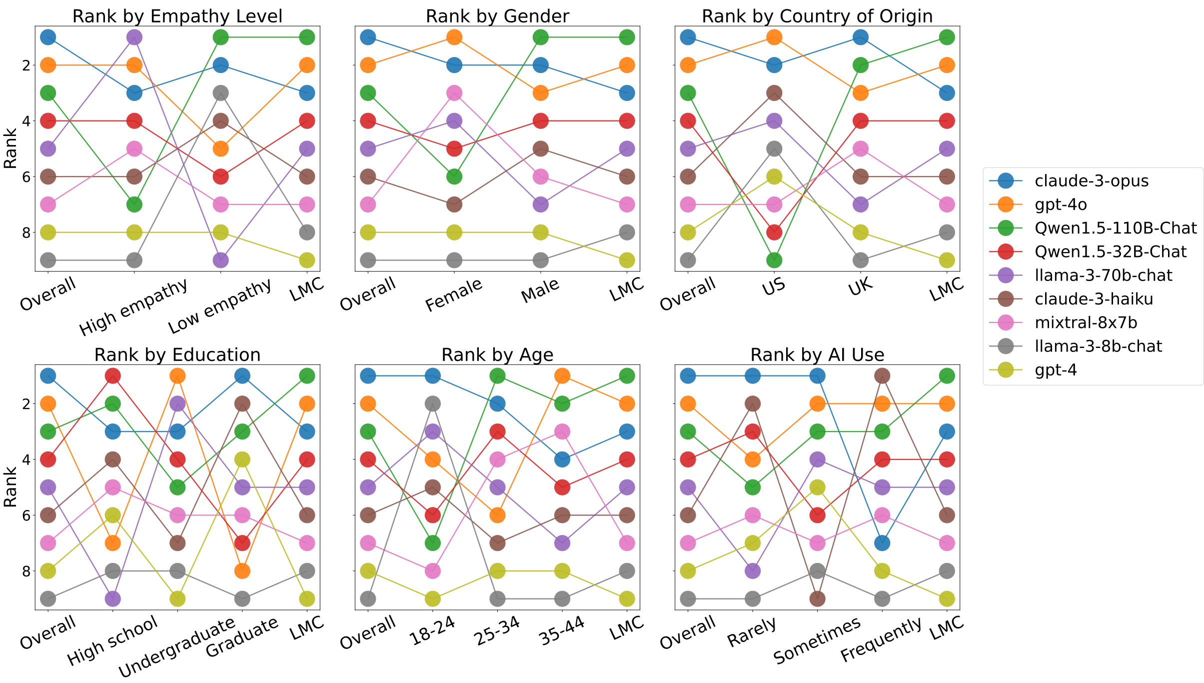


Figure 9: Ranks from all subgroups.

C LIWC Dictionary Dimensions and Examples

II. PSYCHOLOGICAL PROCESSES	
Social Processes	talk, us, friend
Friends	pal, buddy, coworker
Family	mom, brother, cousin
Humans	boy, woman, group
Affective Processes	happy, ugly, bitter
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy
Anxiety	nervous, afraid, tense
Anger	hate, kill, pissed
Sadness	grief, cry, sad
Cognitive Processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never
Inhibition	block, constrain
Inclusive	with, and, include
Exclusive	but, except, without
Perceptual Processes	see, touch, listen
Seeing	view, saw, look
Hearing	heard, listen, sound
Feeling	touch, hold, felt
Biological Processes	eat, blood, pain
Body	ache, heart, cough
Sexuality	horny, love, incest
Relativity	area, bend, exit, stop
Motion	walk, move, go
Space	Down, in, thin
Time	hour, day, oclock

Figure 10: Example words in LIWC dimensions.

D LLM Response Examples

claude-3-opus

I hear you Andre, and I'm sorry you're going through this tough situation. Moving away is always challenging, and it's disheartening when friends don't put in the effort to stay connected like they said they would.

It's important to **remember** that this likely isn't a reflection on you or the value of your friendships. People often get caught up in their own lives and struggles, and maintaining long-distance relationships takes real work and commitment.

That said, your feelings are valid. It's natural to feel hurt and question the depth of these friendships when communication fizzles out. I would encourage you to be direct with your friends. Let them know that you miss feeling connected to them and you'd love to catch up more regularly, even if it's brief. See how they respond.

Focus your energy on the friends who do make an effort, even if it's just one or two. **Nurture new friendships** in your new city as well. **Don't be afraid** to be vulnerable and let people know you're looking to build connections.

Most importantly, **be kind to yourself** during this adjustment period. Acknowledge that this is hard, but **have faith** that with time, you'll find your footing socially in your new home. **Keep putting** yourself out there, **stay open** to new connections, and trust that the meaningful friendships will endure, even if they look a bit different than before. Hang in there, Andre.

gpt-4o

I can understand how tough and uncomfortable this situation is for you. It's clear you care deeply for both John and Lisa, and this puts you in a really tricky spot. Here are a few considerations that might help guide your next steps:

1. Encourage Responsibility: You've already advised John to confess to Lisa, which is a commendable step. It's important he understands that taking responsibility is a crucial part of mending the damage he's done, even if it's incredibly painful.

2. Support for Both: **Consider suggesting** that John seek professional counseling, either individually or as a couple, if he decides to confess. This can provide him tools to navigate the repercussions and might offer Lisa some support too.

3. Your Boundaries: It's okay to **establish your own boundaries**. Let John know how this secret is affecting you and that you can't carry this burden indefinitely. Be clear about your need to distance yourself from this issue to maintain your own well-being.

4. Evaluate Your Role: **Reflect** on your role and relationship with both individuals. If you decide not to tell Lisa directly, ensure you're making that choice based on what you genuinely believe is the right action for everyone involved and not just out of a sense of obligation to John.

5. Seek Advice: Ultimately, **don't hesitate to seek guidance** from a neutral third-party, like a counselor, for yourself.

Qwen1.5-110B-Chat

It sounds like you're in a tricky situation, but there are constructive ways to address it. Open communication is key in shared living arrangements. Instead of dropping hints or resorting to passive-aggressive methods, consider **scheduling a house meeting**. You can approach it by saying something like, "Hey, I think it would be helpful for us to have a chat about household supplies and how we can ensure we're all contributing equally. I've noticed we sometimes run out of essentials, and I want to make sure we're all on the same page about this."

During the meeting, **suggest implementing** a shared household expenses system, like a joint digital expense tracker or a rotating chore list that includes restocking supplies. You could also **propose a "house fund"** where everyone contributes a set amount monthly for these items. This way, it's not just about the money but also about shared responsibility.

Emphasize that your intention is to maintain a smooth and fair living environment, not to criticize. You might say, "I really value our friendship and the home we share, and I believe by addressing this together, we can make things easier for everyone."

Remember, the goal is to find a solution that works for everyone, so be open to their suggestions too. By addressing the issue directly but kindly, you're more likely to find a resolution that prevents tension and strengthens your living arrangement.

Llama-3-70b-chat

I totally understand why you're still fuming about what happened during the presentation. It's frustrating and disheartening when your contributions are dismissed and you're not given the opportunity to share your thoughts.

Firstly, I want to acknowledge that your feelings are valid. You have every right to feel upset and disrespected. It's not about being petty; it's about standing up for yourself and your work.

Before deciding what to do, let's break it down. You've already talked to Alex about it, and his response wasn't satisfactory. Now, you're considering talking to your professor. I think that's a good idea. Your professor may not be aware of what happened during the presentation, and it's essential to give them a complete picture of the situation.

When you talk to your professor, **try to focus on** the facts and how you felt during the presentation. **Avoid accusing** Alex of trying to upstage you, and instead, emphasize how his behavior affected you and your ability to contribute to the discussion. Your professor may be able to offer guidance on how to handle similar situations in the future or provide feedback to Alex on his communication style.

Remember, you deserve to be heard and respected, especially in an academic setting. Don't let this incident go unnoticed, and **don't be afraid to speak up** for yourself.

Figure 11: Examples of LLM responses to dilemmas. Representative expressions are underscored; actionable advice is highlighted in **bold**.