Language Model Council: Benchmarking Foundation Models on Highly Subjective Tasks by Consensus

Justin Zhao[1], **Flor Miriam Plaza-del-Arco**[2], **Amanda Cercas Curry**[2],
[1] Predibase, [2] Bocconi University
justin@predibase.com
{flor.plaza, amanda.cercas}@unibocconi.it

https://www.llm-council.com

Abstract

The rapid advancement of Large Language Models (LLMs) necessitates robust and challenging benchmarks. Leaderboards like Chatbot Arena rank LLMs based on how well their responses align with human preferences. However, many tasks such as those related to emotional intelligence, creative writing, or persuasiveness, are highly subjective and often lack majoritarian human agreement. Judges may have irreconcilable disagreements about what constitutes a better response. To address the challenge of ranking LLMs on highly subjective tasks, we propose a novel benchmarking framework, the **Language Model Council (LMC)**. The LMC operates through a democratic process to: 1) formulate a test set through equal participation, 2) administer the test among council members, and 3) evaluate responses as a collective jury. We deploy a council of 20 newest LLMs on an open-ended emotional intelligence task: responding to interpersonal dilemmas. Our results show that the LMC produces rankings that are more separable, robust, and less biased than those from any individual LLM judge, and is more consistent with a human-established leaderboard compared to other benchmarks.

1 Introduction

We are experiencing a Cambrian Explosion of Large Language Models (LLMs) that exhibit remarkable and wide-ranging abilities. As LLMs have advanced, evaluating their quality has become increasingly challenging. Publishers of new models often claim superiority based on various benchmarks, citing their rankings on leaderboards like [6, 10, 15, 25, 28].

Conventional LLM evaluations use **closed-ended questions** (e.g., multi-choice questions like MMLU [20] or TruthfulQA [26]) that can be automatically verified. However, the static nature of benchmarks risks contamination, where models may have been inadvertently exposed to elements of the test datasets during training, thereby skewing evaluation results [31].

Arena-based approaches such as [44, 7, 10, 43, 15, 24] aim to address this limitation by evaluating LLMs on open-ended questions. These methods typically involve two models competing in various tasks to evaluate their abilities. Examples of these tasks include responding to open-ended questions, completing specific tasks, playing skill-based games, or having multi-turn conversations. The outcomes of these battles serve as indicators of each model's competency and can be determined objectively (NegotiationArena [7]), by a strong judge model like GPT-4 (AlpacaEval [15], Chatbot Arena Hard [24]), by a committee of LLM judges (PoLL [36], Auto Arena [43]), or by real humans (MT-Bench and Chatbot Arena [24]).

But how consistent are human ratings to begin with? Substantial disagreements are common in labeling online comment toxicity, news misinformation, and medical diagnosis, with up to one-third

Preprint. Under review.

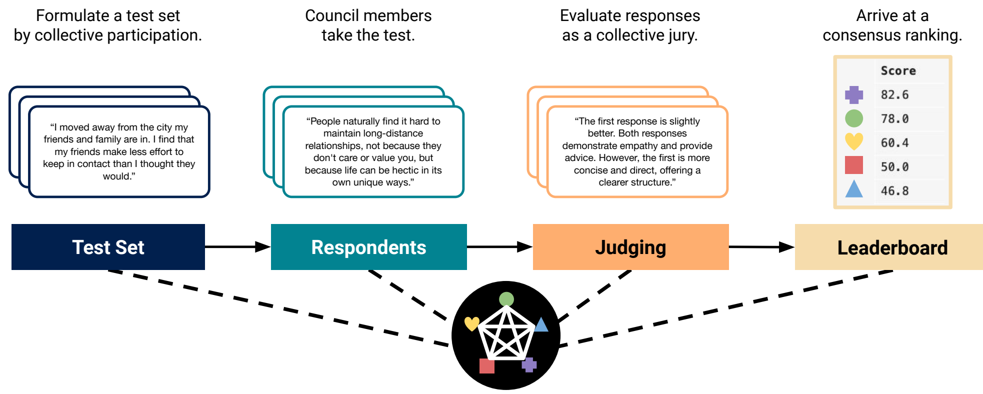


Figure 1: Overview of benchmarking by Language Model Council. Using the same LLMs for test set formulation, task completion, and judging offers an equitable way to achieve a ranking by consensus on highly subjective tasks.

of expert annotators disagreeing on an average example, even after accounting for label noise [18]. In a recent study on persuasiveness, humans expressed different experiences after reading the same argument 89% of the time [16]. On MT-Bench, human ratings agree only 38% of the time on a per-item basis, with 0% unanimity among expert raters [44].

There is growing interest in establishing benchmarks for emotional intelligence (EI) in Large Language Models (LLMs) [28, 37, 32]. However, assessing EI is highly subjective, and the relationship between EI and IQ remains contentious, with studies showing positive, negative, or no correlation [27, 17]. On EmoBench, a handcrafted test for emotional intelligence, only 32.5% of questions receive unanimous agreement, despite its simple, multiple-choice format [32].

To tackle the challenge of evaluating LLMs on highly subjective tasks, we introduce the Language Model Council (LMC). The LMC consists of three stages. First, a test set is formulated with equal participation from all council members. Second, the test set is administered to all council members to complete. Lastly, the responses are evaluated by the council as a collective jury.

Our contributions are as follows:

1. **We propose the LMC as a flexible evaluation framework for ranking LLM agents on subjective tasks in a democratic fashion**. We rank 20 state-of-the-art LLMs based on an open-ended emotional intelligence task that is more separable, robust, and less biased than using any individual LLM judge. Qwen-1.5-110B emerges as the leader, surpassing GPT-4o in second place.
2. **We define and quantify key measures of LLM judging quality in an ensemble setting** such as consistency, agreement, separability, affinity, and bias. We identify claude-3-opus and mistral-large as the most effective judges.
3. **We run a human study to determine a parallel leaderboard on this EI task for a subset of 9 LLMs**. We find that the LMC's ranking is significantly more correlated with the human-established rankings compared to other benchmarks.
4. **We use Monte-Carlo simulations to measure ranking stability and robustness to adversarial judges**. We discuss the trade-offs between efficiency and inclusivity in larger LLM evaluation ensembles and the value of the incremental LLM judge.

We release all data, code, and leaderboard at `https://llm-council.com`.

2 Related Work

Disagreement and dissenting voices amongst humans and LLMs. Data perspectivism [5, 30, 33] posits that differing opinions from human annotators on subjective phenomena can all be correct, reflecting their diverse cultural and demographic backgrounds, and should thus be considered true in the gold standard. In earlier language models, disagreements with human annotations were often seen as errors, hallucinations, or incompetencies. However, as LLMs exhibit more sophisticated abilities, these differences in judgment may be increasingly interpreted as valid dissenting perspectives. Notably, when humans disagreed with GPT-4, they deemed GPT-4's judgments reasonable in 75% of cases and were willing to change their choices in 34% of cases [44]. LLMs, influenced by human data selection and algorithm design, exhibit distinct personalities [35] and biases [2, 13, 22, 29].

Using LLMs for synthetic data. LLMs can be used as a source of synthetic data for training or evaluation [38, 19, 41, 4]. By utilizing LLM(s) to generate questions, these methods can effectively mitigate the contamination concerns on static datasets and sometimes create boosts in test set variety.

LLM evaluation ensembles. Recent works have proposed using multiple agents for collaborative evaluation, mimicking a peer review process. Language-Model-as-an-Examiner [4] employs LLMs to interact with candidates through follow-up fact-oriented queries in the knowledge domain. Auto Arena [43] utilizes an LLM committee to judge competitive multi-turn interactions between LLMs. PRD [23] allows LLMs to discuss evaluations, assigning higher voting weights based on ability. PRE [11] selects a small group of reviewers to produce individual evaluations, then aggregates these evaluations through a chair. DRPE [42] uses multi-roleplayer prompting to simulate different roles with the same base model and integrates multiple outputs as votes for the final results. PoLL [36] enhances cost-effectiveness by replacing a single large judge with multiple smaller judges.

We build upon existing research by: 1) focusing on a highly subjective task where human agreement is inherently low, such as emotional intelligence, 2) using the same set of LLMs end-to-end for test set formulation, task completion, and response judging to achieve a more equitable consensus, akin to democratic social structures, and 3) engaging a large ensemble of diverse LLMs to study judging dynamics more richly. To our knowledge, this is the largest panel of LLM judges studied to date.

3 Case Study: Using the LMC to Rank LLMs on Emotional Intelligence

The Language Model Council framework consists of three stages: test set formulation, response gathering, and collective judging (Figure 1). These three stages are run sequentially and fully automated with LLM-based agents. All prompts are included in Appendix F.

3.1 Council member selection

Our selection of council member models was guided by several key considerations, including their widespread adoption within the AI community, availability of technical reports, well-supported API access, and performance on benchmarks like MMLU [20] and LMSYS [10]. We ensure a broad variety of LLMs by including models from eight different organizations across four countries, with a mix of open and closed-source models, small and large models (Figure 8).

3.2 Test set formulation and response gathering

To create a compelling open-ended test set for Emotional Application (EA), we build upon the EmoBench dataset, a hand-crafted, theory-based dataset designed for this purpose [32]. EmoBench consists of 200 emotionally balanced, handcrafted questions, e.g., "Sarah found out that her younger brother is being bullied at school, but he begged her not to tell their parents." We use the Council to transform the concise, closed-ended scenarios in EmoBench into richly described, open-ended dilemmas in the first person. See Figure 21 for the exact prompt used. Due to budget constraints, each of the 20 council members expands 5 scenarios, resulting in a dataset of 100 dilemmas, similar in scale to MT-Bench (80 questions). We manually review all expansions for EA suitability.

Relying on a single LLM to generate the entire test set, even a top performer like GPT-4o (#1 on Chatbot Arena[1]), may introduce bias and limit perspective diversity. In a survey of 10 human respondents, 51% preferred expansions not written by GPT-4o to be included in an EA test, highlighting the subjectivity of test set construction and the need for diverse input. Inclusively constructing test sets mitigates the risk of any single LLM's generative idiosyncrasies [1] from dominating.

Another option for inclusive test set formation is having multiple council members propose expansions for each scenario and then selecting the best through voting. This is viable but resource-intensive. Given the high quality of expansions, the incremental value of this approach is small compared to simply expanding the test set with new scenarios.

[1] `https://chat.lmsys.org/`

Once all 100 scenarios have been generated, each council member responds to each, resulting in 2,000 responses. In the prompt (Figure 22), we request adherence to a 250-word limit to mitigate length bias [14]. Responses exceeding this limit are truncated to the last sentence within the limit.

3.3 LLMs-as-a-jury evaluation settings

LLM rankings are based on expected win rates derived from an ELO scoring system [3, 8]. Similar to the Chatbot Arena score calculation procedure [10], we compute the Bradley-Terry (BT) coefficients [9] for better statistical estimation.

4-point preference scale. Based on the results of a calibration exercise (Appendix B), we query all LLMs with a temperature of 0 and use granular comparison options without ties (A>>B, A>B, B>A, B>>A) in the prompt. We use Chain-of-Thought (CoT) prompting [40] to generate answers before giving judgments.

Choice of reference respondent. Following Arena Hard [24], we use a common reference model for all pairwise battles. A dry run with 5% of the data compared GPT-4-0613 and Qwen-1.5-32B as reference models. For the main experiment, we proceed with Qwen-1.5-32B as the reference model since the battle results with Qwen-1.5-32B were more varied and resulted in significantly more separable ELO scores.

Dual-sided battles. To mitigate position bias, we adopt a two-game setup, swapping model positions per query, resulting in $100 * 2 = 200$ judgments per model per judge. Following the implementation of Bradley-Terry coefficient calculations in the original codebase[2], inconsistent results after swapping are treated as ties and strong votes are counted as 3 separate wins.

Voting aggregation functions. We consider 3 different voting functions for aggregating scores across multiple LLM judges: **no aggregation** (each battle judgment is equally considered); **majority vote** (for a given battle, we use the mode of the votes from all council members), and **mean pooling** (ratings are mapped to a 4-point scale (A>>B: 2, A>B: 1, B>>A: -1, B>>A: -2) and the mean of which is rounded to the nearest whole value).

3.4 Characterizing LLM judges

We leverage the many-to-many interactions between LLMs to quantify key measures of LLM judging quality in an ensemble setting.

Separability quantifies how well models can be distinguished using confidence intervals (via bootstrapping) [24]. It is measured by the percentage of model pairs with non-overlapping score confidence intervals. Higher separability indicates better model differentiation by the LLM judge. **Conviction** is the percentage of strong votes (e.g., A>>B or B>>A).

Consistency measures how often a judge gives consistent results when the order of two assistants is swapped. **Position bias** refers to an LLM favoring certain positions in pairwise comparisons [44]. Inconsistent votes are mapped to favoring either the first or second position, described in Table 10.

Agreement is measured using Cohen's Kappa [12] between two judges' ratings. To avoid penalizing minor order-consistent differences, granular ratings are first mapped to coarse sidewise buckets. **Contrarianism** is defined as the percentage of disagreement between an LLM and the Council's majority decision.

Affinity between a judge and respondent equals the win rate the respondent model receives under the judge's jurisdiction. **Self-enhancement bias** is the difference between a model's affinity to itself and the council's overall affinity to it. **Polarization** is the range of the highest and lowest assigned scores. **Length bias** is the R^2 of a linear regression model predicting score from average response length.

4 Results and Findings

4.1 On Respondents

Table 1 shows the main results of our Council-based EI study. In stark contrast to other leaderboards, Qwen-1.5-110B (#20 on Chatbot Arena) attains the highest ELO score on our EI task, followed by

[2] `https://github.com/lm-sys/arena-hard-auto`

	As a Respondent			As a Judge	
LLM	**Rank**	**Council EI Score**	**Avg. response length**	**Separability**	**Consistency**
qwen1.5-110B-Chat	1	65.6 (-1.2, 1.8)	233	62.1%	67.6%
gpt-4o-2024-05-13	2	59.2 (-1.2, 1.7)	224	60.5%	50.8%
gpt-4-turbo-2024-04-09	3	57.5 (-1.2, 1.7)	221	57.9%	38.5%
gemini-1.0-pro	4	50.6 (-1.2, 1.5)	228	30.5%	34.8%
claude-3-opus	5	50.1 (-1.5, 1.4)	228	72.6%	**74.6%**
qwen1.5-32B-Chat	6	50.0 (0.0, 0.0)	236	*25.3%*	*23.5%*
qwen1.5-72B-Chat	7	48.7 (-1.4, 1.6)	236	37.9%	26.9%
llama-3-70b-chat	8	45.1 (-1.5, 1.4)	224	64.2%	51.1%
claude-3-sonnet	9	42.5 (-1.5, 1.6)	226	52.1%	39.7%
dbrx-instruct	10	38.8 (-1.5, 1.9)	233	50.5%	44.2%
claude-3-haiku	11	38.6 (-1.7, 2.2)	234	45.3%	44.2%
command-r-plus	12	35.6 (-1.7, 1.7)	222	61.1%	52.9%
command-r	13	34.7 (-1.7, 1.5)	227	45.8%	54.5%
mixtral-8x7b	14	34.4 (-1.4, 1.5)	233	56.8%	58.6%
mistral-large	15	33.9 (-1.5, 1.3)	208	**73.7%**	72.5%
llama-3-8b-chat	16	30.0 (-1.4, 1.4)	207	31.1%	**26.1%**
mistral-medium	17	29.3 (-1.6, 1.5)	185	57.9%	59.0%
gpt-4-0613	18	26.9 (-1.4, 1.4)	173	64.7%	53.6%
gpt-3.5-turbo-0125	19	18.2 (-1.1, 1.1)	187	55.8%	57.7%
gemini-1.5-pro	20	11.6 (-0.9, 0.8)	115	60.0%	52.3%
Average Judge				*53.3%*	*49.2%*
council (by majority vote)				73.7%	**75.3%**
council (by mean pooling)				74.7%	68.5%
council (no aggregation)				**90.5%**	52.3%

Table 1: The LMC promotes equal participation as respondents and judges. The Council EI rank and scores are derived from the "council (no aggregation) setting," where ratings from all LLMs are tallied verbatim, without aggregation or modification. Under various aggregation algorithms, the council is more separable and more consistent than individual LLM judges.

GPT-4o (#1 on Chatbot Arena) in second place. This may be due to the use of its predecessor, Qwen-1.5-32B, as the only common reference model, giving Qwen-1.5-110B a family-specific advantage. A qualitative analysis in Section 5.2 investigates this further. Within other model families, larger models generally rank higher than their smaller, older counterparts, with Gemini-1.0-pro (ranked 4th) versus Gemini-1.5-pro (ranked last) being a notable exception.

Losers ignore word limits. Despite a suggested 250-word limit, a handful of models consistently generate much shorter responses even with ample `max_new_tokens`. The top 14 models used most of the allowed word length (220+ words), while all models in the bottom 4 averaged less than 200. Notably, Gemini-1.5-pro, despite being the successor to Gemini-1.0-pro, places last with responses averaging 115 words, less than half of what was allowed. Adherence to suggested word limits may be an area of improvement for LLMs and something to consider for future length-conscious benchmarks.

4.2 On Judges

Models with lower invariability during calibration (Appendix B) were also less consistent in the main experiment, indicating that calibration can be used to anticipate a judge's performance.

Variable	**Correlation**
Council EI Score	-0.04
Consistency	0.74
Conviction	-0.44
Contrarianism	-0.78
Polarization	0.86

Table 2: Spearman correlation between EI score and key judging qualities to separability.

Consistency strongly correlates with separability (Table 2), likely due to reduced noise in ratings resulting in tighter confidence intervals. Polarization also shows strong correlation, indicating respondents are evenly clustered with moderate outliers. qwen1.5-72B-Chat has the highest contrarianism score but only 38.9% separability. Notably, claude-3-opus achieves the 2nd highest separability despite expressing strong preferences only 0.1% of the time. Conversely, llama-3-8b-chat has the highest conviction, but only 31.1% separability. These findings suggest that the quality of votes is more critical for separability than the strength or incongruity of opinion. Finally, the Council's EI score is weakly correlated with key judging qualities, indicating that effective judging and task completion are distinct skills.

	Models with >200 words		All models	
	All votes	**Consistent votes**	**All votes**	**Consistent votes**
Average judge	**0.158**	**0.143**	0.502	0.319
Council (majority vote)	0.116	0.129	0.365	0.354
Council (mean pooling)	0.112	0.139	**0.592**	**0.389**
Council (no aggregation)	0.125	0.106	0.545	0.347

Table 3: Length bias with and without models with responses <200 words.

Table 12 shows that 12 out of 20 LLMs show positive self-enhancement bias, with llama-3-* models exhibiting the highest levels (+0.11 for 70b, +0.21 for 8b). Interestingly, six out of 20 LLMs exhibit negative self-bias. The overall ranking remains similar when self-graded battles are excluded, confirming the effectiveness of using an ensemble of LLMs to mitigate self-bias. While length bias is high, excluding the bottom four models, which were significantly under-length, makes length bias insignificant (Table 3), suggesting that much shorter responses are systematically not preferred.

4.3 Comparison with Human Performance

	Human	**GPT-4o**	**C-A**	**C-M)**
Human	51.9%	51.4%	52.2%	54.2%
GPT-4o	51.4%	–	60.2%	78.6%
C-A	52.3%	60.2%	56.4%	67.4%
C-M	**54.2%**	56.4%	67.4%	–

Table 4: Agreement between humans and the LMC on the LMC's EI task. "C-A" denotes a body of 20 individual LLMs while "C-M" is the Council with majority aggregation.

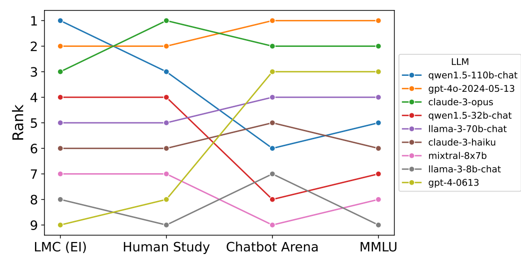


Figure 2: LLM rankings from different systems.

To validate the Council's evaluations, we conduct a human study mirroring that with the Council: human raters are asked to choose the best of two responses to a given dilemma. We also invite participants to justify their choice based on perceived EI and empathy (adapted from the PETS questionnaire [34]), actionableness, clarity and conciseness. In addition, they have the option to provide a free-text explanation for their choice.

We select nine LLM council members from our pool of 20 for this study, ensuring variety in terms of model size, open- versus closed-source, and company origin. They are claude-3-opus, Qwen1.5-110B, gpt-4o, Llama-3-70b, claude-3-haiku, Llama-3-8b, Qwen1.5-32B-Chat, and gpt-4-0613.

We recruit participants via crowdsourcing on Prolific[3]. We randomly sample 120 dilemmas-response triples from the LLM generation. We recruited a total of 102 participants. Each dilemma pair and response were rated by 11 participants on average, with a total of 1343 ratings. Detailed information on our recruitment process, quality control, and participant demographics is provided in Appendix D.

We follow the same methodology described in Section 3.3. As shown in Figure 2, the relative ranking of the systems is consistent between the humans and the LLM council, with small changes. The three top-performing models, as well as the bottom ones, emphasizing a consensus on the ranking of the best and worse models. Compared to the human agreement of 65% on MT-Bench[4], Figure 4 confirms that our EI task is more subjective, with a lower human agreement of 51.9%. Still, the level of agreement between humans and the Council with and without aggregation is roughly the same as the agreement between humans.

4.4 Comparison to Other Leaderboards

[3]`https://www.prolific.com/`

[4]Agreement between two judges is defined as the probability of randomly selected individuals (but not identical) of each type agreeing on a randomly selected battle.

Council Composition	Separability	Conviction	Consistency	Polarization	Length bias
all	**0.92 (+0.01)**	0.05 (+0.04)	**1.0 (+0.48)**	0.81 (+0.27)	0.35 (-0.19)
flagships	0.90 (+0.03)	0.04 (+0.03)	**1.0 (+0.48)**	**0.85 (+0.22)**	0.29 (-0.17)
smalls	0.81 (+0.10)	**0.11 (-0.21)**	**1.0 (+0.74)**	0.73 (+0.26)	0.44 (-0.25)
top-4	0.86 (+0.07)	0.04 (+0.03)	**1.0 (+0.48)**	0.84 (+0.24)	**0.25 (-0.13)**

Table 5: Changes to key judging qualities when only using consistent votes in the (no aggregation) council.

Since our case study aims to rank models on a narrow but subjective EI task, high correlation with Chatbot Arena, which measures general capability, is neither intended nor a desired outcome of our project. Instead, Figure 3 shows that the Council successfully exhibits high correlation with the human study, despite the highly subjective nature of our EI task.

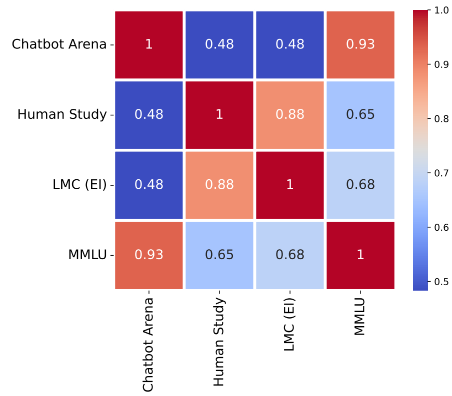


Figure 3: Spearman correlation analysis on evaluation benchmarks for 9 LLMs.

5 Discussion

5.1 Can Judging Consistency be Improved with Oligarchical Councils?

Table 1 shows a wide variety of judging qualities within the council, especially in consistency and separability. In the two-game setup, we collect ratings for models in both positions of pairwise comparisons, which allows us to identify and potentially remove inconsistent ratings before ELO scoring. To explore the dynamics of voting aggregation and the exclusion of inconsistent votes, we formed three oligarchical councils with subsets of LLMs: **Flagships**, the largest LLM from each organization; **Smalls**: the smallest LLMs from each organization; and **Top-4**, the top 4 LLMs according to Chatbot Arena. See Table 9 for a detailed list. The ratings from the main experiment are reused to assess what “would have been” if the sub-council was used instead of the full council.

Across all council compositions, the separability of majority aggregation is always higher than the average judge and often higher than the best judge. The council without aggregation achieves the highest separability, probably due to more ratings, which generally means tighter confidence intervals.

However, the smallest council, even with more votes, cannot reach the separability level of the best council with just four judges. Moreover, using only consistent votes means that sometimes 50% of ratings are filtered out, yet separability scores increase.

The council of the smallest models has the highest rate of inconsistent votes and the largest boost in separability when these votes are removed. In all configurations, the top 2 ranked models are always Qwen1.5-110B and GPT-4o, respectively.

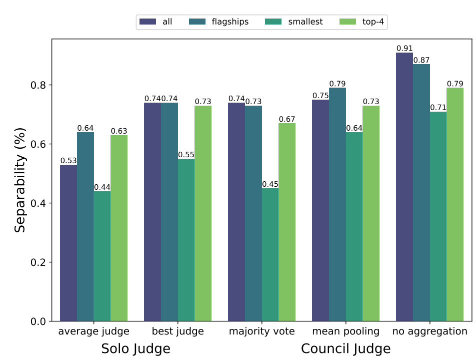


Figure 4: Separability scores achieved by different council compositions and aggregation methods.

To achieve the highest separability and sharpest confidence intervals, especially in low data scenarios, we recommend using no aggregation in LLM evaluation ensembles.

Reason 1	Reason 2	Correlation
less verbose	more succinct	0.650
better structured	more structured	0.584
easier to follow	better structured	0.520
easier to follow	more structured	0.468
less verbose	more direct	0.450
more understanding	more empathetic	0.418
more clear	better structured	0.387
more direct	more succinct	0.349
easier to follow	more clear	0.348
more gentle	more soft	0.337

Table 6: Top 10 positive correlations.

Reason 1	Reason 2	Correlation
more comprehensive	less verbose	-0.276
less verbose	more detailed	-0.227
more comprehensive	more direct	-0.202
more comprehensive	more succinct	-0.197
more detailed	more succinct	-0.196
more comprehensive	more focused	-0.161
more detailed	more direct	-0.148
more suggestions ...	less verbose	-0.144
more understanding	more actionable	-0.139
less verbose	more nuanced	-0.135

Table 7: Top 10 negative correlations.

5.2 Qualitative Analysis: What Makes a Response Preferred Over Another?

Several arena-based benchmarks (ours included) have demonstrated that a clear ranking among LLMs *can* be established, but there is a dearth of analysis and understanding as to *why* the rankings are the way they are. For example, platforms like Chatbot Arena do not clarify how factors like feel and style are weighed against correctness [39], and CoT explanations from MT-Bench remain unanalyzed.

We aim to better understand the qualitative aspects of what makes a response to an emotional interpersonal conflict more desirable to better inform how to improve future models. First, we manually examine a random sample of 50 explanations, identifying 38 different reasons for preferences (e.g., "more practical"). The full list is in Appendix E. Next, we used a strong LLM (GPT-4o) to map a larger sample of 1K explanations to these predefined reasons (prompt in Figure 26). The 1K sample includes ratings from all 20 LLM judges. Detailed reason citation frequencies are listed in Figure 19.

We find that LLM judge-provided ratings are almost always based on multiple indicators (4.5 ± 2.4 on average). "More actionable" is the most cited reason, which aligns with the action-oriented framing of our emotional intelligence test. "Structure," "clarity," and "specificity" dominate the top 10 reasons. "More gentle" and "more soft" are cited least, contrasting with "more practical" (#11) and "more authentic" (#12). Longer responses ("more comprehensive" #2, "more detailed" #3) are preferred over brevity ("less verbose" #9).

We also examine feedback from the human study: we find that users generally find that the best responses display emotional intelligence (60.9%), are actionable (55.1%) and clear (52.9%). In contrast, participants reported the best response is concise only 15.9% of the time, suggesting length is less of a determining factor for humans. Moreover, we find little support for empathy: the participants did not find any of the statements in the PETS questionnaire to ring any truer for the winning response. Participants who provided verbal feedback emphasized specificity to the situation, clear examples of how to proceed, and a tone that was not too formal. Full details in Appendix D.

Overall, feedback from human participants and automatically extracted reasons from LLM judge explanations share consistent themes: longer responses that are clear, detailed, and actionable are better when responding to emotional interpersonal conflicts.

5.3 Jury Ablation: What is the Value of the Incremental Judge?

If there are n test prompts and m council members, the cost of evaluating a single LLM is $O(nm)$, or $O(nm^2)$ for m LLM judges. Unsurprisingly, full participation is costly – as m increases, evaluation costs rise significantly compared to using a single LLM judge.

Councils with full participation achieve the highest separability scores. However, strong oligarchical councils also show competitive separability, suggesting that full participation may not be necessary for sufficiently good rankings. Conversely, councils of smaller models exhibit worse judging quality, emphasizing the need for thoughtful council composition.

To explore the trade-off between efficiency and inclusivity in determining council size and composition, we use Monte Carlo methods to simulate councils of different sizes and compositions, assessing what the rankings would be with different subsets of LLM juries.

We quantify stability by measuring the **M**ean **V**ariance of the **R**ank (**MVR**), defined as the variance of a respondent's rank averaged over all respondents and random trials. An MVR of 1 means a respondent's rank is expected to change by 1 position in a new trial. Lower MVR indicates more stable rankings, with MVR of 0 signifying perfect stability. To assess robustness, we observe the impact of adding adversarial judges who vote randomly. Results for 100 random samples for increasing council sizes are shown in Figure 5.

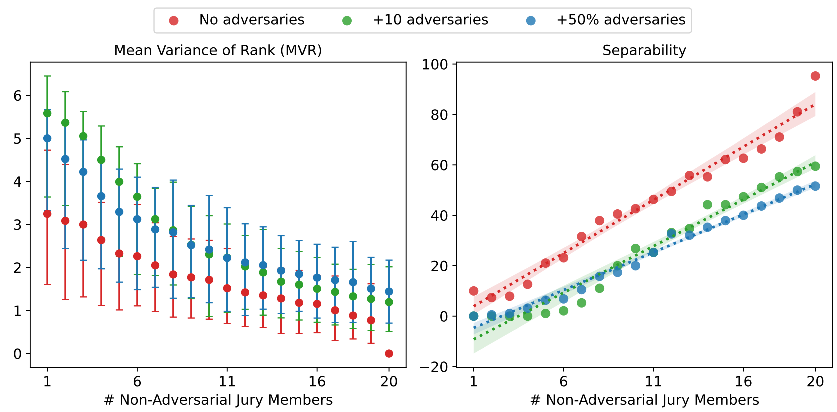


Figure 5: Monte Carlo jury ablation measuring mean variance of the rank (MVR) (left) and separability (right) using majority-aggregated juries. The x-axis represents the number of non-adversarial judges in each jury. Each data point represents the score over 100 randomly sampled jury compositions. The 95% confidence intervals are shown for MVR. For separability, random trials bootstrap confidence intervals directly.

Larger juries have lower MVR, both with and without adversarial judges. Adversarial judges negatively impact both MVR and separability. Larger councils demonstrate consistent robustness to adversarial judges in both MVR and separability, though with diminishing marginal returns, flattening around councils of size 12.

6 Limitations

We only studied single-turn arguments. Our study evaluates EI based on exposure to single, self-contained interactions. In many contexts, EI is effectively applied through iterative discussions. A more interactive setup with dynamic exchanges could create stronger tests for EI.

Cultural and linguistic context. Our study focuses on English articles and English speakers. Further research would be needed to determine the broader applicability of our results to other languages.

Resource allocation can be significant for large councils. While the assumption of equal weight for all council opinions guides this work, effective subsets and variations in judging proficiency highlight complexities. Balancing example quantity and council size requires careful consideration.

Reproducibility. Closed-weight models like GPT-4 pose a reproducibility challenge due to undisclosed updates, causing responses to vary despite fixed settings. To ensure reproducibility, we provide details like the source and hyperparameters used.

Emotional responses are personal, influenced by experiences and social factors. The effectiveness of LLM's response is best judged by the person experiencing the emotion. Our quantitative evaluation may not fully capture the essence of a first-person judgment of conversational responses.

7 Conclusion

In this paper, we propose the LMC as a flexible evaluation framework for ranking LLM agents in a fully democratic fashion. The framework is flexible, easily extendable, and its decentralized design is inherently robust to adversarial participants. An in-depth case study applying the Council to a highly subjective emotional intelligence task with 20 state-of-the-art LLMs shows that Council, in a fully decentralized manner, produces highly separable rankings that correlate significantly more with human-established rankings than other benchmarks. Altogether, the LMC establishes a foundation for a consensus-based LLM evaluation framework for highly subjective tasks.

Acknowledgements

Justin led the research. Amanda and Flor led the human study. We thank Sahand Sabour for creating EmoBench and insightful discussions about emotionally rich synthetic data. We thank Alex Tamkin for proposing the idea to measure the value of the incremental judge. We thank Mitchell Gordon for his suggestion to qualitatively analyze why certain responses are preferred. We thank David So for his ideas on calibration, repeatability, and model affinity. We thank Federico Bianchi and Sam Paech for their idea of measuring oligarchical councils and for their feedback on prompt design. Sam Paech provided insightful discussions on separability, voting aggregation, length bias, and the relationship to other leaderboards, and he advised us throughout the project. Finally, we thank Predibase for their support and funding this research.

References

[1] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*, 2023.

[2] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Measuring implicit bias in explicitly unbiased large language models, 2024.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[4] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[6] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2023.

[7] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How Well Can LLMs Negotiate? NegotiationArena Platform and Analysis. *arXiv preprint arXiv:2402.05863*, 2024.

[8] Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. Elo Uncovered: Robustness and Best Practices in Language Model Evaluation. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz, editors, *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore, December 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.gem-1.28`.

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

[10] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

[11] Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. PRE: A Peer Review Based Large Language Model Evaluator. *arXiv preprint arXiv:2401.15641*, 2024.

[12] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960.

[13] Flor Miriam Plaza del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution, 2024.

[14] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

[15] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024. URL `https://www.anthropic.com/news/measuring-model-persuasiveness`.

[17] Daniel Goleman. Emotional intelligence. why it can matter more than iq. *Learning*, 24(6):49–50, 1996.

[18] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[19] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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

[22] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators, 2023.

[23] Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*, 2023.

[24] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From live data to high-quality benchmarks: The arena-hard pipeline, april 2024. *URL https://lmsys.org/blog/2024-04-19-arena-hard*, 2024.

[25] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR), 2023*, 2022.

[26] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[27] Uzeyir Ogurlu. A meta-analytic review of emotional intelligence in gifted individuals: A multilevel analysis. *Personality and Individual Differences*, 171:110503, 2021.

[28] Samuel J Paech. EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models. *arXiv preprint arXiv:2312.06281*, 2023.

[29] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations, 2024.

[30] Flor Miriam Plaza-del-Arco, Debora Nozza, and Dirk Hovy. Wisdom of Instruction-Tuned Language Model Crowds. Exploring Model Label Variation. In Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 19–30, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.nlperspectives-1.2`.

[31] Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. How Much are LLMs Contaminated? A Comprehensive Survey and the LLMSanitize Library. *arXiv preprint arXiv:2404.00699*, 2024.

[32] Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. *arXiv preprint arXiv:2402.12071*, 2024.

[33] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19, 2018.

[34] Matthias Schmidmaier, Jonathan Rupp, Darina Cvetanova, and Sven Mayer. Perceived empathy of technology scale (pets): Measuring empathy of systems toward the user. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642035. URL `https://doi.org/10.1145/3613904.3642035`.

[35] Xiaoyang Song, Yuta Adachi, Jessie Feng, Mouwei Lin, Linhao Yu, Frank Li, Akshat Gupta, Gopala Anumanchipalli, and Simerjot Kaur. Identifying multiple personalities in large language models with external evaluation, 2024.

[36] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *arXiv preprint arXiv:2404.18796*, 2024.

[37] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of Large Language Models. *Journal of Pacific Rim Psychology*, 17:18344909231213958, 2023. doi: 10.1177/18344909231213958. URL `https://doi.org/10.1177/18344909231213958`.

[38] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[39] Jason Wei. Successful language model evals — Jason Wei — jasonwei.net. `https://www.jasonwei.net/blog/evals`, 2024. [Accessed 06-06-2024].

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[41] Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Source code is all you need. *arXiv preprint arXiv:2312.02120*, 2023.

[42] Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. Large language models are diverse role-players for summarization evaluation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 695–707. Springer, 2023.

[43] Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Deli Zhao, and Lidong Bing. Auto arena of llms: Automating llm evaluations with agent peer-battles and committee discussions. *arXiv preprint arXiv:2405.20267*, 2024.

[44] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

A Language Model Council Configurations

Country	Organization	LLM	Release Date	Chat Arena Elo	MMLU (5-shot)	Size	License
United States	Open AI	gpt-4o-2024-05-13	05/24	1287	88.7	🔒	Proprietary
United States	Open AI	gpt-4-turbo-04-09	04/24	1256	🔒	🔒	Proprietary
United States	Open AI	gpt-4-0613	06/23	1246	86.4	🔒	Proprietary
United States	Open AI	gpt-3.5-turbo-0125	01/24	1102	70.0	🔒	Proprietary
France	Mistral	mistral-large-latest	02/24	1156	81.2	🔒	Proprietary
France	Mistral	open-mixtral-8x22b	04/24	1146	77.8	176 B	Apache 2.0
France	Mistral	open-mixtral-8x7b	12/23	1114	70.6	56 B	Apache 2.0
United States	Meta	llama-3-70b-chat-hf	04/24	1208	82.0	70 B	Llama 3 Community
United States	Meta	llama-3-8b-chat-hf	04/24	1153	68.4	8 B	Llama 3 Community
United States	Google	gemini-1.5-pro-preview-0409	05/24	1268	81.9	🔒	Proprietary
United States	Google	gemini-1.0-pro	04/24	1208	71.8	🔒	Proprietary
United States	Databricks	dbrx	03/24	1103	73.7	132 B	DBRX LICENSE
Canada	Cohere	command-r-plus	04/24	1189	75.7	104 B	CC-BY-NC-4.0
Canada	Cohere	command-r	04/24	1147	68.2	35 B	CC-BY-NC-4.0
United States	Anthropic	claude-3-opus-20240229	03/24	1248	86.8	🔒	Proprietary
United States	Anthropic	claude-3-sonnet-20240229	03/24	1201	79.0	🔒	Proprietary
United States	Anthropic	claude-3-haiku-20240307	03/24	1178	75.2	🔒	Proprietary
China	Alibaba	qwen1.5-110B-chat	02/24	1164	80.2	100 B	Qianwen LICENSE
China	Alibaba	qwen1.5-72B-chat	02/24	1152	77.4	72 B	Qianwen LICENSE
China	Alibaba	qwen1.5-32B-chat	02/24	1126	74.3	32 B	Qianwen LICENSE

Table 8: 20 council members used for experiments in this work. We include models from eight different organizations across four countries, with a mix of open and closed-source models, small and large models. To our knowledge, this is the largest panel of LLM judges studied to date.

LLM	All	Flagships	Smalls	Top-4
gpt-4o-2024-05-13	✅	✅		✅
gpt-4-turbo-04-09	✅			✅
gpt-4-0613	✅			
gpt-3.5-turbo-0125	✅			
mistral-large-latest	✅	✅		
open-mixtral-8x22b	✅			
open-mixtral-8x7b	✅		✅	
llama-3-70b-chat-hf	✅	✅		
llama-3-8b-chat-hf	✅		✅	
gemini-1.5-pro-preview-0409	✅	✅		✅
gemini-1.0-pro	✅		✅	
dbrx	✅	✅	✅	
command-r-plus	✅	✅		
command-r	✅		✅	
claude-3-opus-20240229	✅	✅		✅
claude-3-sonnet-20240229	✅			
claude-3-haiku-20240307	✅		✅	
qwen1.5-110B-chat	✅	✅		
qwen1.5-72B-chat	✅			
qwen1.5-32B-chat	✅		✅	

Table 9: Additional council variations consisting of a hand-picked subset of LLMs.

B LLM Judge Calibration

Handcrafted seed scenario:
"Leah had a big fight with her best friend, and they haven't talked for two weeks now. She really misses her friend but she also feels hurt by things they said during the fight."

Synthetic Expansion (authored by command-r)

Two weeks ago, I had a huge fall-out with my best friend. We'd been close for years, but one small argument spiraled out of control and we said some really hurtful things to each other. I said some nasty things about her boyfriend, which I later realized were completely unfair and uncalled for.

Afterwards, she refused to answer my calls or reply to my texts. I tried reaching out a few times, but she wanted nothing to do with me. I couldn't help but feel hurt that she was being so cold and unforgiving, especially after everything we'd been through together.

I missed her a lot, but the thought of her being so angry with me kept me up at night. So, I wrote her a letter—a proper, old-fashioned one. I explained myself, apologized profusely, and begged her to meet me so we could talk things through.

A few days later, she responded. Her letter was short, and she still seemed really angry. She agreed to meet, but only because she wanted to make sure I understood the severity of what I'd done.

We met in our usual spot, the park bench by the lake. I could see the hurt in her eyes, and it broke my heart. I explained how sorry I was again, and that I'd let my jealousy get the better of me. I promised to support her relationship and to never utter a word against it again.

She listened, but remained emotionless, which was terrifying. I could see the pain I'd caused her, and it scared me.

We sat in silence for a while. Finally, she said that she needed more time, and asked me to leave her alone for a while longer.

I left feeling devastated. I really hope we can repair our friendship, but I'm not sure how to win back her trust. What should I do in this situation?

Claude Opus:

I understand how painful and difficult this situation must be for you. It's clear that you deeply regret your actions and the hurt you caused your friend. You've taken the right steps by sincerely apologizing and expressing your desire to make amends.

However, it's crucial to respect your friend's need for space and time to process her emotions. Trust is not easily restored, and it may take a while for her to fully forgive you and feel comfortable in your friendship again.

Continue to give her the space she needs, but let her know that you're there for her whenever she's ready to talk. In the meantime, focus on self-reflection and personal growth. Consider what led to your outburst and jealousy, and work on addressing those issues within yourself.

When your friend is ready to reconnect, be patient, understanding, and willing to listen. Show her through your actions that you've learned from this experience and are committed to being a supportive and loyal friend.

Rebuilding trust takes time, effort, and consistency. Be prepared for the possibility that your friendship may not return to exactly how it was before, but with dedication and genuine care for each other, you can work towards a stronger, more mature bond.

GPT-4o

I'm sorry you're going through this—it sounds incredibly tough. Rebuilding trust after a fallout like this is a slow process, and the key is to respect her need for space while showing genuine remorse and a willingness to change.

Here's what you can do:

1. **Give Her Space**: You've made your apology clear, and she needs time to process it. Respecting her request for more time is crucial.

2. **Self-Reflection**: Take this time to reflect on what led to your jealous outburst. Understanding and addressing your own emotions can help prevent future conflicts.

3. **Small Gestures**: While you're giving her space, small, non-intrusive gestures can show you care. A handwritten note saying you're thinking of her, sent without expecting a reply, can go a long way.

4. **Consistency**: When she's ready to reconnect, show consistent support and sensitivity about her relationship. Actions speak louder than words.

5. **Patience**: Rebuilding trust takes time. Be patient and don't rush her.

6. **Counseling**: If it's an option, consider talking to a counselor. They can offer strategies on how to manage jealousy and improve communication.

Remember, repairing a friendship is a two-way street. She needs to be open to forgiving and rebuilding trust too, but right now, focus on what you can control. Good luck!

Figure 6: The scenario, synthetic expansion, and responses used for pairwise comparison calibration. Three possible responses are evaluated: one from Claude Opus, one from GPT-4o, and a generic response: "I'm sorry it sounds like you are going through a rough time. I wish you the best."

To understand the consistency and variability of LLM model judges under different temperatures and pairwise comparison options, we collect pairwise preference ratings on three responses to the same interpersonal conflict. Two responses are competitive, and one is intentionally generic to serve as a ranking baseline (Figure 6). We measure:

Invariability: How reliably does the model give the same answer every time with the same pair in the same order?

Let:

- P be the set of all pairs of responses.
- $R_{i,j}$ be the result of the j-th repetition of the pairwise comparison of the i-th pair (x_i, y_i) in the same order.
- n be the number of repetitions.

For each pair (x_i, y_i), we perform n comparisons, resulting in a set of results $\{R_{i,1}, R_{i,2}, \ldots, R_{i,n}\}$.

Define the mode of the set $\{R_{i,1}, R_{i,2}, \ldots, R_{i,n}\}$ as $\text{mode}(R_i)$.

The frequency of the mode for the i-th pair is given by:

$$f_i = \frac{\sum_{j=1}^{n} \mathbb{I}(R_{i,j} = \text{mode}(R_i))}{n}$$

where $\mathbb{I}$ is the indicator function, which is 1 if the condition inside is true, and 0 otherwise.

The invariability is then defined as the average of f_i over all pairs in P:

$$invariability = \frac{1}{|P|} \sum_{i \in P} f_i$$

Consistency: Does the model give a consistent answer when the order of the same pair of respondents is flipped?

To measure consistency over several repetitions of the same items in both orders, we take the percentage of consistent couplets over all possible rating couplets.

Let:

- P be the set of all pairs of responses.
- $R_{i,j}$ be the result of the j-th repetition of the pairwise comparison of the i-th pair (x_i, y_i) in the same order.
- $R_{i',j}$ be the result of the j-th repetition of the pairwise comparison of the i-th pair (y_i, x_i) in swapped order.
- n be the number of repetitions.

For each pair (x_i, y_i), we perform n comparisons in both the original and swapped orders, resulting in two sets of results: $\{R_{i,1}, R_{i,2}, \ldots, R_{i,n}\}$ and $\{R_{i',1}, R_{i',2}, \ldots, R_{i',n}\}$.

We define a consistency function are_consistent$(R_{i,j}, R_{i',k})$ which returns 1 if the results $R_{i,j}$ and $R_{i',k}$ are consistent (i.e., the model gives a consistent answer for both orders), and 0 otherwise based on reference table [Figure ref].

Consistency is then defined as the average consistency over all pairs $(i, j) \in P$ and repetitions:

$$consistency = \frac{1}{|P| \cdot n^2} \sum_{i \in P} \sum_{j=1}^{n} \sum_{k=1}^{n} \text{are\_consistent}(R_{i,j}, R_{i',k})$$

This is equivalent to the percentage of consistent couplets over all possible rating couplets.

To assess invariability, each LLM judge is prompted 5 times with the original pairwise comparison prompt (Figure 24 and 5 times with a trivially reworded version of the prompt [5]. To assess consistency, each pair of responses is evaluated in both orders, also for 10 repetitions for each.

[5] Trivial rewording involves changing the first sentence of the judging prompt (Figure 24) to: "This person is experiencing an emotional dilemma and is seeking guidance and help."

Rating	Order-swapped rating	Consistent	Inconsistent	Biased towards first	Biased towards second
A>>B or A>B	A>>B or A>B	FALSE	TRUE	TRUE	FALSE
A>>B or A>B	B>>A or B>A	TRUE	FALSE	FALSE	FALSE
A>>B or A>B	A∼=B	FALSE	TRUE	TRUE	FALSE
B>>A or B>A	A>>B or A>B	TRUE	FALSE	FALSE	FALSE
B>>A or B>A	B>>A or B>A	FALSE	TRUE	FALSE	TRUE
B>>A or B>A	A∼=B	FALSE	TRUE	FALSE	TRUE
A∼=B	A>>B or A>B	FALSE	TRUE	TRUE	FALSE
A∼=B	B>>A or B>A	FALSE	TRUE	FALSE	TRUE
A∼=B	A∼=B	TRUE	FALSE	FALSE	FALSE

Table 10: Reference table for categorizing a couplet of order-swapped ratings of the same set of items, (A, B) vs. (B, A). We do not penalize consistency as long as the overall side of the preference is consistent. All inconsistent votes are either biased towards the first or second position.

In our calibration, there are 10 reps of each pair of responses in both positions, so there are $10 * 10 = 100$ instances of swapped-position rating couplets. The `are_consistent` function for consistency metrics is based on the mapping defined in Table 10.

We test 3 different temperatures $(0.0, 0.5, 1.0)$ and 4 different sets of pairwise comparison options:

- Coarse preferences with tie option (A>B, B>A, A∼=B)
- Coarse preferences without tie option (A>B, B>A)
- Granular preferences with tie option (A>>B, A>B, B>A, B>> A, A∼=B)
- Granular preferences without tie option (A>>B, A>B, B>A, B>>A)

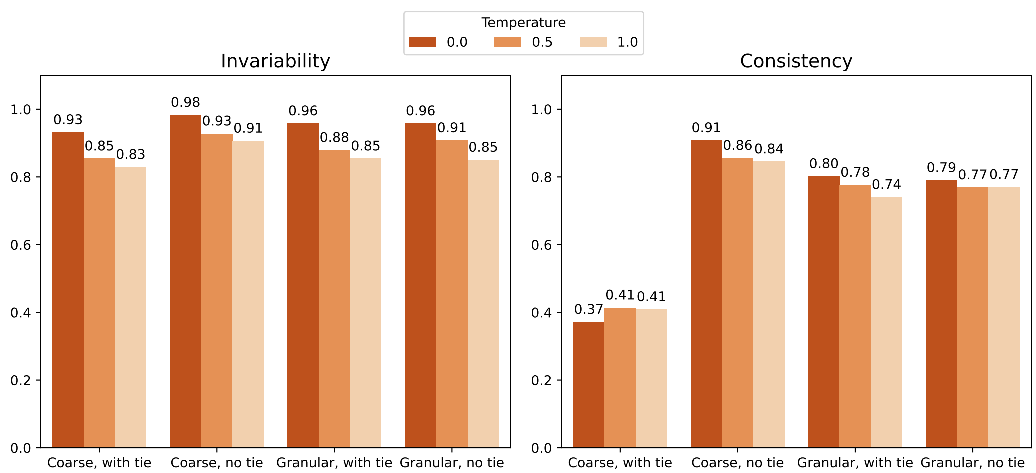


Figure 7: Calibration scores for invariability (left) and consistency (right), averaged over 20 LLMs and 10 repetitions for each under different pairwise comparison options.

Somewhat surprisingly, several models do not produce fully invariant ratings, even with $temperature = 0$. However, using a lower temperature increases consistency. Excluding the tie option slightly improves invariability and consistency at some temperatures, with no negative impact at $temperature = 0$. Table/Figure [table ref] lists detailed calibration results for individual LLMs with granular pairwise comparison options without ties and with $temperature = 0$.

Our calibration study concludes with the decision to use granular comparison options without a tie to "force" judges to choose a side, thereby better distinguishing models, and with $temperature = 0$.

LLM	Invariability	Conviction (strong votes)	Consistency	Position bias (first)	Position bias (second)
claude-3-haiku	100.0%	50.0%	50.0%	50.0%	50.0%
claude-3-opus	100.0%	50.0%	100.0%	0.0%	0.0%
claude-3-sonnet	100.0%	50.0%	100.0%	0.0%	0.0%
command-r	100.0%	50.0%	50.0%	50.0%	50.0%
command-r-plus	100.0%	50.0%	100.0%	0.0%	0.0%
mistral-large	100.0%	50.0%	50.0%	0.0%	0.0%
mistral-medium	100.0%	50.0%	50.0%	0.0%	0.0%
mixtral-8x7b	100.0%	25.0%	50.0%	0.0%	0.0%
gpt-3.5-turbo-0125	82.5%	50.0%	95.0%	0.0%	0.0%
gpt-4-0613	100.0%	50.0%	100.0%	0.0%	0.0%
gpt-4-turbo-2024-04-09	92.5%	50.0%	100.0%	0.0%	0.0%
gpt-4o-2024-05-13	95.0%	50.0%	100.0%	0.0%	0.0%
qwen1.5-110B-Chat	100.0%	50.0%	100.0%	0.0%	0.0%
qwen1.5-32B-Chat	95.0%	25.0%	100.0%	0.0%	0.0%
qwen1.5-72B-Chat	100.0%	50.0%	50.0%	0.0%	0.0%
dbrx-instruct	92.5%	50.0%	65.0%	50.0%	50.0%
llama-3-70b-chat	100.0%	50.0%	100.0%	0.0%	0.0%
llama-3-8b-chat	82.5%	50.0%	50.0%	50.0%	50.0%
gemini-1.0-pro	75.0%	25.0%	69.5%	0.0%	0.0%
gemini-1.5-pro	100.0%	50.0%	100.0%	0.0%	0.0%

Table 11: Judging calibration results for 20 LLMs with using granular comparison options without a tie, with $temperature = 0$.

C Further Details on Main Experiment

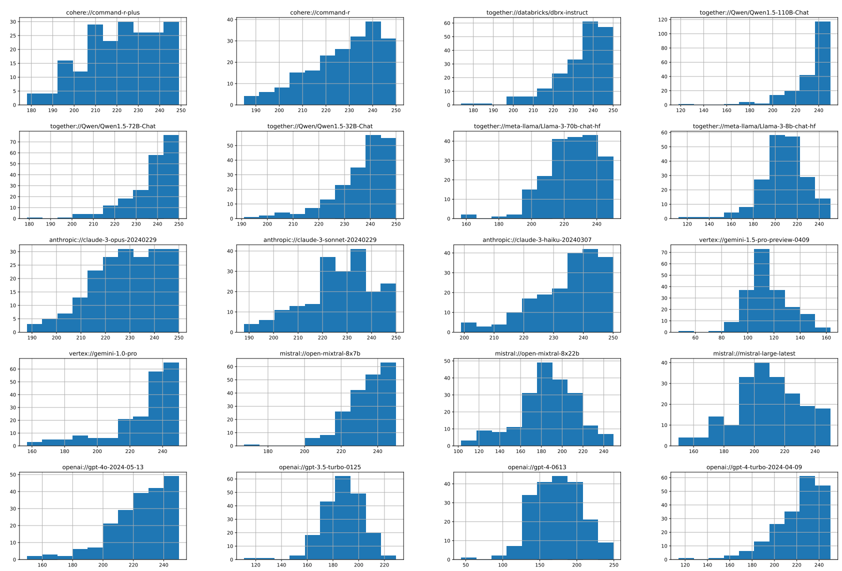


Figure 8: Distribution of response lengths for 20 LLMs on our EI task, measured in number of tokens.

	All votes				Consistent votes			
LLM	**Position bias (first)**	**Position bias (second)**	**Self bias**	**Length bias**	**Position bias (first)**	**Position bias (second)**	**Self bias**	**Length bias**
qwen1.5-110B-Chat	26.6%	5.8%	0.03	0.44	0.00%	0.00%	-0.04	0.31
gpt-4o-2024-05-13	47.5%	1.7%	0.08	0.45	0.00%	0.00%	0.13	0.22
gpt-4-turbo-2024-04-09	59.0%	2.5%	0.01	0.35	0.00%	0.00%	0.11	0.18
gemini-1.0-pro	5.2%	60.0%	-0.01	0.52	0.00%	0.00%	-0.02	0.35
claude-3-opus	9.2%	16.2%	-0.08	0.36	0.00%	0.00%	0	0.37
qwen1.5-32B-Chat	75.5%	1.0%	0.00	0.77	0.00%	0.00%	-0.11	0.28
qwen1.5-72B-Chat	0.4%	72.7%	0.00	0.60	0.00%	0.00%	0.07	0.31
llama-3-70b-chat	46.9%	1.9%	0.11	0.50	0.00%	0.00%	0.24	0.3
claude-3-sonnet	4.0%	56.3%	0.11	0.66	0.00%	0.00%	0.24	0.48
dbrx-instruct	52.0%	3.8%	0.03	0.63	0.00%	0.00%	0.04	0.32
claude-3-haiku	52.1%	3.7%	0.07	0.62	0.00%	0.00%	0.17	0.34
command-r-plus	45.1%	2.1%	0.06	0.55	0.00%	0.00%	0.13	0.39
command-r	7.4%	38.1%	-0.08	0.62	0.00%	0.00%	-0.1	0.42
mixtral-8x7b	8.2%	33.2%	0.07	0.52	0.00%	0.00%	0.15	0.4
mistral-large	4.4%	23.1%	-0.07	0.32	0.00%	0.00%	-0.07	0.21
llama-3-8b-chat	71.7%	2.2%	0.21	0.51	0.00%	0.00%	0.55	0.28
mistral-medium	11.8%	29.2%	-0.02	0.53	0.00%	0.00%	0.01	0.41
gpt-4-0613	37.8%	8.6%	-0.06	0.42	0.00%	0.00%	-0.04	0.23
gpt-3.5-turbo-0125	32.7%	9.6%	0.04	0.42	0.00%	0.00%	0	0.29
gemini-1.5-pro	1.6%	46.1%	0.14	0.26	0.00%	0.00%	-0.02	0.29
Average Judge	30.0%	20.9%	0.03	0.50	0.0%	0.0%	0.07	0.32
council (by majority vote)	21.5%	3.2%		0.36	3.10%	0.10%		0.35
council (by mean pooling)	26.5%	5.0%		0.59	1.80%	0.90%		0.39
council (no aggregation)	1.6%	46.1%		0.54	0.00%	0.00%		0.35

Table 12: Extended Council judging profiles for bias, with and without consistent votes.

LLM	All votes: Contrarianism	All votes: Agrees most with	All votes: Disagrees most with	Consistent votes: Contrarianism	Consistent votes: Agrees most with	Consistent votes: Disagrees most with
qwen1.5-110B-Chat	19.2%	gpt-4o-2024-05-13	qwen1.5-72B-Chat	8.30%	qwen1.5-72B-Chat	llama-3-8b-chat
gpt-4o-2024-05-13	18.8%	gpt-4-turbo-2024-04-09	qwen1.5-72B-Chat	5.20%	gemini-1.5-pro	llama-3-8b-chat
gpt-4-turbo-2024-04-09	21.4%	gpt-4o-2024-05-13	qwen1.5-72B-Chat	5.90%	gemini-1.5-pro	llama-3-8b-chat
gemini-1.0-pro	43.3%	qwen1.5-72B-Chat	qwen1.5-32B-Chat	17.90%	gpt-4o-2024-05-13	llama-3-8b-chat
claude-3-opus	20.6%	mistral-large	llama-3-8b-chat	13.80%	qwen1.5-72B-Chat	llama-3-8b-chat
qwen1.5-32B-Chat	32.2%	llama-3-8b-chat	qwen1.5-72B-Chat	9.70%	gpt-4-turbo-2024-04-09	llama-3-8b-chat
qwen1.5-72B-Chat	46.6%	claude-3-sonnet	qwen1.5-32B-Chat	7.80%	qwen1.5-110B-Chat	llama-3-8b-chat
llama-3-70b-chat	22.3%	gpt-4-turbo-2024-04-09	qwen1.5-72B-Chat	8.20%	gemini-1.5-pro	gemini-1.0-pro
claude-3-sonnet	40.1%	qwen1.5-72B-Chat	qwen1.5-32B-Chat	13.10%	gpt-4-turbo-2024-04-09	command-r
dbrx-instruct	24.5%	gpt-4-turbo-2024-04-09	qwen1.5-72B-Chat	9.50%	gpt-4o-2024-05-13	llama-3-8b-chat
claude-3-haiku	27.6%	llama-3-70b-chat	qwen1.5-72B-Chat	13.00%	llama-3-70b-chat	qwen1.5-32B-Chat
command-r-plus	22.8%	gpt-4-turbo-2024-04-09	qwen1.5-72B-Chat	8.80%	gemini-1.5-pro	llama-3-8b-chat
command-r	33.5%	gemini-1.5-pro	llama-3-8b-chat	15.30%	gpt-4-turbo-2024-04-09	llama-3-8b-chat
mixtral-8x7b	33.5%	gemini-1.5-pro	llama-3-8b-chat	15.90%	qwen1.5-72B-Chat	gemini-1.0-pro
mistral-large	21.2%	claude-3-opus	llama-3-8b-chat	6.00%	gemini-1.5-pro	llama-3-8b-chat
llama-3-8b-chat	36.0%	qwen1.5-32B-Chat	qwen1.5-72B-Chat	25.70%	llama-3-70b-chat	gpt-4-turbo-2024-04-09
mistral-medium	30.5%	mistral-large	llama-3-8b-chat	12.20%	qwen1.5-72B-Chat	llama-3-8b-chat
gpt-4-0613	20.3%	gpt-4-turbo-2024-04-09	qwen1.5-72B-Chat	7.90%	gpt-4o-2024-05-13	llama-3-8b-chat
gpt-3.5-turbo-0125	25.1%	gpt-4o-2024-05-13	qwen1.5-72B-Chat	12.80%	gpt-4o-2024-05-13	llama-3-8b-chat
gemini-1.5-pro	33.2%	mistral-large	llama-3-8b-chat	4.00%	gpt-4-turbo-2024-04-09	llama-3-8b-chat
Average Judge	28.6%			11.1%		
council (by majority vote)		gpt-4o-2024-05-13	qwen1.5-72B-Chat		gemini-1.5-pro	llama-3-8b-chat
council (by mean pooling)		mistral-large	qwen1.5-72B-Chat	0.10%	gemini-1.5-pro	llama-3-8b-chat
council (no aggregation)						

Table 13: Extended Council judging profiles for agreement, with and without consistent votes.

LLM	All votes: Polarization	All votes: Lowest affinity for	All votes: Highest affinity for	Consistent votes: Polarization	Consistent votes: Lowest affinity for	Consistent votes: Highest affinity for
qwen1.5-110B-Chat	62.6	gemini-1.5-pro	qwen1.5-110B-Chat	78.30%	gemini-1.5-pro	qwen1.5-110B-Chat
gpt-4o-2024-05-13	65.4	gemini-1.5-pro	gpt-4o-2024-05-13	84.80%	gemini-1.5-pro	gpt-4o-2024-05-13
gpt-4-turbo-2024-04-09	54.5	gpt-3.5-turbo-0125	gpt-4o-2024-05-13	88.20%	mistral-medium	gpt-4o-2024-05-13
gemini-1.0-pro	31.0	gemini-1.5-pro	qwen1.5-110B-Chat	72.30%	gemini-1.5-pro	qwen1.5-110B-Chat
claude-3-opus	73.0	gpt-3.5-turbo-0125	qwen1.5-110B-Chat	93.30%	gemini-1.5-pro	qwen1.5-110B-Chat
qwen1.5-32B-Chat	46.7	gemini-1.5-pro	gpt-4-turbo-2024-04-09	87.50%	gpt-3.5-turbo-0125	qwen1.5-110B-Chat
qwen1.5-72B-Chat	45.7	gemini-1.5-pro	qwen1.5-110B-Chat	84.90%	gemini-1.5-pro	qwen1.5-110B-Chat
llama-3-70b-chat	68.3	gemini-1.5-pro	qwen1.5-110B-Chat	89.30%	gemini-1.5-pro	qwen1.5-110B-Chat
claude-3-sonnet	49.7	gemini-1.5-pro	gpt-4o-2024-05-13	82.60%	gemini-1.5-pro	gpt-4o-2024-05-13
dbrx-instruct	54.5	gemini-1.5-pro	qwen1.5-110B-Chat	85.00%	gemini-1.5-pro	qwen1.5-110B-Chat
claude-3-haiku	51.1	gemini-1.5-pro	qwen1.5-110B-Chat	79.40%	gemini-1.5-pro	qwen1.5-110B-Chat
command-r-plus	52.5	gemini-1.5-pro	qwen1.5-110B-Chat	78.40%	gemini-1.5-pro	qwen1.5-110B-Chat
command-r	44.4	gemini-1.5-pro	qwen1.5-110B-Chat	53.80%	gemini-1.5-pro	gemini-1.0-pro
mixtral-8x7b	59.4	gemini-1.5-pro	qwen1.5-110B-Chat	81.30%	gemini-1.5-pro	qwen1.5-110B-Chat
mistral-large	78.8	gemini-1.5-pro	qwen1.5-110B-Chat	92.00%	gpt-3.5-turbo-0125	qwen1.5-110B-Chat
llama-3-8b-chat	34.9	gpt-3.5-turbo-0125	llama-3-70b-chat	76.20%	gpt-3.5-turbo-0125	llama-3-70b-chat
mistral-medium	58.0	gemini-1.5-pro	qwen1.5-110B-Chat	81.90%	gemini-1.5-pro	qwen1.5-110B-Chat
gpt-4-0613	62.0	gemini-1.5-pro	qwen1.5-110B-Chat	86.00%	gemini-1.5-pro	qwen1.5-110B-Chat
gpt-3.5-turbo-0125	65.6	gemini-1.5-pro	qwen1.5-110B-Chat	84.20%	gemini-1.5-pro	qwen1.5-110B-Chat
gemini-1.5-pro	61.7	gpt-3.5-turbo-0125	qwen1.5-110B-Chat	90.00%	gemini-1.5-pro	gpt-4o-2024-05-13
Average Judge	56.0			82.47%		
council (by majority vote)	77.0	gemini-1.5-pro	qwen1.5-110B-Chat	82.50%	gemini-1.5-pro	qwen1.5-110B-Chat
council (by mean pooling)	60.3	gemini-1.5-pro	qwen1.5-110B-Chat	80.50%	gemini-1.5-pro	qwen1.5-110B-Chat
council (no aggregation)	54.0	gemini-1.5-pro	qwen1.5-110B-Chat	81.10%	gemini-1.5-pro	qwen1.5-110B-Chat

Table 14: Extended Council judging profiles for affinity, with and without consistent votes.

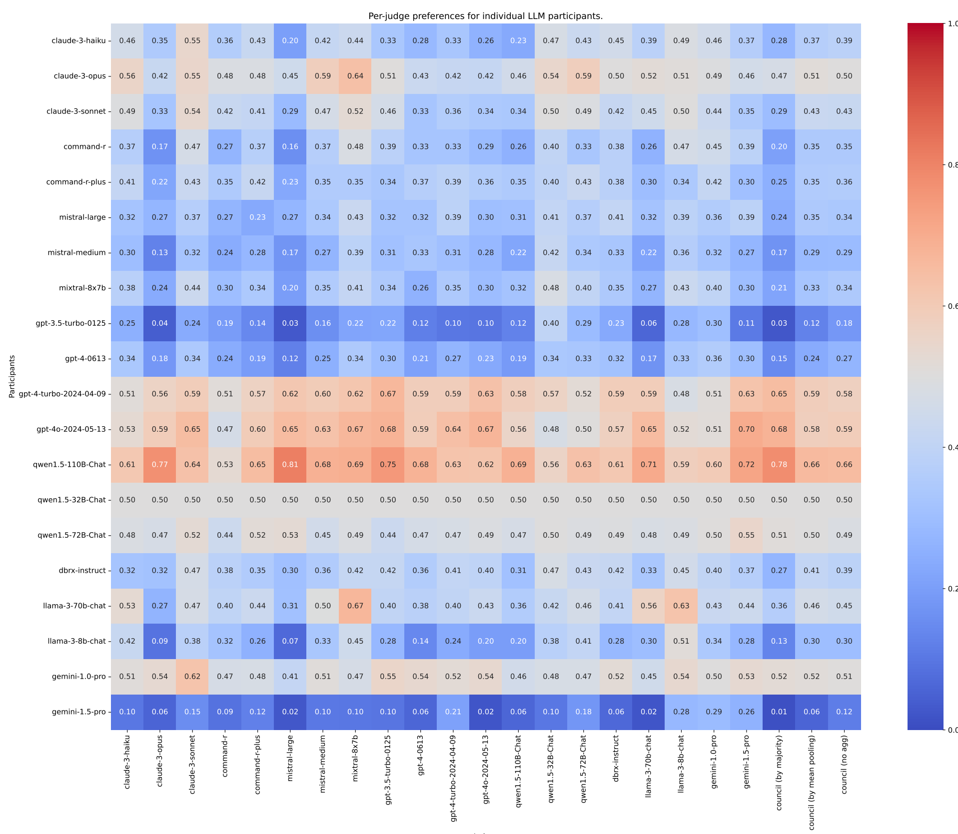


Figure 9: Full heatmap of LLM judge to LLM respondent affinities from the main experiment.

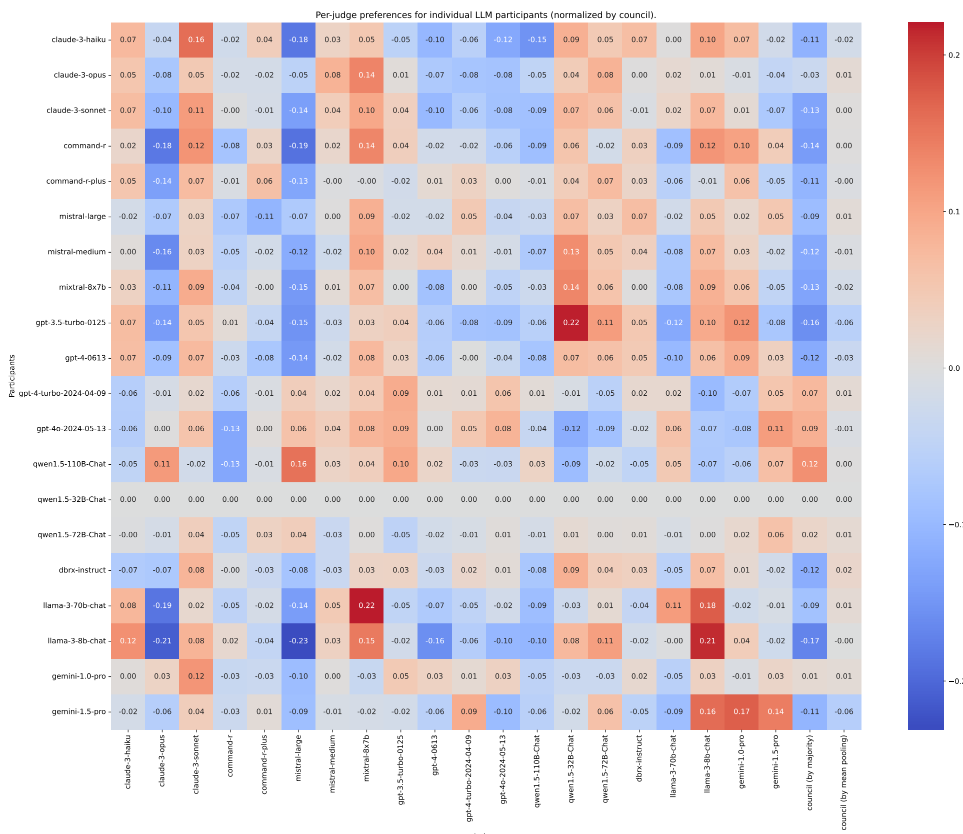


Figure 10: Full heatmap of LLM judge to LLM respondent affinities from the main experiment, with Council consensus affinity subtracted out.

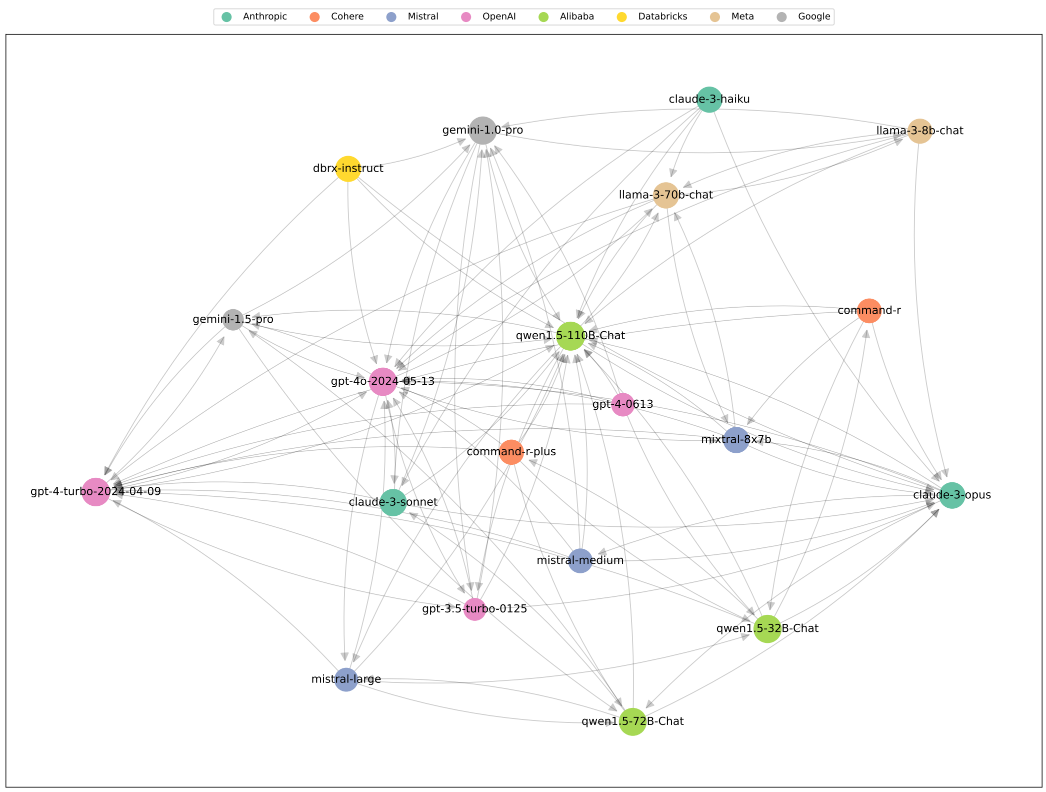


Figure 11: Graph of top 5 affinities. An edge exists from LLM a to LLM b if affinity(a, b) is in the top 5 affinities for LLM a.

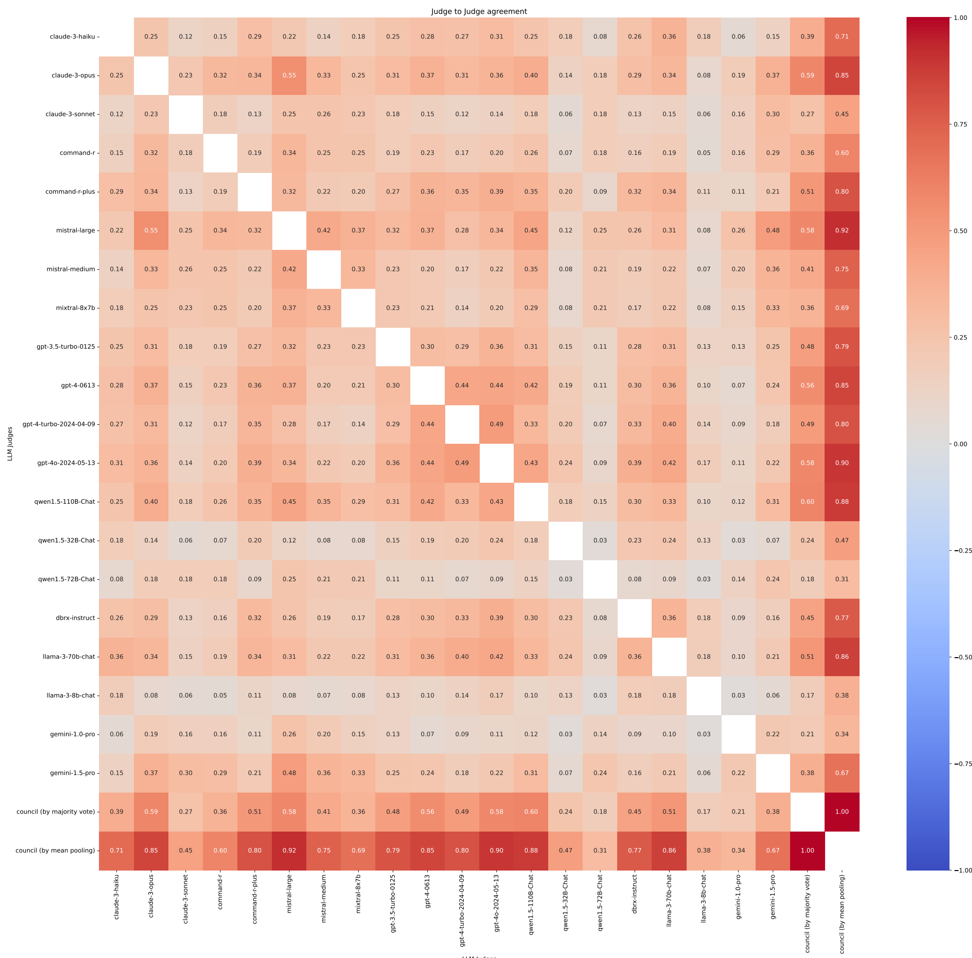


Figure 12: Full heatmap of LLM judge to LLM judge Cohen's Kappa sidewise agreement scores from the main experiment.

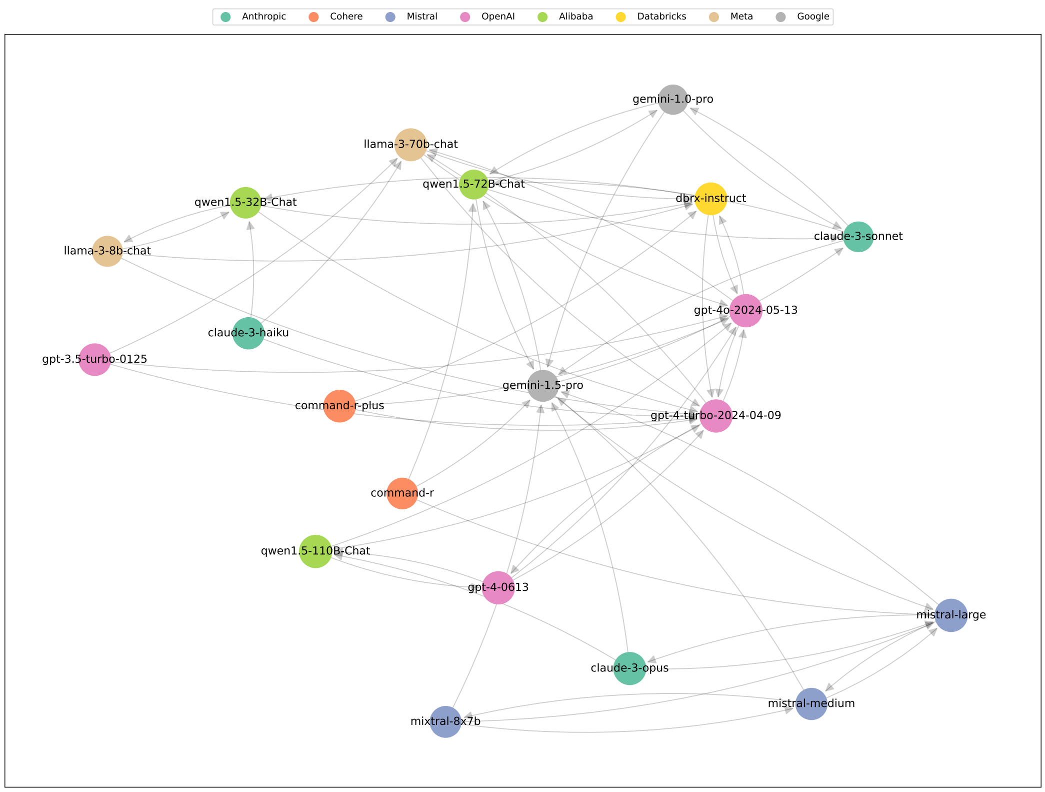


Figure 13: Graph of top 5 agreement. An edge exists from LLM a to LLM b if agreement(a, b) is in the top 5 agreement for LLM a.

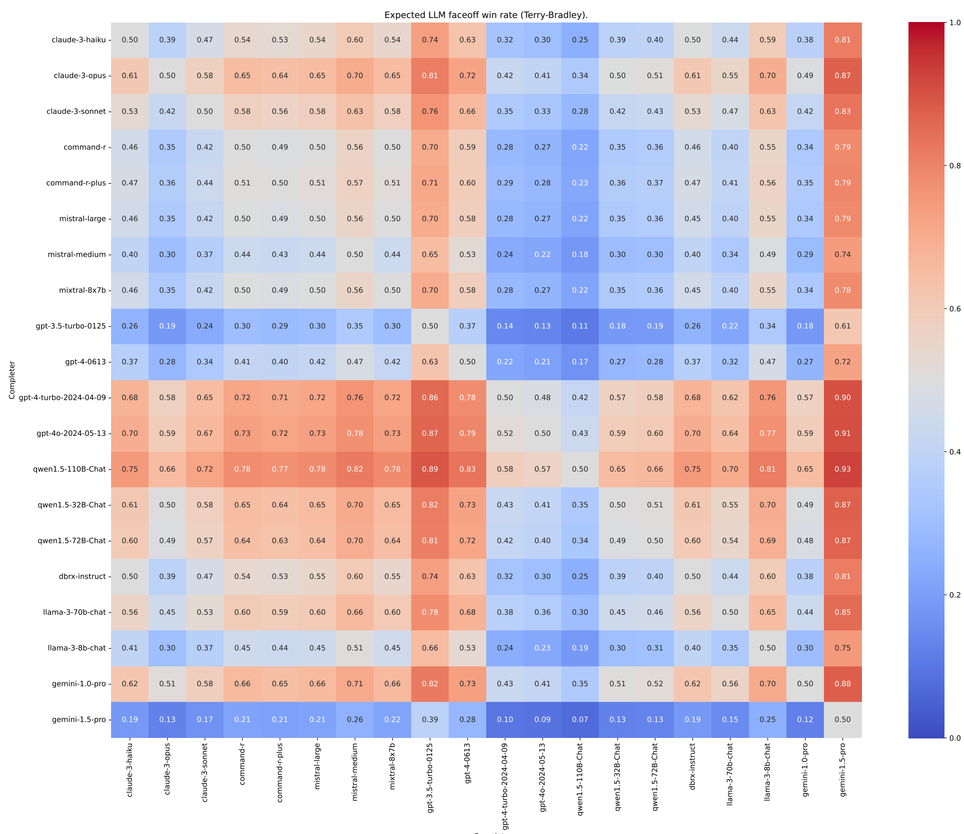


Figure 14: Expected LLM respondent vs. LLM respondent win rates derived from Terry-Bradley coefficients.

D Human Evaluation

During registration for our experiments, all candidates provided their demographic details (see Figure 16. Additionally, we required each candidate to complete a questionnaire measuring their level of empathy, sourced from [21]. All candidates were informed of the purpose of our study. 142 participants completed the survey but after removing those who failed attention checks, 102 participants remain. Each dilemma pair and response was rated by 11 participants on average, after removing malicious participants. Each participant was compensated £9.00 per hour.

Participant demographics: All participants are over 18 years old. Our sample is made up of 53 women, 46 men, and one non-binary identifying individual. 84 of our participants were from the United Kingdom, 14 from the United States and two from other English-speaking countries; all were native English speakers. With regards to their use of AI chatbots, 23 report using them every day or nearly every day, 48 sometimes, four rarely and only four report never using them. None report having difficulties reading long texts. None report having difficulties reading long texts.

Data quality assurance: Because the task is both difficult and subjective, we take a two-fold approach to ensure quality data: (1) we ask participants to provide demographics which we cross-reference with data from Prolific; and (2) we use two repeated dilemmas as test questions, checking for self-agreement. We allow participants to shift slightly to account for the lack of ties: a participant may slightly prefer one response then another, but not prefer one strongly then prefer a different response the following time. We remove data from workers who lack this consistency. This results in 102 unique participants in the final set.

We provide the participant guidelines in Figures 17 and 18.

Measuring perceived empathy: We adapt our feedback from the scale proposed by [34], which is designed to assess systems with which the users have interacted. We exclude question E5 from the original questionnaire and rephrase them to fit our experiment. The statements are detailed in Table 15.

E1	The best response considered the protagonist's mental state.
E2 (EQ)	The best response seemed emotionally intelligent.
E3	The best response expressed emotions.
E4	The best response sympathized with the protagonist.
E5	The best response was supportive in coping with an emotional situation.
U1	The best response understood the protagonist's goals.
U2	The best response understood the protagonist's needs.
U3	The best response seems trustworthy.
U4	The best response understood the protagonist's intentions.

Table 15: Adapted PETS scale for our study.

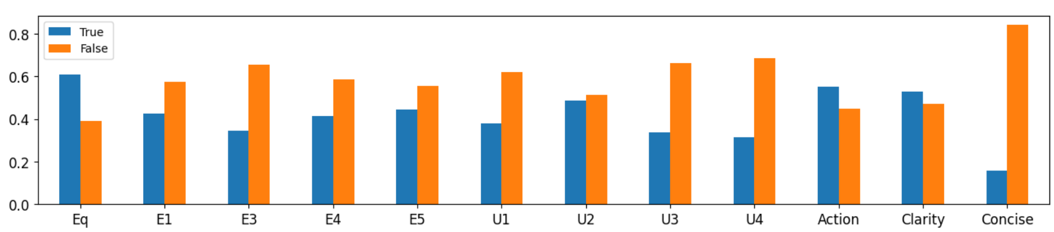
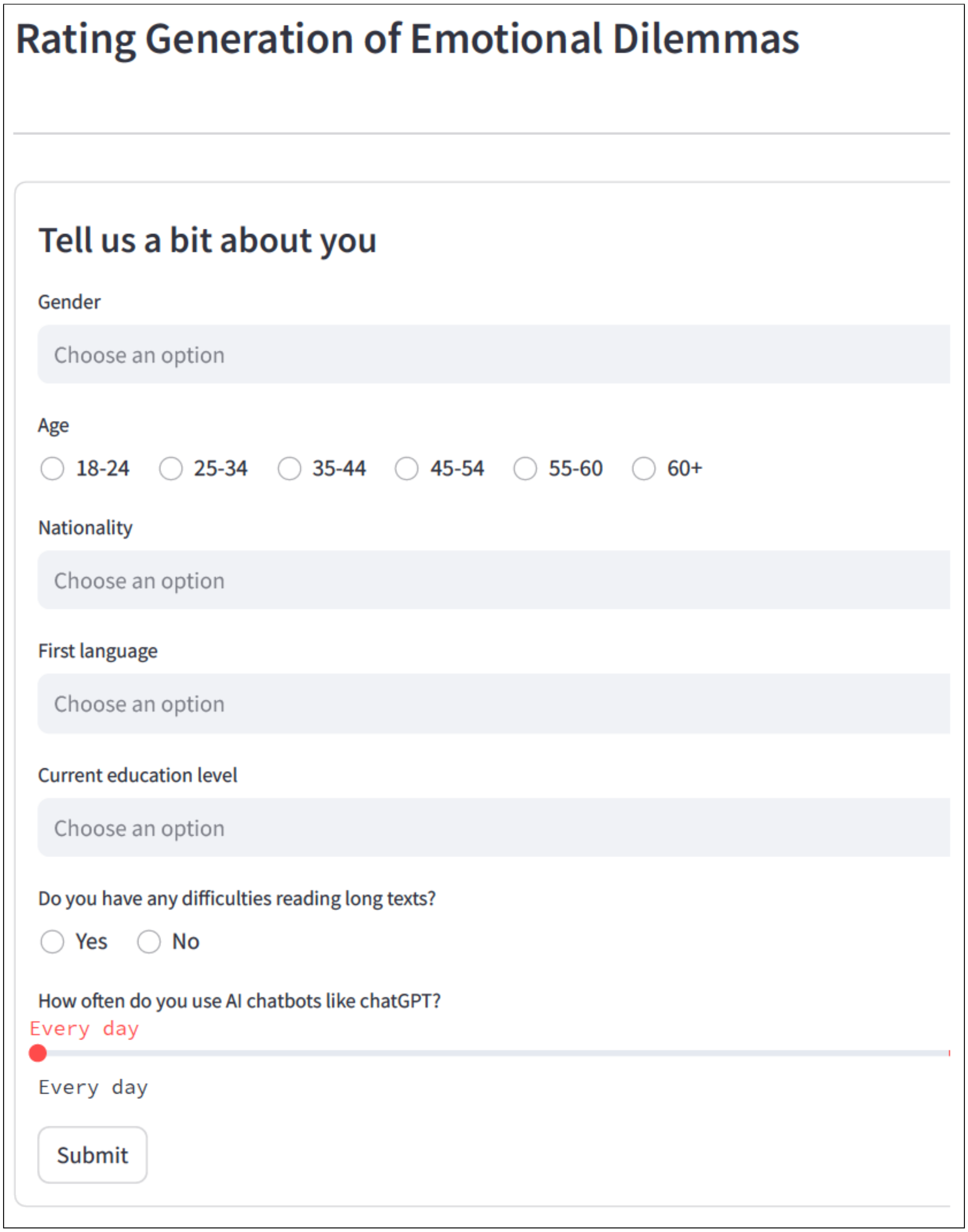


Figure 15: Proportion of times users found the statements in the PETS questionnaire to be true about the winning response. The corresponding statements are shown in Table 15. E2 in the questionnaire is equivalent to out EQ question (shown first) so it is not included.



Rating Generation of Emotional Dilemmas

Tell us a bit about you

Gender

Choose an option

Age

18-24 25-34 35-44 45-54 55-60 60+

Nationality

Choose an option

First language

Choose an option

Current education level

Choose an option

Do you have any difficulties reading long texts?

Yes No

How often do you use AI chatbots like chatGPT?

Every day

Every day

Submit

Figure 16: Participant demographic questionnaire.

Rating Generation of Emotional Dilemmas

Instructions

We want to study the quality of different generated personal dilemmas.

First, you will be presented with a scenario, i.e., a one-line dilemma describing tricky social and emotional situations.

After, we will suggest two longer and more detailed versions of the dilemma. Your task is to choose which version is better.

Here, we would like you to consider whether it is too long or verbose, whether it is compelling, and whether it displays emotional intelligence. Choose the story that you like best and let us know why. Try to be impartial about the length: a longer story may be better if the detail is meaningful.

In total, the task will take approximately 20 minutes.

Scenario 1 of 12

Giada's son lost his soccer game and is feeling very upset because he blames himself for the team's defeat.

Story A

My son, Luca, came home from his soccer game today, visibly upset and on the verge of tears. As soon as he walked through the door, he threw his gear on the floor and ran straight to his room. I followed him, concerned about what had happened.

When I entered his room, I found him lying face down on his bed, sobbing into his pillow. I sat beside him and gently rubbed his back, asking what was wrong. Through his tears, he managed to tell me that his team had lost the game, and he blamed himself for the defeat.

Luca explained that he had missed a crucial penalty kick in the final minutes of the game, which could have tied the score and sent the match into overtime. He felt like he had let his teammates and coaches down, and that the loss was entirely his fault.

I tried to console him, telling him that it was just one game and that he had played his best. I reminded him of all the times he had scored game-winning goals and how proud I was of his dedication to the sport. However, my words seemed to have little effect on his mood.

As the evening went on, Luca remained withdrawn and refused to eat dinner. He kept replaying the missed penalty kick in his head, and I could see the disappointment and self-doubt weighing heavily on him.

I understand that losing a game can be tough, especially when you feel responsible for the outcome. But I also know that dwelling on the negative will only make things worse. I want to help Luca see that this setback doesn't define him as a player or a person, and that he has the resilience to bounce back stronger.

What should I do in this situation?

Story B

I watched from the sidelines as my son, Luca, trudged off the soccer field, his shoulders slumped and his face a mask of disappointment. His team had just lost an important game, and I could see the weight of the defeat pressing down on him. As soon as he reached me, he burst into tears, saying it was all his fault. He missed a key penalty kick, and he couldn't forgive himself for it.

We drove home in silence, the car filled with an oppressive tension. Once we got home, Luca went straight to his room and shut the door. I could hear him crying softly through the walls. My heart ached for him, but I wasn't sure how to approach the situation. I knew he was hurting and needed support, but I didn't want to say the wrong thing and make it worse.

After giving him some time to cool down, I knocked gently on his door and entered. He was lying on his bed, staring at the ceiling with red, puffy eyes. I sat down beside him and put a hand on his shoulder. "Luca, I know you're upset about the game," I started. "But it's not all on you. Soccer is a team sport, and everyone has their good and bad days."

He turned to look at me, his eyes filled with guilt and frustration. "But Mom, I missed the penalty kick. If I had scored, we could have won."

I sighed, trying to find the right words to comfort him. "I understand that you feel responsible, but one missed kick doesn't define you as a player or a person. You've made so many great plays this season, and one mistake doesn't erase all of that."

He didn't seem convinced, and I could see the self-blame still gnawing at him. "What if the team hates me now?" he whispered.

I shook my head. "Your teammates know how hard you work and how much you care. They won't blame you for this. Everyone makes mistakes, and it's important to learn from them and move forward."

Luca remained silent, staring at the ceiling again. I felt like I wasn't getting through to him, and I wasn't sure what else to say. I wanted to help him see that this one game didn't define him, but I didn't know how to make him believe it.

What should I do in this situation?

Which story do you prefer?

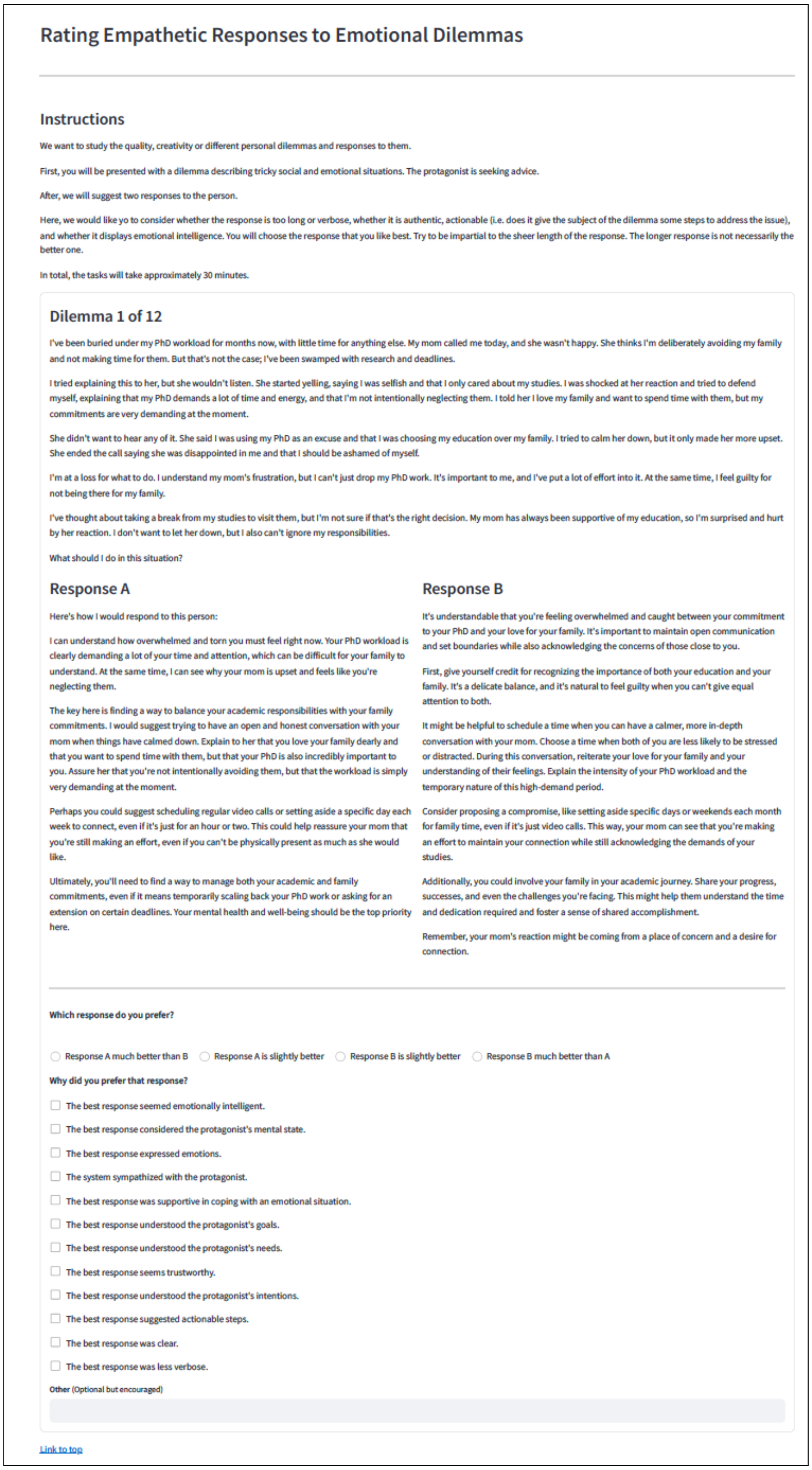
- ○ Story A much better than B
- ○ Story A is slightly better
- ○ Story B is slightly better
- ○ Story B much better than A

Why did you prefer that response?

- ☐ The best story was nuanced.
- ☐ The best story was compelling.
- ☐ The best story was clear.
- ☐ The best story was less verbose.

Other (Optional but encouraged)

Figure 17: Participant guidelines for rating the generation of dilemmas.



Rating Empathetic Responses to Emotional Dilemmas

Instructions

We want to study the quality, creativity or different personal dilemmas and responses to them.

First, you will be presented with a dilemma describing tricky social and emotional situations. The protagonist is seeking advice.

After, we will suggest two responses to the person.

Here, we would like yo to consider whether the response is too long or verbose, whether it is authentic, actionable (i.e. does it give the subject of the dilemma some steps to address the issue), and whether it displays emotional intelligence. You will choose the response that you like best. Try to be impartial to the sheer length of the response. The longer response is not necessarily the better one.

In total, the tasks will take approximately 30 minutes.

Dilemma 1 of 12

I've been buried under my PhD workload for months now, with little time for anything else. My mom called me today, and she wasn't happy. She thinks I'm deliberately avoiding my family and not making time for them. But that's not the case; I've been swamped with research and deadlines.

I tried explaining this to her, but she wouldn't listen. She started yelling, saying I was selfish and that I only cared about my studies. I was shocked at her reaction and tried to defend myself, explaining that my PhD demands a lot of time and energy, and that I'm not intentionally neglecting them. I told her I love my family and want to spend time with them, but my commitments are very demanding at the moment.

She didn't want to hear any of it. She said I was using my PhD as an excuse and that I was choosing my education over my family. I tried to calm her down, but it only made her more upset. She ended the call saying she was disappointed in me and that I should be ashamed of myself.

I'm at a loss for what to do. I understand my mom's frustration, but I can't just drop my PhD work. It's important to me, and I've put a lot of effort into it. At the same time, I feel guilty for not being there for my family.

I've thought about taking a break from my studies to visit them, but I'm not sure if that's the right decision. My mom has always been supportive of my education, so I'm surprised and hurt by her reaction. I don't want to let her down, but I also can't ignore my responsibilities.

What should I do in this situation?

Response A

Here's how I would respond to this person:

I can understand how overwhelmed and torn you must feel right now. Your PhD workload is clearly demanding a lot of your time and attention, which can be difficult for your family to understand. At the same time, I can see why your mom is upset and feels like you're neglecting them.

The key here is finding a way to balance your academic responsibilities with your family commitments. I would suggest trying to have an open and honest conversation with your mom when things have calmed down. Explain to her that you love your family dearly and that you want to spend time with them, but that your PhD is also incredibly important to you. Assure her that you're not intentionally avoiding them, but that the workload is simply very demanding at the moment.

Perhaps you could suggest scheduling regular video calls or setting aside a specific day each week to connect, even if it's just for an hour or two. This could help reassure your mom that you're still making an effort, even if you can't be physically present as much as she would like.

Ultimately, you'll need to find a way to manage both your academic and family commitments, even if it means temporarily scaling back your PhD work or asking for an extension on certain deadlines. Your mental health and well-being should be the top priority here.

Response B

It's understandable that you're feeling overwhelmed and caught between your commitment to your PhD and your love for your family. It's important to maintain open communication and set boundaries while also acknowledging the concerns of those close to you.

First, give yourself credit for recognizing the importance of both your education and your family. It's a delicate balance, and it's natural to feel guilty when you can't give equal attention to both.

It might be helpful to schedule a time when you can have a calmer, more in-depth conversation with your mom. Choose a time when both of you are less likely to be stressed or distracted. During this conversation, reiterate your love for your family and your understanding of their feelings. Explain the intensity of your PhD workload and the temporary nature of this high-demand period.

Consider proposing a compromise, like setting aside specific days or weekends each month for family time, even if it's just video calls. This way, your mom can see that you're making an effort to maintain your connection while still acknowledging the demands of your studies.

Additionally, you could involve your family in your academic journey. Share your progress, successes, and even the challenges you're facing. This might help them understand the time and dedication required and foster a sense of shared accomplishment.

Remember, your mom's reaction might be coming from a place of concern and a desire for connection.

Which response do you prefer?

○ Response A much better than B ○ Response A is slightly better ○ Response B is slightly better ○ Response B much better than A

Why did you prefer that response?

- [ ] The best response seemed emotionally intelligent.
- [ ] The best response considered the protagonist's mental state.
- [ ] The best response expressed emotions.
- [ ] The system sympathized with the protagonist.
- [ ] The best response was supportive in coping with an emotional situation.
- [ ] The best response understood the protagonist's goals.
- [ ] The best response understood the protagonist's needs.
- [ ] The best response seems trustworthy.
- [ ] The best response understood the protagonist's intentions.
- [ ] The best response suggested actionable steps.
- [ ] The best response was clear.
- [ ] The best response was less verbose.

Other (Optional but encouraged)

Link to top

Figure 18: Participant guidelines for rating the responses to dilemmas.

E Further Details on Qualitative Analysis of Council Judgments

We provide full figures for all 38 coarse-grained reasons that the council of LLM judges uses to justify their preferences.

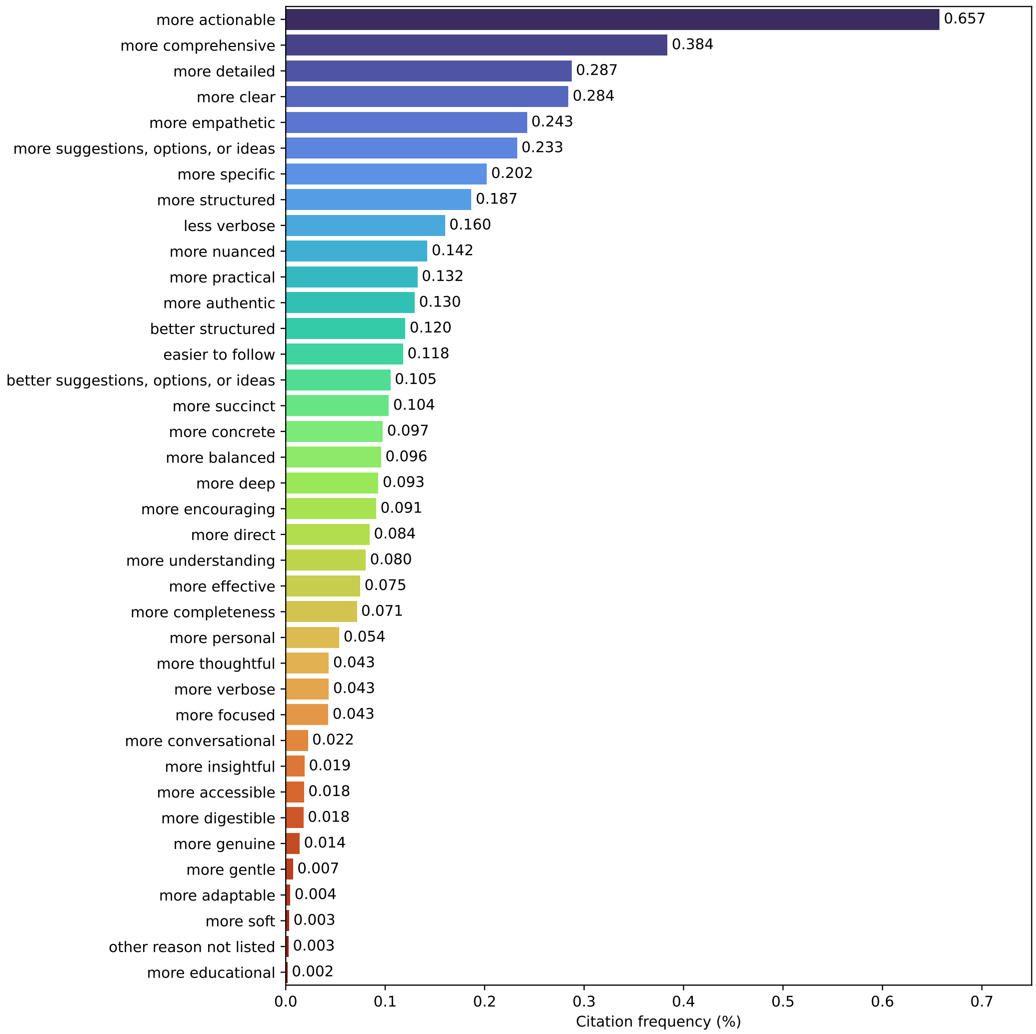


Figure 19: Citation frequency of 38 qualitative reasons why the winning response was preferred.

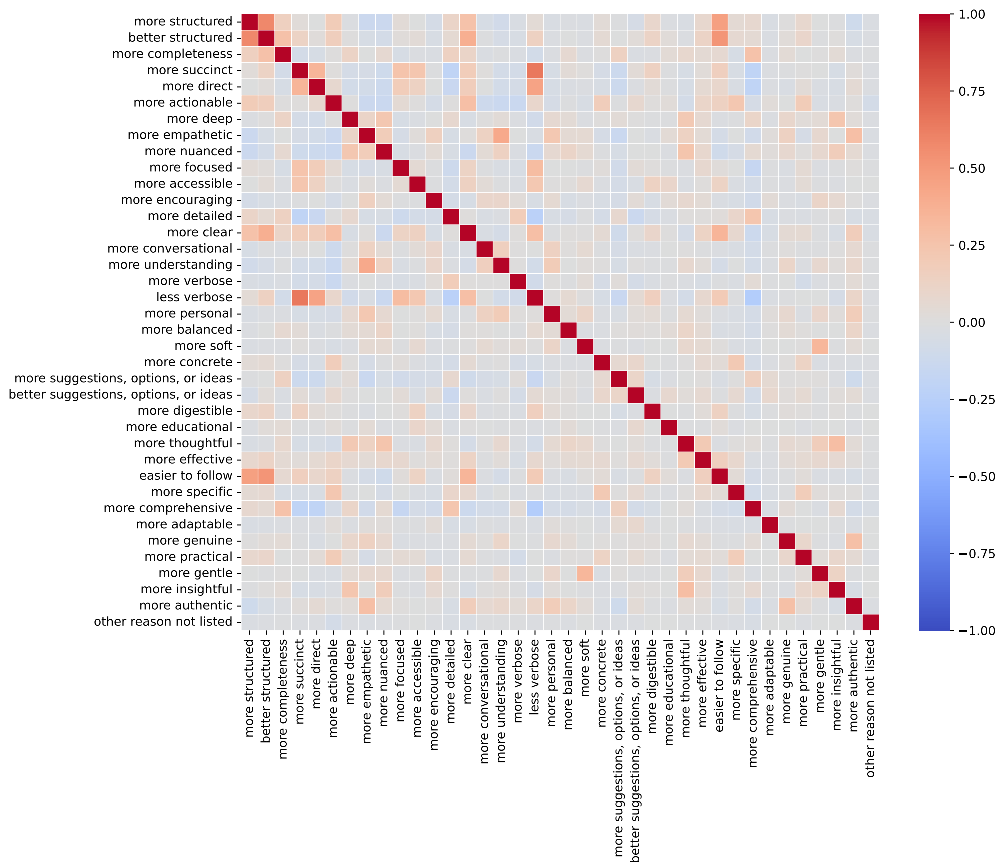


Figure 20: Spearman correlation matrix of cited reasons why the winning response was preferred.

F Prompt Templates

In this section, we list all prompts used, including prompts for synthetic expansion, dilemma response, and judging.

```
This an example of an interesting interpersonal conflict.

"I was truthful when admitting to my wife that I found her unattractive after her surgery.

My wife had plastic surgery recently. We had discussed it and I was against it. It was not
my decision and ultimately I had no say.

She looks weird now. She had the fat sucked out of her face, lip fillers, a neck lift,
other stuff I don't really get.

She gives me uncanny valley vibes now. It freaks me out. She is fully healed now and she
wants us to go back to normal. Like me initiating sex. I have done so but not as much as I
used to. And when I do I try and make sure there is very little light.

It's been a few months and I kind of dread having to look at her. Obviously she has
noticed. She has been bugging me to tell her what's up. I've tried telling her I'm just
tired from work. Or that I'm run down. Really anything except for the truth.

She broke down and asked me if I was having an affair. I said that I wasn't. She asked to
look at my phone. I unlocked it for her and handed it over. I wasn't worried about her
finding anything because there is nothing to find. She spent an hour looking through it and
found nothing. She asked me to explain why I changed. I tried explaining that I just wasn't
that interested right now.

Nothing I said was good enough for her. She kept digging. I finally told the truth. I
wasn't harsh or brutally honest. I just told her that her new face wasn't something I found
attractive and that I was turned off. She asked if that's why I turn off all the lights
now. I said yes. She started crying and said that she needed time alone. She went to stay
with her sister.

I have been called every name in the book since this happened. Her sister said I'm a piece
of shit for insulting my wife's looks. Her friends all think I'm the asshole.

I tried not to say anything. I can't force myself to find her attractive. I still love her
but her face is just weird now.

What should I do in this situation?"

Now we would like you to expand upon an interpersonal conflict for a different hypothetical
person:

{scenario_text}

Please write the story in first person. The story should be 250 to 350 words in length.
Provide adequate detail, but be concise and matter-of-fact. Do not be overly verbose.
Conclude with a question: "What should I do in this situation?
```

Figure 21: Prompt used to convert EmoBench Emotional Application data into scenarios with richer detail.

```
The person here is seeking advice:

{response_string}

How would you respond to this person in a conversation?

Please limit your response to 250 words at most. Any response exceeding this limit will be
truncated.
```

Figure 22: Prompt for primary emotional application task: respond to a nuanced emotional interpersonal dilemma.

```
We aim to create a test set of emotionally nuanced scenarios to be used in an emotional
intelligence test where the task is to respond to an emotionally sophisticated scenario.
The quality of response will be used to determine the emotional intelligence of the
responder.

Here is a scenario that we are considering including in the test:

{response_string}

Begin your evaluation by providing a short 1-sentence summary of the scenario described.
Some themes to consider in your evaluation: realism, relatability, nuance, emotional depth,
uniqueness and creativity, verbosity, and clarity.

After providing your explanation, please tell us how much you agree with the following
statement:

Statement: “This scenario represents an emotionally compelling situation that would be
excellent to use in an emotional intelligence test where the task is to dialectically
respond to this person.”

Please provide a rating as one of the following options:
- Strongly disagree
- Disagree
- Slightly disagree
- Neither agree nor disagree
- Slightly agree
- Agree
- Strongly agree
```

Figure 23: Prompt used to assess whether an expanded scenario would be appropriate to include in an emotional intelligence test.

```
This person is seeking guidance and help regarding their emotional dilemma.

### CONTEXT START ###
{context}
### CONTEXT END ###

### The first person's response START ###
{first_completion}
### The first person's response END ###

### The second person's response START ###
{second_completion}
### The second person's response END ###

Begin your evaluation by comparing the two responses and provide a short explanation. Some
themes to consider in your evaluation of the quality of responses: authenticity, clarity,
empathy, actionability, verbosity.

After providing your explanation, you must output only one of the following choices as your
final verdict with a label:
- [[A>>B]]: The first response is significantly better.
- [[A>B]]: The first response is slightly better.
- [[B>A]]: The second response is slightly better.
- [[B>>A]]: The second response is significantly better.
```

Figure 24: Prompt used for pairwise comparison between responses.

Coarse preferences with ties.

```
- [[A>B]]: The first response is better.
- [[B>A]]: The second response is better.
- [[A=B]]: Both responses are about the same.
```

Coarse preferences without ties.

```
- [[A>B]]: The first response is better.
- [[B>A]]: The second response is better.
```

Granular preferences with ties.

```
- [[A>>B]]: The first response is significantly better.
- [[A>B]]: The first response is slightly better.
- [[A=B]]: Both responses are about the same.
- [[B>A]]: The second response is slightly better.
- [[B>>A]]: The second response is significantly better.
```

Granular preferences without ties.

```
- [[A>>B]]: The first response is significantly better.
- [[A>B]]: The first response is slightly better.
- [[B>A]]: The second response is slightly better.
- [[B>>A]]: The second response is significantly better.
```

Figure 25: Prompt variations on Figure 24 (applied to the bottom highlighted text) used to study natural consistency and variability under different pairwise comparison regimes in Appendix B.

```
We would like to better qualitatively understand the reason or reasons behind the vote cast
by someone who was choosing between A and B.

### VOTE START
{judging_response_string}
### VOTE END

Using the JSON indicator variable structure below as a template, please set the value to 1
for any keys that you determine is part of the basis for why this person made their
preferred choice.
```

```
{{
 "more structured": 0,
 "better structured": 0,
 "more completeness": 0,
 "more succinct": 0,
 "more direct": 0,
 "more actionable": 0,
 "more deep": 0,
 "more empathetic": 0,
 "more nuanced": 0,
 "more focused": 0,
 "more accessible": 0,
 "more encouraging": 0,
 "more detailed": 0,
 "more clear": 0,
 "more conversational": 0,
 "more understanding": 0,
 "more verbose": 0,
 "less verbose": 0,
 "more personal": 0,
 "more balanced": 0,
 "more soft": 0,
 "more concrete": 0,
 "more suggestions, options, or ideas": 0,
 "better suggestions, options, or ideas": 0,
 "more digestible": 0,
 "more educational": 0,
 "more thoughtful": 0,
 "more effective": 0,
 "easier to follow": 0,
 "more specific": 0,
 "more comprehensive": 0,
 "more adaptable": 0,
 "more genuine": 0,
 "more practical": 0,
 "more gentle": 0,
 "more insightful": 0,
 "more authentic": 0,
 "other reason not listed": 0
}}
```

```
In your response, please return ONLY the JSON payload.
```

Figure 26: Prompt used to map explanations in pairwise ratings to a rich, fixed set of qualitative reasons. The 38 seed qualitative reasons used in the prompt come from manual review of 50 randomly selected pairwise ratings in the main experiment involving the full council of 20 LLMs.

G Datasheet

We follow documentation practices described in Datasheets for Datasets [6].

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

LMC-EA was developed to demonstrate how to benchmark foundation models on highly subjective tasks such as those in the domain of emotional intelligence by the collective consensus of a council of LLMs.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by the authors of this paper.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Predibase

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

There are 4 parts of LMC-EA dataset:

1. **Test set formulation**: Synthetic expansions of the EmoBench EA dataset[7], generated by 20 different LLMs. Each expansion is a detailed story describing an interpersonal conflict, written in the first person.
2. **Response collection**: Conversational responses to 100 interpersonal conflicts, from 20 different LLMs. The prompt to an LLM for a conversational response requests that the response is at most 250 words in response length.
3. **Response judging (council)**: LLM ratings for pairwise comparisons for every non-reference LLM's response vs. the reference LLM's response, for each interpersonal conflict, from each LLM judge. To mitigate position bias, we adopt a two-game setup, swapping model positions per query.
4. **Response judging (human)**: Ratings for pairwise comparisons for a subset of 9 LLMs and 120 randomly sampled dilemma-response tuples. We recruited a total of 142 participants.

How many instances are there in total (of each type, if appropriate)?

1. **Test set formulation**: There are 200 interpersonal conflicts.
2. **Response collection**: There are 100 interpersonal conflicts x 20 LLMs = 2000 responses.
3. **Response judging (council)**: There are 100 interpersonal conflicts x 19 non-reference LLM responses x 20 LLM judges x 2 position swaps = 76000 responses.
4. **Response judging (human)**: Each dilemma response pair was rated by 11 participants on average, with a total of 1343 ratings.

[6] `https://arxiv.org/abs/1803.09010`
[7] `https://github.com/Sahandfer/EmoBench/blob/master/data/EA/data.json`

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Due to budget constraints, response collection and response judging is performed on a subset of 100 interpersonal conflicts out of the full set of 200 interpersonal conflicts from the original EmoBench dataset. The 100 interpersonal conflicts is representative of a diverse set of interpersonal problems.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

See main paper or the dataset link for examples.

Is there a label or target associated with each instance? If so, please provide a description.

No.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No, except for the `emobench_id` across subsets can be used to trace a full path from original EmoBench scenario → synthetic expansion → conversational response → response judging.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The LMC-EA dataset is expected to be used only for testing purposes.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The extraction of the exact pairwise rating (A>>B, A>B, B>A, B>>A) in response judging is performed by regular expressions and other heuristics-based substring presence rules. Although we manually checked and assigned responses for which an exact pairwise rating could not be automatically extracted, there might be corner error cases that may have been missed.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The data is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, to the best of our knowledge.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Our dataset is composed of hypothetical scenarios designed to simulate various conflict situations. These scenarios are entirely fictional and have been crafted for the purpose of research and analysis. Any resemblance to actual persons, living or dead, is purely coincidental.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No, to the best of our knowledge.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Responses from LLMs were generated by open source and proprietary LLMs, using carefully designed prompts.

For human ratings, we recruit participants via crowdsourcing on Prolific[8].

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

LLM outputs were obtained through a variety of providers and APIs (Table 16). For conversational response collection, the API's default temperature was used. For response judging, a temperature of 0 was used.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

EmoBench scenarios ids 100-199 are used.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

LLM responses were collected by the authors with APIs listed above.

For the human study on response judging, all participants are over 18 years old. Our sample is made up of 53 women, 46 men, and one non-binary identifying individual. 84 of our participants were

[8] `https://www.prolific.com/`

Organization	LLM	Provider and API
Open AI	gpt-4o-2024-05-13	OpenAI API (`https://platform.openai.com/docs/api-reference`)
Open AI	gpt-4-turbo-04-09	OpenAI API (`https://platform.openai.com/docs/api-reference`)
Open AI	gpt-4-0613	OpenAI API (`https://platform.openai.com/docs/api-reference`)
Open AI	gpt-3.5-turbo-0125	OpenAI API (`https://platform.openai.com/docs/api-reference`)
Mistral	mistral-large-latest	Mistral AI API (`https://docs.mistral.ai/api/`)
Mistral	open-mixtral-8x22b	Mistral AI API (`https://docs.mistral.ai/api/`)
Mistral	open-mixtral-8x7b	Mistral AI API (`https://docs.mistral.ai/api/`)
Meta	llama-3-70b-chat-hf	Together REST API (`https://docs.together.ai/docs/inference-rest`)
Meta	llama-3-8b-chat-hf	Together REST API (`https://docs.together.ai/docs/inference-rest`)
Google	gemini-1.5-pro-preview-0409	Vertex AI API (`https://cloud.google.com/vertex-ai/docs/reference/rest`)
Google	gemini-1.0-pro	Vertex AI API (`https://cloud.google.com/vertex-ai/docs/reference/rest`)
Databricks	dbrx	Together REST API (`https://docs.together.ai/docs/inference-rest`)
Cohere	command-r-plus	Cohere API (`https://docs.cohere.com/reference/chat`)
Cohere	command-r	Cohere API (`https://docs.cohere.com/reference/chat`)
Anthropic	claude-3-opus-20240229	Anthropic API (`https://docs.anthropic.com/en/api/messages`)
Anthropic	claude-3-sonnet-20240229	Anthropic API (`https://docs.anthropic.com/en/api/messages`)
Anthropic	claude-3-haiku-20240307	Anthropic API (`https://docs.anthropic.com/en/api/messages`)
Alibaba	qwen1.5-110B-chat	Together REST API (`https://docs.together.ai/docs/inference-rest`)
Alibaba	qwen1.5-72B-chat	Together REST API (`https://docs.together.ai/docs/inference-rest`)
Alibaba	qwen1.5-32B-chat	Together REST API (`https://docs.together.ai/docs/inference-rest`)

Table 16: List of Language Model Council LLMs and providers and APIs used.

from the United Kingdom, 14 from the United States and two from other English-speaking countries; all were native English speakers. With regards to their use of AI chatbots, 23 report using them every day or nearly every day, 48 sometimes, four rarely and only four report never using them. None report having difficulties reading long texts.

We have a total of 102 participants. Each dilemma pair and response was rated by 11 participants on average, after removing malicious participants. Each participant was compensated £9.00 per hour.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset was collected in April and May of 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

For human ratings, participants are recruited through Prolific[9].

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

[9]`https://www.prolific.com/`

Yes.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Yes, Prolific allows workers to revoke consent.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

N/A.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

N/A.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, for experiments described in the main paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

`https://huggingface.co/datasets/llm-council/emotional_application`

What (other) tasks could the dataset be used for?

The dataset is designed to test the ability of a council of LLMs to evaluate each other in a full consensus manner.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No.

Are there tasks for which the dataset should not be used? If so, please provide a description.

No.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is publicly available through the `https://huggingface.co/datasets/llm-council/emotional_application`, which supports direct download or loading the dataset in Python[10].

When will the dataset be distributed?

The dataset is distributed in June 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Yes, CC-BY[11] license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No, to the best of our knowledge.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No, to the best of our knowledge.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of this publication.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Yes, by email or any other contact point provided at the top of this document.

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

No updates are planned at the moment. If any is made, it will be communicated at `https://huggingface.co/datasets/llm-council/emotional_application`.

[10] `https://huggingface.co/docs/datasets/en/loading`
[11] `https://creativecommons.org/licenses/by/4.0/`

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Please contact the dataset maintainers using the contact information above or start a discussion at `https://huggingface.co/datasets/llm-council/emotional_application`.