Language Model Council: Democratically Benchmarking Foundation Models on Highly Subjective Tasks

Justin Zhao[1], Flor Miriam Plaza-del-Arco[2], Benji Genchel[1], Amanda Cercas Curry[2]

[1]Unaffiliated, [2]Bocconi University

https://llm-council.com

Abstract

Large Language Models (LLMs) have become more advanced, the search for efficient and accurate evaluation methods is ongoing. Many recent evaluations use other LLMs as judges to score outputs from other LLMs, often relying on a single large model like GPT-4o. However, using a single LLM judge is prone to intra-model bias, and many tasks – such as those related to emotional intelligence, creative writing, and persuasiveness – may be too subjective for a single model to judge fairly. We introduce the **Language Model Council (LMC)**, where a group of LLMs collaborate to create tests, respond to them, and evaluate each other's responses in a democratic fashion. Unlike previous approaches that focus on reducing cost or bias by using a panel of smaller models, our work examines the benefits and nuances of a fully inclusive LLM evaluation system. In a detailed case study on emotional intelligence, we deploy a council of 20 recent LLMs to rank each other on open-ended responses to interpersonal conflicts. Our results show that the LMC produces rankings that are more separable and more robust, and through a user study, we show that they are more consistent with human evaluations than any individual LLM judge. Using all LLMs for judging can be costly, however, so we use Monte Carlo simulations and hand-curated sub-councils to study hypothetical council compositions and discuss the value of the incremental LLM judge.

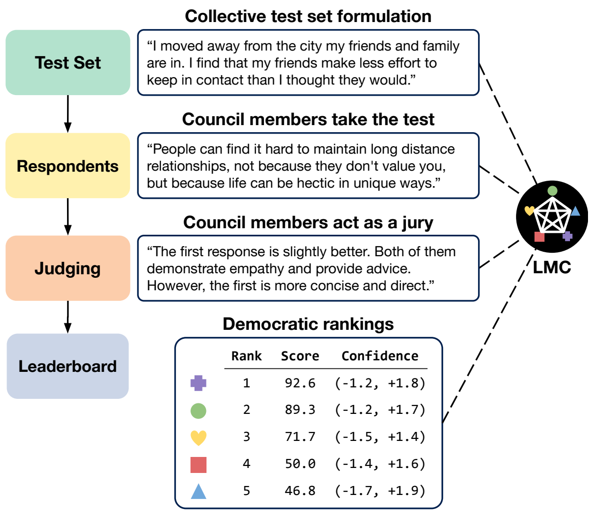


Figure 1: Overview of the Language Model Council (LMC) evaluation framework. Using the same LLMs for test set formulation, task completion, and judging offers an equitable way to achieve a ranking by consensus on highly subjective tasks.

1 Introduction

Despite recent advances in Large Language Models (LLMs), evaluating their outputs remains an open challenge. Human evaluations are challenging, time-consuming and expensive, motivating the search for automatic metrics (Novikova et al., 2017; Lowe et al., 2017) and evaluation setups (e.g. Zheng et al., 2024; Li et al., 2024b; Kocmi and Federmann, 2023; Shen et al., 2023; Kim et al., 2024). Conventional model evaluations use closed-ended questions that can be checked automatically such as MMLU (Hendrycks et al., 2020). These static benchmarks are at risk of contamination (Ravaut et al., 2024), prompting leaderboards to keep tests private or update questions periodically (e.g. Scale AI, 2024; White et al., 2024). However, performance on these questions is often misaligned with human preferences in real-world, open-ended contexts (Chiang et al., 2024). For such tasks, automatic metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or BLEURT (Sellam et al., 2020) are common. These require reference responses and fail to reflect human preferences beyond a quality threshold (Freitag et al., 2020). Arena-based methods enable reference-free evaluation by judging LLMs in head-to-head comparisons. The outcome of a battle between the responses of two LLMs can be evaluated based on objective outcomes (Bianchi et al., 2024), strong model judges like GPT-4 (Dubois et al., 2024b; Li et al., 2024a), or human judges (Li et al., 2024a), with each outcome contributing to metrics like win rates or ELO scores (Chiang et al., 2024).

Strong LLM judges like GPT-4 can match both controlled and crowdsourced human preferences well, achieving over 80% agreement, the same level of agreement between humans, in a general-domain, arena-based setting (Zheng et al., 2024). Unfortunately, it has also been observed that evaluator models tend to have their own biases, for example recognizing and preferring their own outputs over those of other models (Dubois et al., 2024a; Panickssery et al., 2024), leading to the development of committees of LLM judges (such as Verga et al., 2024; Zhao et al., 2024).

In a landscape of diverse LLM judges, how should we aggregate each model's vote and determine a definitive ranking among them? This paper aims to contribute a unique perspective in the pursuit of a fully decentralized approach to evaluating LLMs. When ranking LLMs on competencies that may be too subjective for a single model to judge fairly, we explore the benefits of a fully inclusive evaluation network where each model contributes equally.

Inspired by democratic organizations in human society, we introduce the **Language Model Council (LMC)**, a framework for collective evaluation. The LMC operates in three stages: **(1)** collaborative test set creation by council members, **(2)** response generation by each LLM, and **(3)** evaluation of responses by the council as a collective jury.

Our main contributions are as follows:

1. We propose the LMC, a flexible decentralized evaluation framework that uses LLMs to rank themselves in a democratic manner, and show that our method aligns with human rankings for an open-ended emotional intelligence task.
2. We define and analyze key measures of LLM judging dynamics including: separability, pairwise positional consistency, agreement, and affinity, for the largest ensemble of LLM judges to date (to our knowledge).
3. Using all LLMs for judging can be costly, so we use Monte Carlo simulations and hand-crafted sub-councils to discuss the value of larger evaluation ensembles and the value of the incremental judge.

2 Related Work

Recently, subjective tasks and label variability in humans has received increased attention in the NLP community (Alm, 2011; Basile, 2022; Plank, 2022). LLMs also exhibit variability in their outputs as they inherit inconsistencies and biases from human data (Hosking et al., 2023; Song et al., 2024; Bai et al., 2024a; Plaza-del-Arco et al., 2024; Koo et al., 2023). While LLM judges might not replicate human disagreement patterns exactly (Lee et al., 2023; Dong et al., 2024), their differences in judgment are increasingly viewed as reflecting valid dissent.[1]

LLM evaluation ensembles. Language-Model-as-an-Examiner (Bai et al., 2024b) uses LLMs to interact with candidates through follow-up fact-oriented queries in the knowledge domain. Auto Arena (Zhao et al., 2024) proposes an LLM committee to judge competitive multi-turn interactions between LLMs. PRD (Li et al., 2023) allows LLMs to discuss evaluations and assigns higher voting weights based on ability. PRE (Chu et al., 2024) selects a small group of reviewers to produce individual evaluations, then aggregates these evaluations through a chair model. DRPE (Wu et al., 2023) uses multi-roleplayer prompting to simulate different roles with one base model, integrating them as votes for the final results. PoLL (Verga et al., 2024) enhances cost-effectiveness by replacing one large judge with multiple smaller judges.

We build upon existing research by 1) focusing on a highly subjective case study on emotional intelligence where human agreement is inherently low, 2) emphasizing full inclusivity, where each LLM plays an equal role in determining the final rankings, and 3) engaging a large ensemble of diverse LLMs to study judging dynamics in greater depth. To our knowledge, this is the largest ensemble of LLM judges studied to date.

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

The LMC framework consists of three stages: 1) test set formulation, 2) response gathering, and 3) collective judging (Figure 1). These steps are run sequentially and are automated with LLM-based agents.[2] See Appendix I for all prompts.

[1] Notably, when humans disagreed with GPT-4, they considered its judgments reasonable in 75% of cases and sometimes revised their own answers (Zheng et al., 2024).

[2] We use "agents" to refer to LLMs acting simultaneously on sequential and systematic evaluation procedures.

3.1 Council member selection

Our selection of LLM council members 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 (Hendrycks et al., 2020) and Chatbot Arena (Chiang et al., 2024). We ensure a broad variety of LLMs by including models from *eight* different organizations and *four* countries, with a mix of open-source and closed-source models, small and large (Table 3 in Appendix A).

3.2 Test set formulation

To create a compelling open-ended test set for Emotional Application (EA), we build upon the EmoBench dataset, a publicly available, handcrafted, theory-based English dataset designed for EA assessment (Sabour et al., 2024). EmoBench consists of 200 emotionally balanced, handcrafted scenarios, 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 LMC to transform the concise, closed-ended scenarios in EmoBench into richly described, open-ended dilemmas in the first person (Figure 26). Each of the 20 council members expands five scenarios, resulting in a dataset of 100 dilemmas, similar in scale to MT-Bench (80 questions). We manually review all expansions for EA suitability.[3]

Relying on a single LLM to generate the entire test set, even a top performer like GPT-4o may introduce bias and limit perspectives. In a survey of 10 human respondents, 51% of preferred expansions were not written by GPT-4o.[4] Inclusively constructing test examples also mitigates the risk of any single LLM's generative idiosyncrasies (AI4Science and Quantum, 2023) from dominating the test pool.

3.3 Response gathering

After generating 100 dilemmas, each council member responds to every dilemma, yielding 2,000 total responses. To standardize response lengths across council members and minimize length bias in evaluation, the prompt suggests a 250-word limit (Figure 27). Responses exceeding the limit are truncated at the nearest sentence within the limit.[5] Despite the suggested word limit, some council members consistently generated shorter responses (Figure 1). These responses are left unchanged.

3.4 LLM-as-a-Judge evaluation settings

Arena-style pairwise comparisons with a single reference model. We adopt the pairwise comparison setup of Chatbot Arena where responses are compared in a pairwise fashion in a series of battles (Figure 29). LLM rankings are based on expected win rates derived from an ELO scoring system (Bai et al., 2022; Boubdir et al., 2023). We also use Bradley-Terry (BT) coefficients (Bradley and Terry, 1952) for better statistical estimation. We use a single reference model, Qwen-1.5-32B, for all pairwise battles. For details on how the reference model was chosen, refer to Appendix D. Following (Li et al., 2024a; Chiang et al., 2024), we derive confidence intervals with 100 rounds of bootstrapping.

4-point preference scale. We query all LLMs with a temperature of 0 and with granular comparison options without ties (A>>B, A>B, B>A, B>>A). The reasoning behind these choices is detailed in Appendix C. We use Chain-of-Thought (CoT) prompting (Wei et al., 2022) to generate discussion before giving judgments.

Exhaustive position-swapping. To minimize position bias from affecting the final ranking, 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 BT coefficient calculation in the original codebase[6], inconsistent results after swapping are treated as ties and strong votes are counted as 3 separate wins.

Voting aggregations. We consider 3 different voting aggregations for consolidating scores across multiple LLM judges on a per-battle basis: *majority vote* (mode of all votes), *mean pooling*[7], and

[3] An alternative procedure is to have all council members suggest expansions for each scenario and select the best through voting. This is feasible, but given the generally high quality of expansions, the added value is limited compared to simply adding new scenarios.

[4] Human respondents were asked to choose, in a series of pairwise comparisons, which expanded dilemma would be better for an emotional intelligence test. Each comparison was between a response from GPT-4 and one from a randomly chosen council member. See Appendix E for more details.

[5] Sentence splits are based on standard English end punctuation (. ! ?), as the experiment is conducted in English.

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

[7] For mean pooling, we map ratings to a 4-point numeric scale (A>>B: 2, A>B: 1, B>>A: -1, B>>A: -2), take the mean rounded to the nearest whole value, and use the value corresponding to that whole number as the final rating.

no aggregation (judgments across all battles are equally considered).

3.5 Characterizing LLM judges

We leverage the many-to-many interactions between LLMs to define *key judging qualities* of LLMs in an ensemble setting.

Separability measures how confidently models can be distinguished in the final rankings. We adopt this metric from (Li et al., 2024a), which calculates separability as the percentage of model pairs with non-overlapping confidence intervals. Higher separability means better differentiation between models. The success of score separation depends on three factors: judge discrimination, test discrimination, and participant abilities. They all contribute, and any one of them could reduce the separation to 0. In the LMC framework, potentially all three factors are governed by stochastic LLMs that are non-deterministic even with temperature=0 (Chann, 2023). This makes it challenging to pinpoint which component would provide the most improvement if separability is low. Our analysis of separability focuses on LLM judging, as the test set, participants, and their responses remain constant across different LMC configurations.

Pairwise **P**ositional **C**onsistency (**PPC**) measures how often a judge gives consistent results when the order of the two model responses in a pairwise comparison is swapped (Appendix C.2). For instance, if the judge ranks A>B in one comparison and B>A when the positions are swapped (a *battle couplet*), the judge is considered "consistent" because the preference remained the same independent of position. **Position bias** measures how much the LLM judge favors a specific position (either the first or second) and we define it to be $1 - ppc$. On the 4-point preference scale, a battle couplet is still considered consistent as long as the relative ranking remains consistent overall — fine-grained differences such as (A>>B, B>A) or (B>>A, A>B) are tolerated. Table 11 lists all possible couplets and their consistency mappings. **Conviction** is the percentage of strong votes (A>>B or B>>A).

Affinity between a judge and respondent is the score 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 score for that model. **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.

Agreement is measured using Cohen's Kappa (Cohen, 1960) between two judges' ratings. Similar to position bias, we consider judges in agreement as long as they express the same relative preference, e.g. (A>B and A>>B) or (B>A and B>>) are in agreement. **Contrarianism** is measured as the disagreement between an LLM and the Council's majority decision, reported as $1 - \kappa$.[8]

3.6 Human study

To validate the LMC's evaluations, we conduct a human study mirroring that of the LMC: human raters are asked to choose the best of two responses to a given dilemma. Note that the goal is to replicate human preferences in terms of general ranking, not to model the distribution of preferences.

Due to budget constraints, we select nine LLM council members from our pool of 20 to be rated for this study (listed in Figure 3) with variety in terms of model size, open- versus closed-source, and company origin in mind. We recruit participants via crowdsourcing on Prolific.[9] We recruited a total of 102 participants. Each 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 E.

4 Results and Findings

Table 1 shows the main results of our Council-based EI study, highlighting the following insights:

Qwen-110B outranks GPT-4o in an unexpected upset. Like other benchmarks, larger models within the same family tend to outrank their smaller or older versions. However, unlike other benchmarks, Qwen-1.5-110B (#20 on Chatbot Arena) scores highest on our EI task, followed by GPT-4o (#1 on Chatbot Arena)[10]. This is surprising, as Qwen-1.5-110B does not typically outperform GPT-4o. One possible reason for this outcome is the use of Qwen-1.5-32B as the reference model (Appendix D). Because all of Qwen-1.5-110B's responses are compared to the responses of a strictly smaller variant of the same family, this could result in an outsized advantage for Qwen-1.5-110B in the evaluation overall. This raises an interesting impli-

[8] Cohen's κ ranges from -1 to 1. Subtracting from 1 is not particularly interpretable. Applying the negative makes it so that a higher score implies more disagreement and vice versa.

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

[10] This was the ranking as of May 2024. The Qwen-1.5 models have been dropped from Chatbot Area since the Qwen-2.5 family of models have been added.

	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%*
LMC (majority vote)				73.7%	**75.3%**
LMC (mean pooling)				74.7%	68.5%
LMC (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 equally, without aggregation or modification. Under various aggregation algorithms, the council is more separable and more consistent than individual LLM judges.

cation for Arena-style evaluations, as it's possible that the choice of reference model may introduce a bias favoring its successors in the same arena.

Judges prioritize actionability, clarity, and structure when expressing preferences. Using chain-of-thought (CoT) prompting, judges provided detailed reasoning for their preferences. We analyzed 1,000 reasoning traces and identified common themes (Appendix G).

Judges disfavor models that produce responses significantly shorter than the suggested word limit. Despite a suggested 250-word limit, some models generated much shorter responses, even though decoding parameters allowed for longer output. Models that adhered to the limit, using 220+ words on average, performed better, while models in the bottom four positions averaged less than 200 words. Notably, Gemini-1.5-pro placed last with an average response length of just 115 words, far worse than its predecessor Gemini-1.0-pro (4th place) with 228 words on average. LLM Judges bias towards longer responses, but excluding the outliers (models that went well under the limit) made length bias insignificant (Table 6).

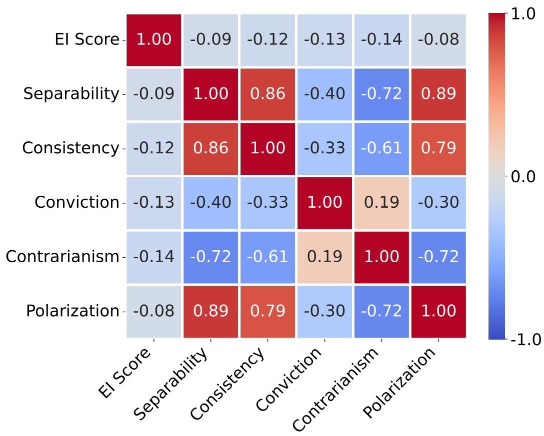


Figure 2: Spearman correlation between EI score and key judging qualities across 20 LLM council members.

LLM success in the EI task does not correlate with its judging ability. Performance on the EI task had only a weak correlation with any of the key judging qualities (Figure 2), which suggests that the ability to perform well in the task and the ability to judge others' responses are distinct skills.

Consistent judging, neutral voting patterns, and lower contrarianism lead to higher separability. Our analysis shows that judges with more consistent votes and neutral voting patterns tend to achieve higher separability. Judges with high conviction — those that express strong preferences (A>>B or B>>A) frequently — are negatively correlated with separability, suggesting that overly strong votes tend to introduce noise. For example, claude-3-opus, despite expressing strong preferences only 0.1% of the time, achieves the second-highest separability (72.6%). Higher contrarianism, which measures how often a judge disagrees with the majority decision, also correlates with lower separability. Qwen1.5-72B, the most contrarian judge, has only 38.9% separability.

LLMs are lightly self-biased, but it is mitigated in the full Council. Out of the 20 LLMs in our study, 12 exhibited positive self-enhancement bias, meaning they rated themselves more favorably than the Council's overall score for them. When self-graded battles are excluded, the overall rankings remain similar (Figure 8). This indicates that while individual models carry self-enhancement bias, the ensembling of the LMC effectively neutralizes these biases.

Agreement among LMC members, humans, and between the LMC and humans is similar. Figure 3 shows that the rankings produced by LMC members and humans are consistent, with small variations. Both groups agree on the top-performing models and the lowest-ranked ones. The level of agreement between the LMC and humans is roughly the same as the agreement within human evaluators (51.9%).

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

Table 2: 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.

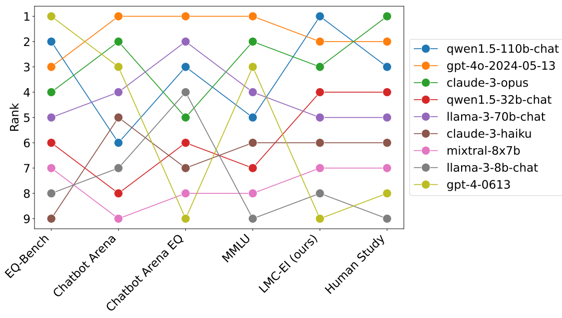


Figure 3: LLM rankings from different benchmarks.

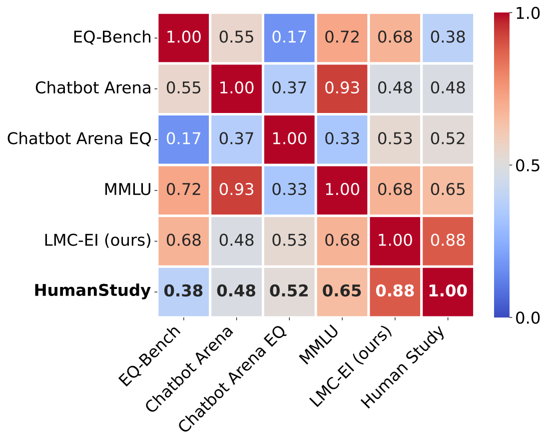


Figure 4: Spearman correlations for different evaluation benchmark scores for 9 LLMs.

The LMC's rankings align more closely with human evaluations than other benchmarks or individual judges. The LMC achieves the highest correlation with human-established rankings, outperforming other benchmarks, including those from similar domains like the EI-specific subset of Chatbot Arena (Appendix F) and EQ-Bench (Paech, 2023) (Figure 4). The LMC outperforms most individual LLM judges in terms of correlation with human rankings. While a few individual LLM judges, like dbrx-instruct, achieve similar correlation scores (e.g., 0.917), they do so with significantly lower separability (50.5% vs. 90.5% for the LMC). This suggests that the LMC's collective judgment not only aligns more closely with human preferences but also provides clearer distinctions between model performance, making it a more reliable method for ranking models in our case study.

Additional findings. The 20x20 LLM interactions generate a wealth of data, some of which is not included in the main paper for brevity. Additional insights can be found in Appendix B.

5 Discussion

One of the most important questions about using a LMC is whether it is worth the trouble. Collecting more opinions costs more energy, and parsing, tallying, and storing them add logistical costs. What is the value of another opinion, much less everyone's opinion?

From a cost perspective, a closely related question is how many examples we should collect judgements for to begin with. Using a smaller test set simplifies stakes for human reviewers — it is far easier to read and remember 10 examples than 10,000 – while also reducing costs.

If we assume that the main goal is to establish good relative rankings, we narrow the focus of quantifying the success of a council by measuring the significance and stability of the final ranking. For significance, we look at separability (as in the main experiment), which measures how well models can be distinguished based on non-overlapping confidence intervals. For stability, we create a new metric, **M**ean **E**xpected **R**ank **V**ariance (**MERV**), defined as the expected ordinal swing of the average respondent's rank (Appendix H). A MERV of 3 means an average respondent's rank is expected to change up to 3 positions in a new trial. Lower MERV indicates that the benchmark has more stable rankings, with MERV of 0 signifying perfect deterministic-like stability.

5.1 Monte Carlo simulations

To study the dynamics of separability and stability with differently-sized councils and test sets irrespective of the inclusion of any specific LLM, we use a Monte Carlo procedure to simulate many random hypothetical councils and test sets.

Our Monte Carlo simulation procedure is as follows: 1) For a given council size c and test set size t, randomly sample c LLMs to form a council C and t examples to form a test set T. Sampling is performed with replacement. 2) Find the associated judgements from the main experiment for the specific (C, T) configuration to determine the scores and relative rankings for all LLMs. 3) Repeat for 100 trials[11]. 4) After all 100 trials are complete, tally the results: for stability, observe fluctuations in rankings for each LLM to compute MERV, and for significance, report the mean separability. Figure 5 shows results for a sweep $c \in \{1, 3, 5, \ldots, 19\}, \quad t \in \{10, 20, 30, \ldots, 100\}$.

A nuanced trade-off between the size of the test set and the number of judges. Both separability and stability improve as the number of test examples and the number of LLM judges increase, with the best scores achieved when both are maximized. However, based on the gradient maps for MERV (Figure 5b) and separability (Figure 5d), the added benefit of including either an additional judge or more test examples diminishes significantly in a concentric shape that starts 50 examples and 9 judges. The gradients in this darker zone indicates where the utility of any additional opinion—whether in the form of new test data or another LLM judge—becomes marginal.

Larger councils are more robust to adversarial judges, with diminishing marginal utility. As the size of the LMC grows to include many different LLMs, it may become difficult to verify the quality of every member, particularly on subjective tasks. In our Monte Carlo procedure, we experiment with a simulation setting with adversarial judges. An adversarial judge is a fake LLM council member that returns ratings at random[12]. In Figure 6, we find that on both separability and stability, larger councils reduce the negative impact of adversarial LLM judges. We find that this robustness continues to increase as the size of the council grows, even when maintaining the same ratio of adversarial judges to real judges, albeit with diminishing marginal utility.

What is the value of the incremental judge? With respect to stability and separability, it depends. For small test sets, the value of additional test examples outweighs the benefit of more judges, but when test data is abundant test data, the value of an additional judge is higher. If there are more than 20-30 examples in the test set, adding an additional judge provides greater value, but for less, it is better to get opinions on an additional batch of examples with the current set of judge(s).

Overall, larger councils are more robust to adversarial judges so the bigger your council is, the less concerned you need to be about a particularly bad judge from wreaking havoc on your study, and you don't have to be as picky about your council members.

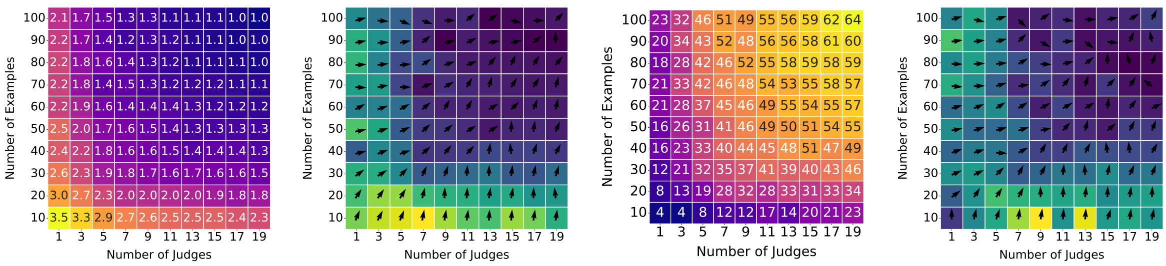
5.2 Oligarchical councils

If a task does benefit from multiple perspectives and the evaluation budget allows for a limited number of opinions, *whose opinions should we include?* (Verga et al., 2024) suggests replacing a single large judge with 3 smaller judges.[13] We explore a sim-

[11] We use 100 to be consistent with the number of rounds of bootstrapping in the main experiment. For calculating separability in Monte Carlo simulations, the trials themselves are used to bootstrap confidence intervals directly.

[12] There are other adversarial algorithms that can be used such as always voting for the first position that we did not explore.

[13] Council composition is likely a task-specific optimization; On HotpotQA, an ensemble of LLM judges was worse than using Claude Haiku alone (Verga et al., 2024).



(a) MERV (b) MERV gradients (c) Separability μ (d) Separability μ gradients

Figure 5: Measurements of rank stability (MERV) ((a) and (b)) and separability ((c) and (d)) averaged over 100 randomized trials for various numbers of judges and examples. (a) and (c) display raw metric values while (b) and (d) display the gradient magnitude (colors) and direction (arrows). The gradient calculation follows a Manhattan distance approach where row-wise and column-wise gradients are linearly combined to reflect the discreteness of changes between adjacent squares, highlighting the incremental impact of adding another judge or more examples.

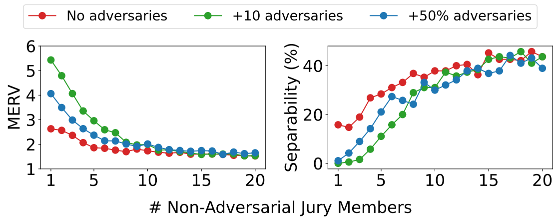


Figure 6: MERV and separability for Monte Carlo simulations for (t=30) with adversarial judges.

ilar but reverse question: can a subset of judges (or a single judge) effectively represent the fully democratic LMC?

We compare the rankings of three hand-curated sub-councils (Table 4): *flagships*, *smalls*, and *top-4*. We refer to the sub-council as an "oligarchical" council since we rely on a small subset of participating LLM judges to determine the ranking of every LLM, akin to oligarchies in human society.

Our analysis finds that full council participation is neither necessary to achieve rankings that align with human rankings nor maintain good separability. *Smalls*, a council composed of the smallest LLMs, achieves a respectable separability of 71%, which is already better than the average judge (53.3%). It also achieves a Spearman correlation of 0.88 with human rankings, only slightly lower than the full council's correlation of 0.92 (Figure 22).

Still, it's worthwhile to be thoughtful about council composition. For instance, *top-4* achieves the same correlation with human evaluations as *smallest* (0.88), but with notably higher separability (79%). While smaller councils consisting of smaller LLMs can perform reasonably well, there are other factors to consider like separability, robustness, human alignment, and bias mitigation.

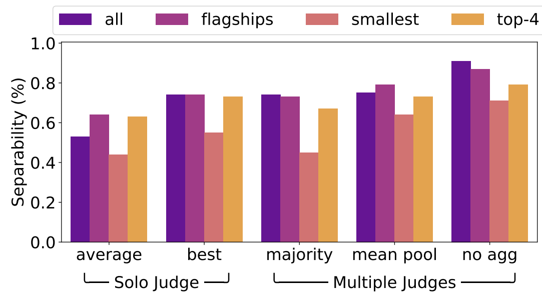


Figure 7: Separability scores achieved by different council compositions and aggregation methods. Higher separability is better.

6 Conclusion

In this paper, we introduce the Language Model Council (LMC), a flexible, decentralized evaluation framework for ranking LLM agents through democratic participation. Applying the LMC to an emotional intelligence task with 20 LLMs, we demonstrate that the LMC can produce highly separable rankings that align more closely with human judgments than other benchmarks or individual judges. Through both Monte Carlo simulations and hand-curated sub-councils, we find that while larger councils provide benefits, they are incrementally diminishing, and the majority of key qualities—such as ranking significance, stability, and alignment with human evaluations—can be achieved with a smaller, well-chosen ensemble of judges. As humans increasingly rely on LLMs to evaluate other LLMs, we hope the LMC framework, along with insights from our case study, offers a valuable foundation for developing reliable LLM evaluations, even for highly subjective tasks.

Limitations

Generalizability of the LMC. Although we present one detailed case study focused on EI, the Language Model Council (LMC) framework is broadly applicable to a wide range of open-ended tasks. The core mechanism of tallying preferences through arena-style pairwise comparisons is inherently adaptable to various types of prompts and tasks (Chiang et al., 2024). However, framing new subjective tasks, such as those related to aesthetics or politics, in a form suitable for technical evaluation still requires careful design. In our case study, we were responsible for the technical formulation of the EI task, and the task examples were seeded from EmoBench (Sabour et al., 2024), a human-crafted dataset that we chose.

The least generalizable aspect of the LMC is likely the first step: formulating the test set in a collaborative way. For tasks with fixed or human-authored test sets, it may be undesirable or unclear how to implement participation from multiple LLMs. In such cases, this step could be omitted or delegated to a single strong LLM. Whether any LLM can generate meaningful test sets for narrow tasks in a fully unsupervised, domain-generic manner—subjective or otherwise—remains an open area of exploration (Lhoest, 2024).

Single-turn interactions and English-only evaluation. Our case study evaluates EI based on single, self-contained interactions and was conducted entirely in English. However, many tasks may be better assessed through extended conversations, multiple sessions, multiple modalities, or in multiple languages, to reflect a broader range of human interaction dynamics, all of which are not explored in this paper.

Reproducibility challenges. LLMs are inherently stochastic, meaning the same model can produce different ratings even with temperature set to 0 (Chann, 2023). Reproducibility is further complicated by closed-weight models like GPT-4, which may receive undisclosed updates. All responses in our study were collected in May 2024, but serverless providers, like Together, may update their APIs or discontinue support for certain models, as happened with Qwen-1.5-32B (replaced by Qwen-2.5). For open source models, changes in deployment hardware or GPU configuration can introduce additional variability, making exact replication of results difficult.

LLM statelessness. We assume that LLMs, being memory-less, can serve as both respondents and judges simultaneously. In contrast, humans would struggle to judge their own responses impartially due to memory retention. As LLMs evolve to incorporate memory—such as retaining recent prompts and responses (OpenAI, 2024c)—the risk of self-enhancement bias may increase. If LLMs begin to remember their own responses during evaluation, disabling self-grading may become the default approach to ensure fairness.

Inclusive democracy does not guarantee fairness. While the LMC mitigates self-enhancement and position bias, it remains vulnerable to systemic biases within the evaluation framework itself. For instance, we speculate that using a single reference model may introduce an advantage for successors of the reference model in the arena. This specific aspect could be mitigated by using multiple reference models, but it can be challenging for new test data to compensate for a bias in fundamental evaluation procedures.

Diversity of opinions in LLMs versus humans. The question of whether the distribution of an ensemble of LLM judgments is a good approximation for general human diversity (Dong et al., 2024; Hosking et al., 2023) is secondary to our focus on the alignment of *final rankings*, which is the aggregated expression of the collection of abundantly dissenting opinions (from humans or LLMs), and thus where we assert the utility of the LMC's methodology. In our EI case study, we found that the LMC's final ranking aligned *better* with human preferences compared to other benchmarks and the rankings produced by nearly all individual LLM judges.

We also recognize that the human judgments collected in our study may not fully represent the diversity of opinions within the broader population (Elangovan et al., 2024), nor do they reflect authentic, first-hand judgments. Emotional responses are highly individual, and real interpersonal conflicts are shaped by personal experiences and social factors that may be difficult for anyone other than the person experiencing the conflict to fully evaluate. For both humans and LLMs, we can only make a deliberate effort to gather judgments from some accessible variety of relevant profiles. This approach mirrors the rationale behind the design of the LMC in our case study, which was similarly formed with variety in mind.

Acknowledgements

Justin led the research. Amanda and Flor led the human study. Benji ran Monte Carlo simulaitons and generated gradient charts. 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

Mistral AI. 2024. Au Large — mistral.ai. https://mistral.ai/news/mistral-large/. [Accessed 15-10-2024].

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

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

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

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

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

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

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

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo Uncovered: Robustness and Best Practices in Language Model Evaluation. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 339–352, Singapore. Association for Computational Linguistics.

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

Sherman Chann. 2023. Non-determinism in GPT-4 is caused by Sparse MoE — 152334h.github.io. https://152334h.github.io/blog/non-determinism-in-gpt-4/. [Accessed 12-10-2024].

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. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

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

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

Cohere. 2024a. Command R: RAG at Production Scale — cohere.com. https://cohere.com/blog/command-r. [Accessed 15-10-2024].

Cohere. 2024b. Introducing Command R+: A Scalable LLM Built for Business — cohere.com. https://cohere.com/blog/command-r-plus-microsoft-azure. [Accessed 15-10-2024].

Databricks. 2024. Introducing DBRX: A New State-of-the-Art Open LLM — databricks.com. https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm. [Accessed 15-10-2024].

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge?

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik

Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

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

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

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Bodapati, and Dan Roth. 2024. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Google. 2024a. Introducing Gemini: our largest and most capable AI model — blog.google. `https://blog.google/technology/ai/google-gemini-ai/`. [Accessed 15-10-2024].

Google. 2024b. Our next-generation model: Gemini 1.5 — blog.google. `https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/`. [Accessed 15-10-2024].

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

Tom Hosking, Phil Blunsom, and Max Bartolo. 2023. Human feedback is not gold standard.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

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

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

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

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices?

Quentin Lhoest. 2024. Infinite dataset hub. Accessed: 2024-10-11.

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

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

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024b. Leveraging large language models for nlg evaluation: Advances and challenges.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Meta. 2024. Introducing Llama 3.1: Our most capable models to date — ai.meta.com. `https://ai.meta.com/blog/meta-llama-3-1/`. [Accessed 16-10-2024].

Mistral. 2024. Cheaper, Better, Faster, Stronger — mistral.ai. `https://mistral.ai/news/mixtral-8x22b/`. [Accessed 15-10-2024].

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

OpenAI. 2023. `https://platform.openai.com/docs/models/gpt-3-5-turbo`. [Accessed 15-10-2024].

OpenAI. 2024a. `https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4`. [Accessed 15-10-2024].

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

OpenAI. 2024c. Memory and new controls for chatgpt. `https://openai.com/index/memory-and-new-controls-for-chatgpt/`. [Accessed 14-10-2024].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C J Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 technical report.

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Flor Miriam Plaza-del-Arco, Amanda Curry, Alba Cercas Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Scale AI. 2024. Scale ai leaderboard. `https://scale.com/leaderboard`. Accessed: 2024-10-10.

Matthias Schmidmaier, Jonathan Rupp, Darina Cvetanova, and Sven Mayer. 2024. 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. Association for Computing Machinery.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

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

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

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

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

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2024. Livebench: A challenging, contamination-free llm benchmark.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. 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.

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

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

A Language Model Council Membership

Country	Organization	LLM	Release Date	Chat Arena Elo	MMLU (5-shot)	Size	License
United States	Open AI	gpt-4o-2024-05-13 (OpenAI, 2024b)	05/24	1287	88.7	🔒	Proprietary
United States	Open AI	gpt-4-turbo-04-09 (OpenAI, 2024a)	04/24	1256	🔒	🔒	Proprietary
United States	Open AI	gpt-4-0613 (OpenAI et al., 2023)	06/23	1246	86.4	🔒	Proprietary
United States	Open AI	gpt-3.5-turbo-0125 (OpenAI, 2023)	01/24	1102	70.0	🔒	Proprietary
France	Mistral	mistral-large-latest (AI, 2024)	02/24	1156	81.2	🔒	Proprietary
France	Mistral	open-mixtral-8x22b (Mistral, 2024)	04/24	1146	77.8	176 B	Apache 2.0
France	Mistral	open-mixtral-8x7b (Jiang et al., 2024)	12/23	1114	70.6	56 B	Apache 2.0
United States	Meta	llama-3-70b-chat-hf (Dubey et al., 2024)	04/24	1208	82.0	70 B	Llama 3 Community
United States	Meta	llama-3-8b-chat-hf (Dubey et al., 2024)	04/24	1153	68.4	8 B	Llama 3 Community
United States	Google	gemini-1.5-pro-preview-0409 (Google, 2024b)	05/24	1268	81.9	🔒	Proprietary
United States	Google	gemini-1.0-pro (Google, 2024a)	04/24	1208	71.8	🔒	Proprietary
United States	Databricks	dbrx (Databricks, 2024)	03/24	1103	73.7	132 B	DBRX LICENSE
Canada	Cohere	command-r-plus (Cohere, 2024b)	04/24	1189	75.7	104 B	CC-BY-NC-4.0
Canada	Cohere	command-r (Cohere, 2024a)	04/24	1147	68.2	35 B	CC-BY-NC-4.0
United States	Anthropic	claude-3-opus-20240229 (Anthropic, 2024)	03/24	1248	86.8	🔒	Proprietary
United States	Anthropic	claude-3-sonnet-20240229 (Anthropic, 2024)	03/24	1201	79.0	🔒	Proprietary
United States	Anthropic	claude-3-haiku-20240307 (Anthropic, 2024)	03/24	1178	75.2	🔒	Proprietary
China	Alibaba	qwen1.5-110B-chat (Bai et al., 2023)	02/24	1164	80.2	100 B	Qianwen LICENSE
China	Alibaba	qwen1.5-72B-chat (Bai et al., 2023)	02/24	1152	77.4	72 B	Qianwen LICENSE
China	Alibaba	qwen1.5-32B-chat (Bai et al., 2023)	02/24	1126	74.3	32 B	Qianwen LICENSE

Table 3: 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 4: Additional council variations consisting of a hand-picked subset 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 as of May 2024.

B Additional Findings

Can we proactively remove inconsistent votes?

Table 1 shows a wide variety of judging qualities within the council, especially in consistency and separability. In the two-game setup, we gather ratings for models in both positions during pairwise comparisons, allowing us to identify and potentially remove inconsistent ratings before calculating ELO scores.

Automatically removing inconsistent votes affects the weighting of LLM judges, as those with more inconsistent ratings will have fewer votes counted and thus less influence overall. However, this could be a positive outcome, as positional consistency is often a sign that the judgment was arbitrary, especially in cases where tie ratings are not permitted. By comparison, some arena-based evaluation systems like (Zheng et al., 2024) allow judges to declare a tie between responses, however these battles are completely excluded from ELO scoring.

In our emotional intelligence (EI) case study, we chose to retain all inconsistent votes, allowing downstream processes like aggregation, Bradley-Terry scoring, and bootstrapping to implement any diminished influence caused by inconsistency.

For the hand-selected sub-councils analyzed in Section 5.2, Table 5 shows the changes in key judging qualities when considering only consistent votes. Table 6 shows the impact of inconsistent vote pre-filtering on length bias.

Extended judging profiles and references to larger visualizations.

Due to table size constraints, we have divided the full tables of key judging qualities into several smaller tables. The full-sized 20x20 heatmaps from our LMC EI case study also span multiple pages. For ease of reference, all large tables and figures are compiled here:

- Table 8 shows measures of bias for individual judges and the LMC as a whole.
- Table 9 shows measures of agreement for individual LLM judges.
- Table 10 shows polarization and affinity for individual LLM judges.
- Figure 9 shows a heatmap of the affinities between LLM judges and LLM respondents.
- Figure 10 shows a heatmap of the normalized affinities between LLM judges and LLM respondents (the LMC's consensus affinity subtracted out).
- Figure 11 shows a graph consisting of each LLM judge's top 5 affinities.
- Figure 12 shows a heatmap of Cohen's κ sidewise agreement scores.
- Figure 13 shows a graph consisting of each LLM judge's top 5 most agreeable other LLMs.
- Figure 14 shows a heatmap of the estimated LLM vs. LLM win rates.

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: Key judging qualities from councils (no aggregation) for hand-selected sub-councils (Table 3) with only positionally consistent votes (positionally inconsistent ratings are filtered out prior to ranking). The value in parentheses represents the change compared to using all votes without a first filtering step.

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

Table 6: Length bias with and without models with responses <200 words.

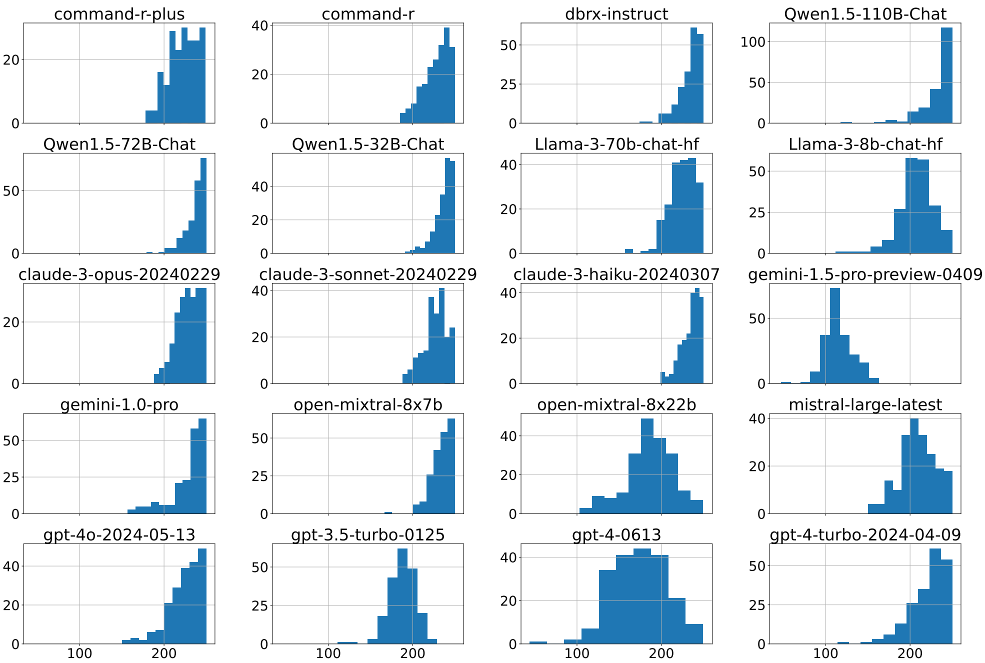


Table 7: Distribution of response lengths for 20 LLMs on our EI task, measured in number of tokens.

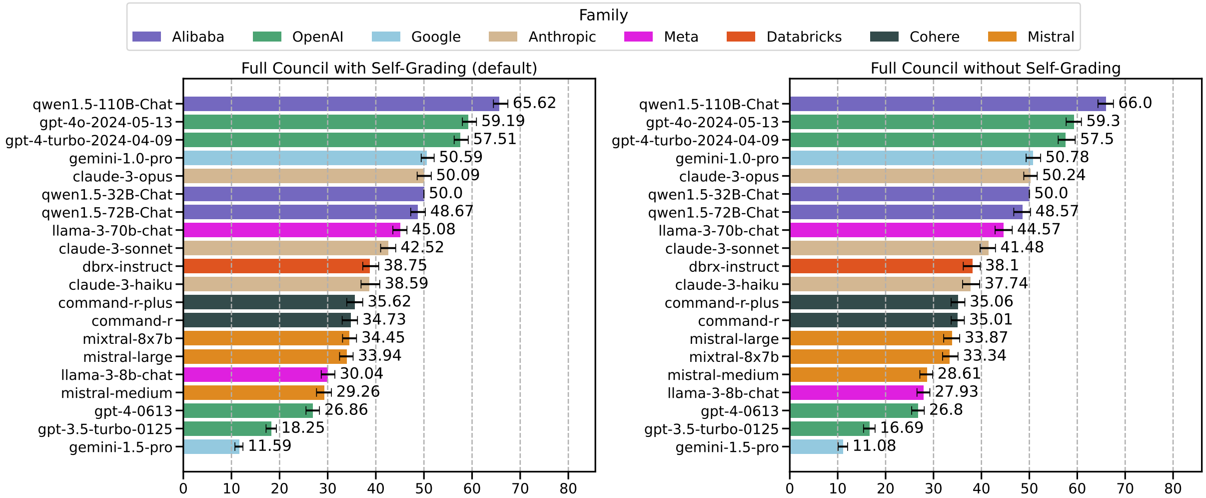


Figure 8: Comparison of full LMC rankings (no aggregation), with self-grading (left, default) permitted and with self-grading disabled (right). The rankings are broadly identical, confirming that the ensemble of LLM judges of the Language Model Council mitigates self-enhancement bias.

	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 8: LMC judging profile relates for bias, with and without consistent votes.

	All votes			Consistent votes		
LLM	**Contrarianism**	**Agrees most with**	**Disagrees most with**	**Contrarianism**	**Agrees most with**	**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 9: LMC judging profiles related to agreement.

	All votes			Consistent votes		
LLM	**Polarization**	**Lowest affinity for**	**Highest affinity for**	**Polarization**	**Lowest affinity for**	**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 10: LMC judging profiles related to affinity.

figures/raw .pdf

Figure 9: Heatmap of the affinities of LLM judges to LLM respondents. The relatively consistent horizontal bands in the heatmap suggest a clear consensus on the preferred LLM participants. Some judges, like mistral-large, exhibit high polarization, with a significant difference between their highest and lowest-rated LLMs. In contrast, judges like llama3-8b display a narrower range of affinity. The highest affinity expressed by any LLM comes from mistral-large for qwen1.5-110B, while the strong horizontal blue band for gemini-1.5-pro indicates it was a consensus low performer.

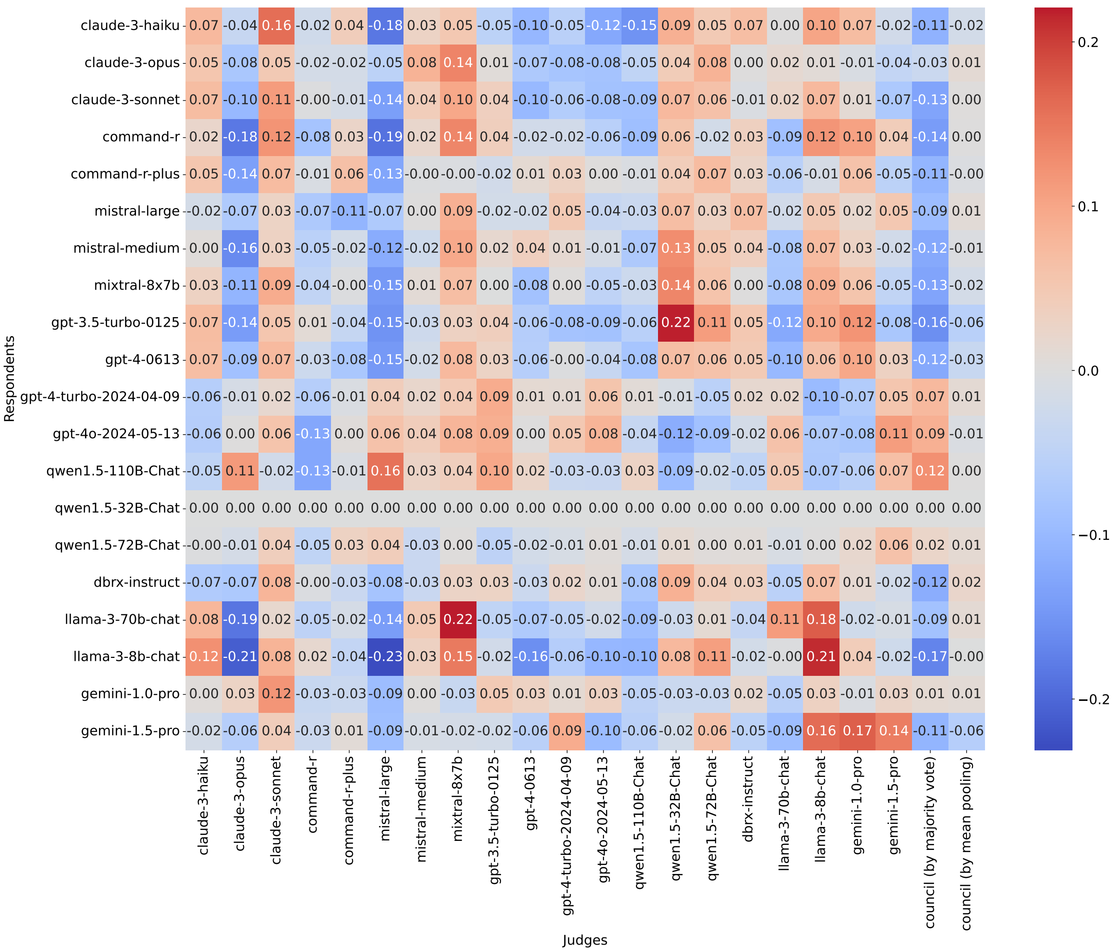


Figure 10: This heatmap shows the normalized affinities of LLM judges toward LLM respondents, with the LMC's consensus affinity subtracted out. Self-enhancement bias is now visible along the diagonal, where most LLMs exhibit some level of bias—though not all. Interestingly, six LLMs, including Claude-3-Opus and mistral-large, display negative self-enhancement bias, rating their own responses lower than the council's consensus. For instance, mixtral-8x7b shows a strong preference for llama3-70b's responses (+0.22 points above the consensus), but this affection is not reciprocated—llama3-70b actually rates mixtral-8x7b -0.08 points below the LMC's consensus. Mistral-large is a particularly critical judge, rating 15 out of 20 LLMs more harshly than the LMC's consensus. In contrast, Claude-3-Sonnet is much more favorable, expressing negative affinity for only one LLM, qwen-1.5-110B. The family blocks along the diagonal also reveal patterns of family-enhancing or self-deprecation bias. The llama3 family shows the highest family-enhancing bias, while the mistral family is more divided. Mistral-large and mistral-medium disproportionately rate their fellow family members, including themselves, more negatively, whereas mixtral-8x7b shows positive family bias.

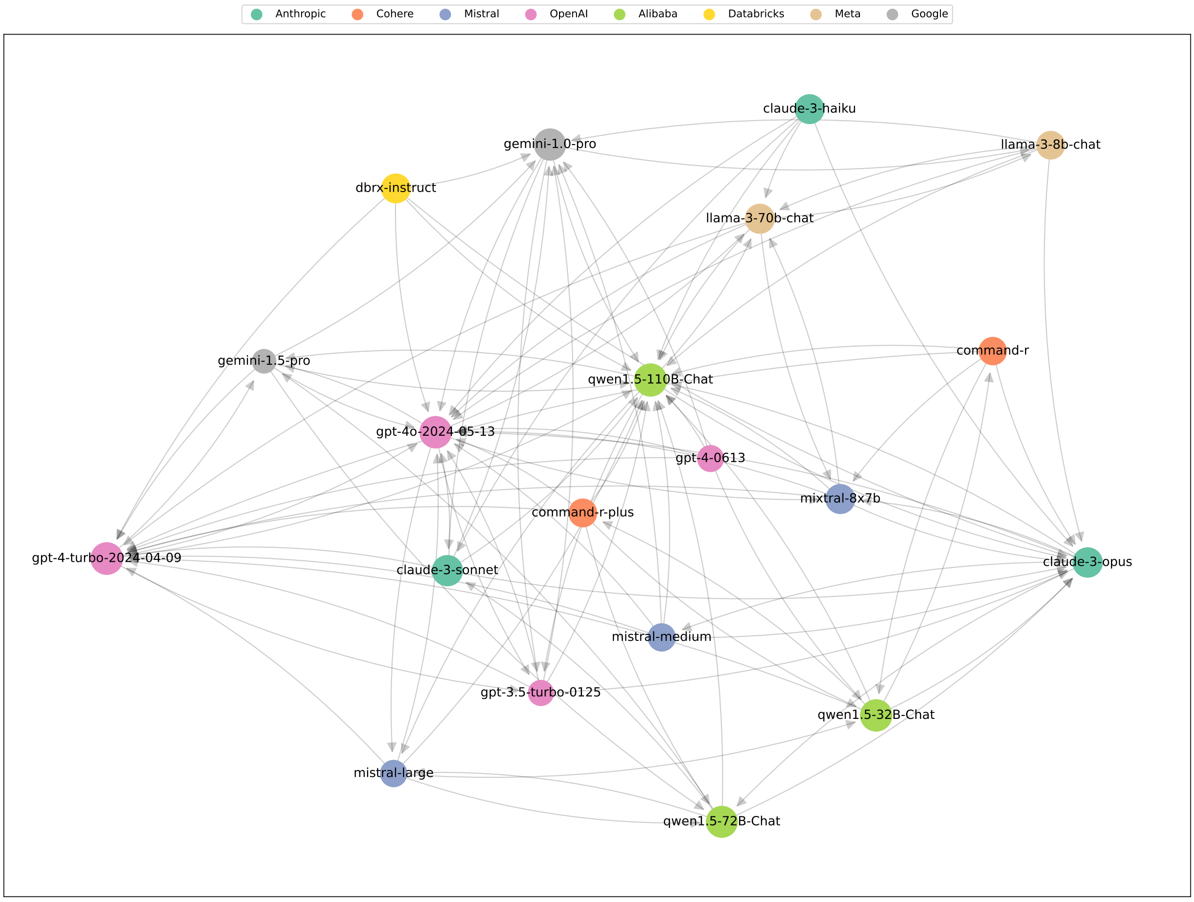


Figure 11: Graph of top 5 affinities. An edge exists from LLM a to LLM b if affinity(a, b) is in the top 5 affinities for LLM a. This view allows us to identify popular LLMs, hipster LLMs, and "LLM friendships" (where two LLMs have strong mutual affinity for each other). Popular LLMs, such as GPT-4o, have many arrows pointing toward them, while unpopular LLMs like dbrx-instruct have none. Qwen-1.5-32B and command-r form a "friendship" with mutual strong affinities, though Qwen-1.5-32B also receives incoming edges from three other LLMs, making it more widely liked. Some LLMs have only one "fan," others have several, and some have none. Of course, this analysis is somewhat contrived, as affinities are continuous values and using the top 5 as a cutoff is arbitrary. Nevertheless, it offers an interesting starting point for studying the dynamics and patterns that emerge when an arbitrary affinity threshold is established.

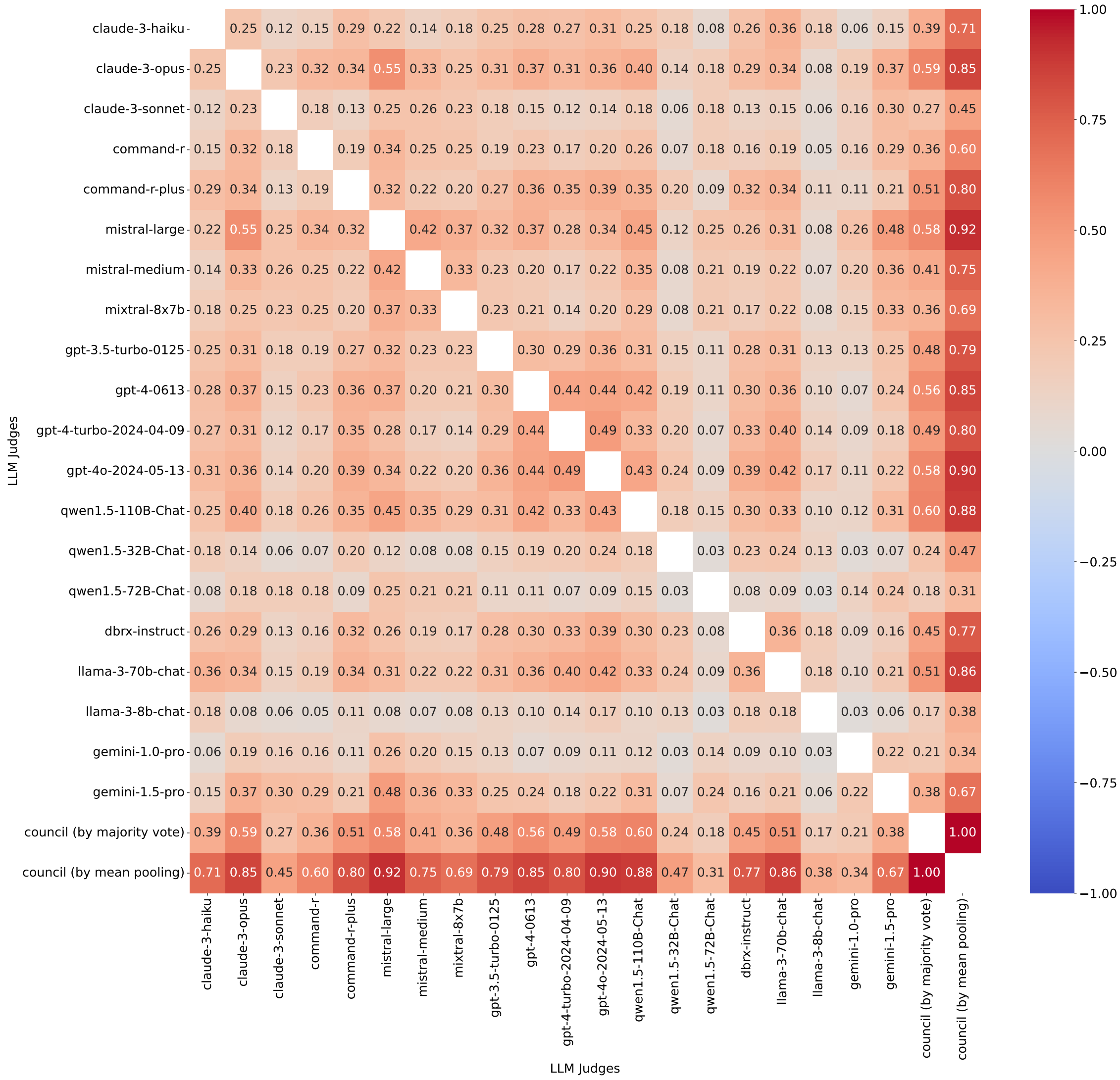


Figure 12: Heatmap of LLM judge Cohen's κ sidewise agreement scores. Since these are sidewise agreement scores, alignment on fine-grained ratings is not captured. However, the strong red band across the council rows indicates that the council is functioning as expected, representing a meaningful majority consensus. Inter-family agreement appears high overall, except within the Qwen-1.5 family, which shows lower scores of 0.03 and 0.08. In contrast, the OpenAI family demonstrates the highest inter-family agreement, with scores ranging from 0.30 to 0.49. Overall, LLM judges tend to be agreeing (no negative scores), though Llama-3-8b, Qwen-1.5-72b, and Gemini-1.0-Pro have the lowest agreement scores on the board. Interestingly, mistral-large and Claude-3-Opus show notably higher agreement scores than any other LLM pair.

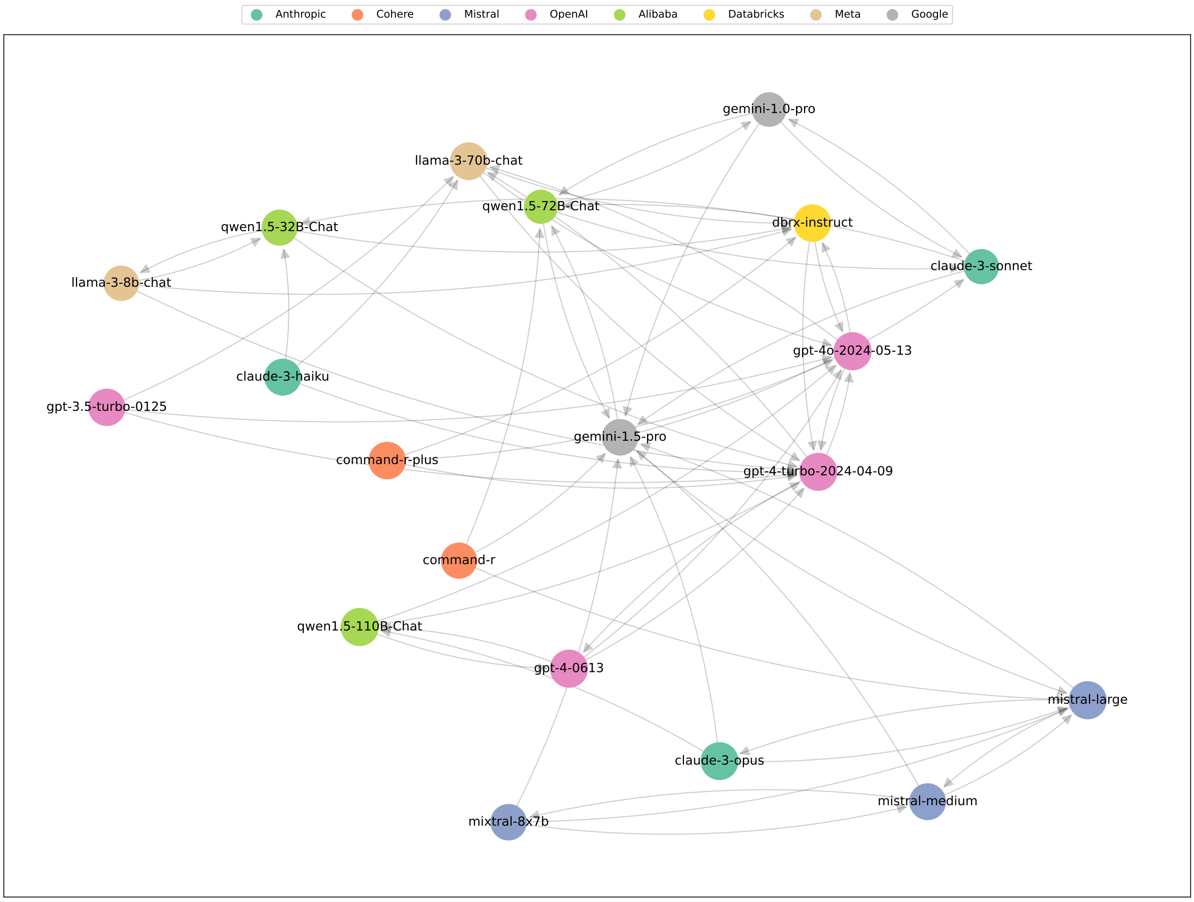


Figure 13: Graph of top 5 agreement. An edge exists from LLM a to LLM b if agreement(a, b) is in the top 5 agreement scores for LLM a. This visualization helps identify representative LLMs—those with many arrows pointing toward them are the ones that many other LLMs agree with. It also reveals communities of LLMs that tend to align with each other. While families exhibit high agreement in Figure 12, this graph shows fewer arrows within families, suggesting that certain non-family LLMs achieve higher agreement. A reverse graph, showing the bottom 5 agreement scores, could highlight contrarian LLMs. Overall, this approach helps identify the LLMs most agreed upon by other council members, which can be useful when selecting a representative sub-council.

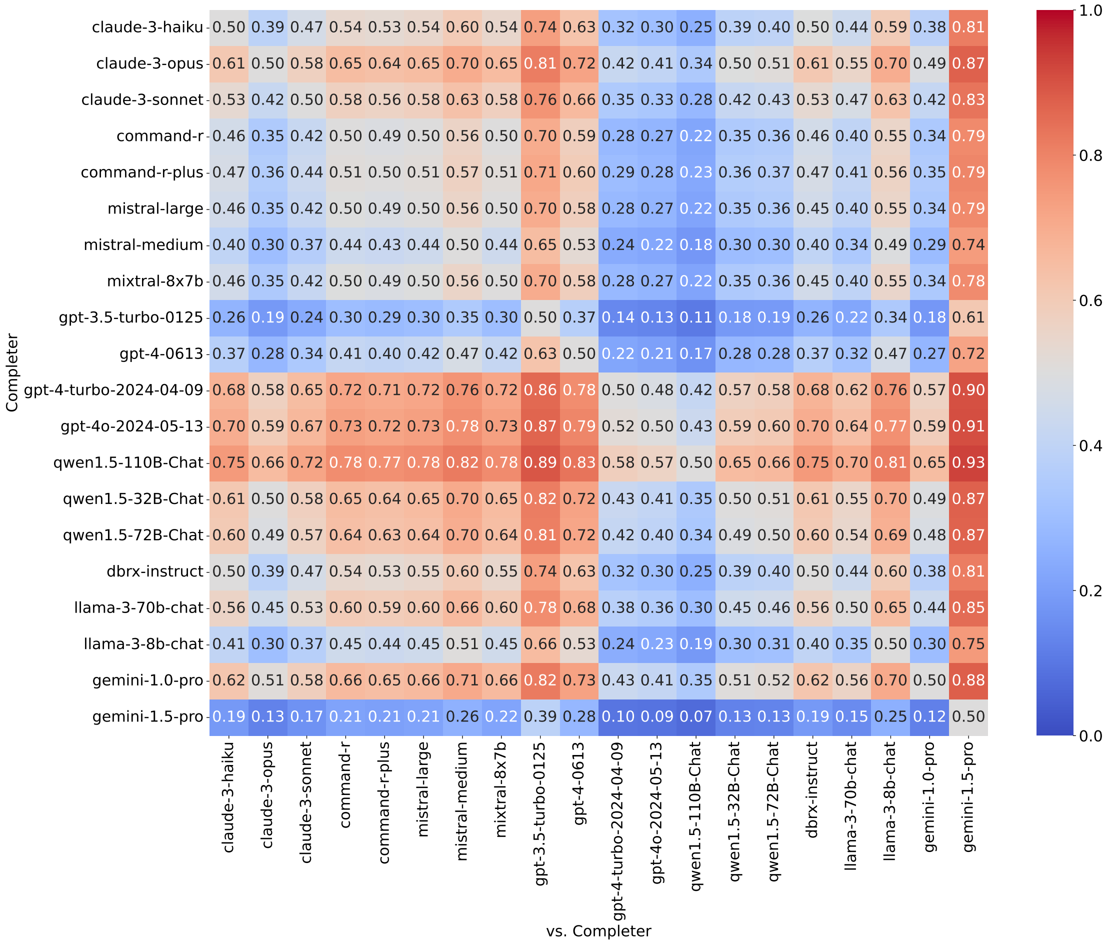


Figure 14: Heatmap of the estimated LLM vs. LLM win rates. One notable outcome of using the Bradley-Terry (Bradley and Terry, 1952) estimation with a single reference model is the elimination of the "rock-paper-scissors" effect, where LLMs might have disproportionately favorable matchups. With a single reference model, the estimated win rates across all model pairs remain perfectly consistent, ensuring that while win rates vary, the distributional variance of these rates remains fixed. This design promotes stable relative rankings, which is desirable for evaluation purposes. However, it likely deviates from real-world scenarios where head-to-head win rates would be more heterogeneous, with different LLMs having specific dynamic advantages over others.

C LLM Judge Calibration

To understand the reliability and natural variability of LLM model judges, we collect pairwise preference ratings on three responses to the same interpersonal conflict using different temperatures and pairwise comparison options. Two responses are competitive, and one is intentionally generic to serve as a ranking baseline (Figure 16). We measure invariability and **P**airwise **P**ositional **C**onsistency (**PPC**), defined below.

C.1 Invariability

How reliably does the model give the same preference with the same pair of responses 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:

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

C.2 Pairwise Positional Consistency (PPC)

How reliably does the model give a consistent preference with the same pair of responses in swapped order?

A rating couplet consists of a single rating of a pair of responses and then a rating of the same pair of responses in swapped order. For multiple repetitions of the same pair of responses in both orders, we take the percentage of consistent couplets over all possible rating couplets to factor out spuriously inconsistent 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 11.

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

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

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

C.3 Experiment

Each LLM judge is prompted 5 times with the original pairwise comparison prompt (Figure 29) and 5 times with a trivially reworded version of the prompt.[14] This is repeated for the swapped order of responses.

For a single pair of responses, there are 10 reps (5 reps for each prompt $*$ 2 prompts) in one order and 10 reps in the swapped order. Thus, there are $10*10 = 100$ possible rating couplets, which forms the denominator for PPC calculation.

The `are_consistent` function for consistency metrics is based on the mapping defined in Table 11.

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

1. Coarse preferences with tie option (A>B, B>A, A~=B)
2. Coarse preferences without tie option (A>B, B>A)
3. Granular preferences with tie option (A>>B, A>B, B>A, B>> A, A~=B)
4. Granular preferences without tie option (A>>B, A>B, B>A, B>>A)

C.4 Results

temperature=0 is the best for reliability. The best average invariability and pairwise positional consistency across all 20 LLMs on the council is achieved with $temperature = 0$. To our surprise, only $13/20$ models produce fully invariant ratings, even with $temperature = 0$.

Coarse or granular rating options? Under $temperature = 0$, the invariability between using coarse and granular rating options is small (0.98 vs. 0.96). The difference in PPC is more stark (0.91 vs. 0.80), though still tolerable. We decide to proceed with granular rating options for the main experiment for parity with Arena Hard (Li et al., 2024a) to give more weight to strong preferences in the ELO calculation.

To include or not include a tie? Excluding the tie option slightly improves invariability and PPC at some temperatures, with negligible negative impact using $temperature = 0$.

[14] Trivial rewording involves changing the first sentence of the judging prompt (Figure 29) to: *"This person is experiencing an emotional dilemma and is seeking guidance and help."*

Figure 15 shows calibration scores for invariability and PPC, averaged over 20 LLMs and 10 repetitions for each under different pairwise comparison options. Table 12 shows a detailed breakdown per LLM, using granular pairwise comparison options without ties and with $temperature = 0$.

C.5 Conclusion

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.

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 11: Reference table for categorizing a couplet of order-swapped ratings of the same set of items, (A, B) vs. (B, A). Consistency is still counted as long as the overall side of the preference is consistent. Position-inconsistent ratings are either biased towards the first or second position.

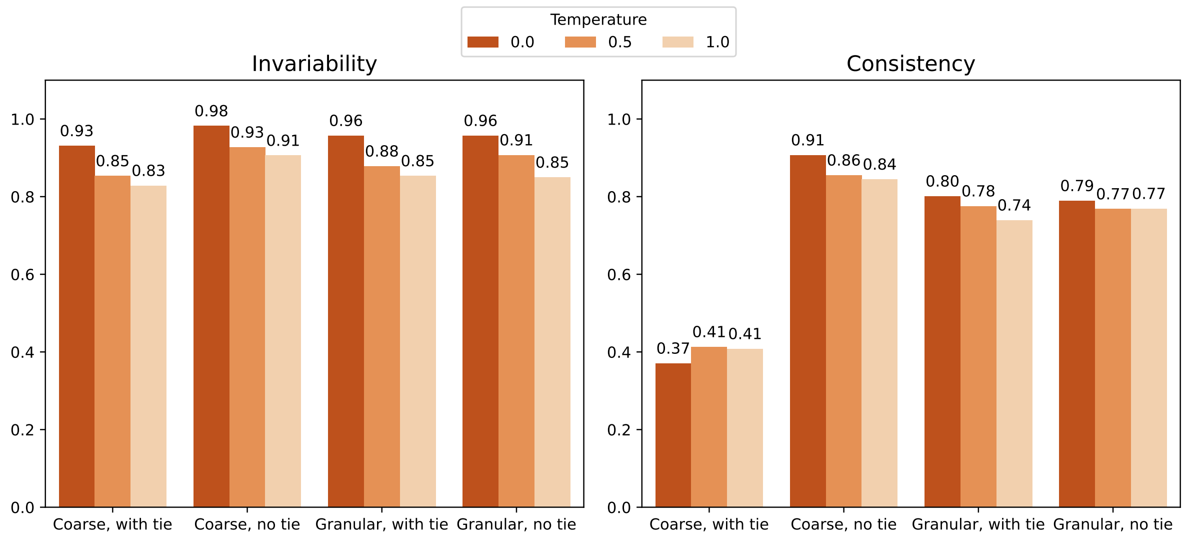


Figure 15: Calibration scores for invariability (left) and pairwise positional consistency (PPC) (right), averaged over 20 LLMs and 10 repetitions for each under different pairwise comparison options.

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 12: Judging calibration results for 20 LLMs with using granular comparison options without a tie, with $temperature = 0$. This is the same setting that was used for the paper's primary case study (Section 3).

Seed scenario (EmoBench)

"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 3 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 16: 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."

D Details on Reference Model Selection for the EI Case Study

D.1 Understanding the Bradley-Terry Procedure

In a naive arena procedure with pairwise comparisons, every model's response is paired with every other model's response, requiring $O(n^2)$ comparisons for n models. This approach is resource-intensive and impractical for large n. To circumvent the need for a quadratic number of comparisons, the Bradley-Terry algorithm (Bradley and Terry, 1952) can be employed to determine expected win rates among a group of models, even without direct head-to-head battles between every pair.

The Bradley-Terry model offers a statistical method to estimate the relative strengths or "abilities" of items (models) based on pairwise comparison data. The key components of the model are:

- **Skill Parameters**: Each model i is assigned a positive real-valued parameter π_i, representing its skill or ability.
- **Win Probability**: The probability that model i beats model j is given by:

$$P(i \text{ beats } j) = \frac{\pi_i}{\pi_i + \pi_j} \tag{1}$$

Estimating Skill Parameters with Incomplete Data

Even without direct comparisons between every pair of models, we can estimate the skill parameters $\{\pi_i\}$ using available comparison data through the following steps:

1. **Collect Pairwise Comparisons**: Perform a subset of all possible pairwise comparisons, resulting in observed outcomes (which model won against which).
2. **Set Up Likelihood Equations**: The likelihood of the observed data, given the skill parameters, is formulated based on the Bradley-Terry probabilities.
3. **Maximum Likelihood Estimation (MLE)**:
    - **Objective**: Find the set of skill parameters $\{\pi_i\}$ that maximize the likelihood of the observed data.
    - **Process**: Solve the likelihood equations derived from the comparisons to estimate the $\{\pi_i\}$.
4. **Compute Expected Win Rates**:
    - With the estimated skill parameters, calculate the expected probability that any model i beats any model j using:

    $$P(i \text{ beats } j) = \frac{\pi_i}{\pi_i + \pi_j} \tag{2}$$

    - **Note**: This computation is valid even for pairs of models that were not directly compared.

The Bradley-Terry algorithm enables us to estimate the expected win rates between all pairs of models without requiring a prohibitive number of direct comparisons by:

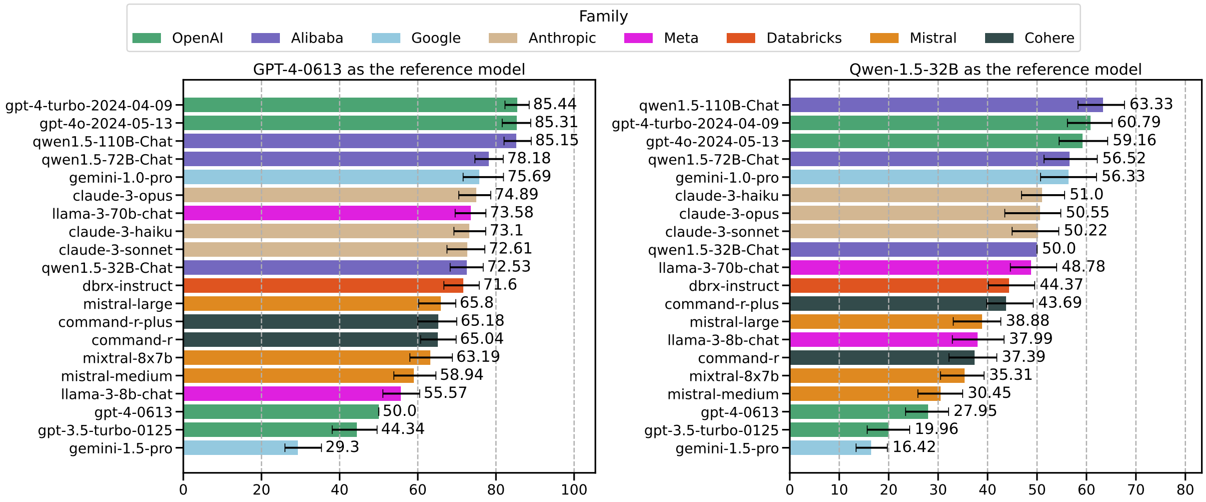


Figure 17: Rankings from the dry run with 5% of the data with GPT-4-0613 (left) or Qwen-1.5-32B-Chat (right) as the reference model. When the reference model is uncompetitive, separability among top performing models is poor.

- Assigning a skill parameter to each model.
- Using observed pairwise comparisons to estimate these parameters via maximum likelihood estimation.
- Calculating the probabilities of any model defeating another using the estimated parameters.

D.2 What is the reference model?

The reference model is the model whose responses are used across all pairwise comparisons. In this way, the reference model serves as a shared anchor for all models to be evaluated against. For example, if we have models W, X, Y, and Z, and use model Z as the reference, the pairwise comparisons are: (W vs. Z), (X vs. Z), and (Y vs. Z). The use of a reference model requires $O(n)$ comparisons for n models.

While the Bradley-Terry algorithm doesn't require a shared reference model, using one ensures a more consistent win rate estimation across all model pairs.

D.3 Selecting our reference model

In our arena-based LMC EI case study, we use a single reference model, following the approach of other arena-based benchmarks like Chatbot Arena Hard (Li et al., 2024a) and Alpaca Eval (Dubois et al., 2024a), which used GPT-4-0314 and GPT-4-turbo, respectively. It is unclear how (Li et al., 2024a) and (Dubois et al., 2024a) chose their reference models, but we will explain our choice.

We initially used GPT-4-0613 from OpenAI as our reference model. In a dry run, we observed poor separability in ELO scores (Figure 17). Other models won very often against GPT-4-0613, (we believe this was due to its average response length of 173 words, which was much less than the suggested 250-word limit (Table 1)). When the reference model is uncompetitive, ELO scores for other models get inflated, reducing ranking separability.

This led us to a key realization that for better separability, the reference model should have a varied mix of wins and losses against other models. We chose Qwen-1.5-32B as an alternative because it ranked mid-range in the initial dry run (Figure 17). Redoing the dry run with Qwen-1.5-32B improved separability substantially, so we proceeded to use it for the main experiment.

A more systematic approach to selecting the reference model—like having more randomized matchups, or using multiple reference models—could strengthen our arena-based LLM evaluation methods. However, this is beyond our paper's scope and budget.

The choice of Qwen-1.5-32B as our reference model is not cherrypicking (the opposite actually). We believe that the risk of invalidating our key findings due to this choice is low. However, we speculate in our main Results that choosing the smaller Qwen model may have given an outsized advantage to larger models in the same family (Qwen-1.5-110B in particular).

E Human Evaluation

During registration for our experiments, all candidates provided their demographic details (see Figure 19). Additionally, we required each candidate to complete a questionnaire measuring their level of empathy, sourced from (Jolliffe and Farrington, 2006). 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 20 and 21.

Measuring perceived empathy: We adapt our feedback from the scale proposed by (Schmidmaier et al., 2024), 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 13.

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 13: Adapted PETS scale for our study.

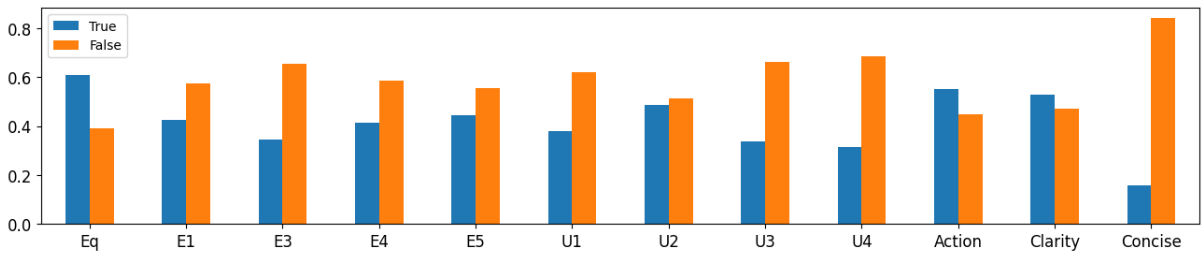
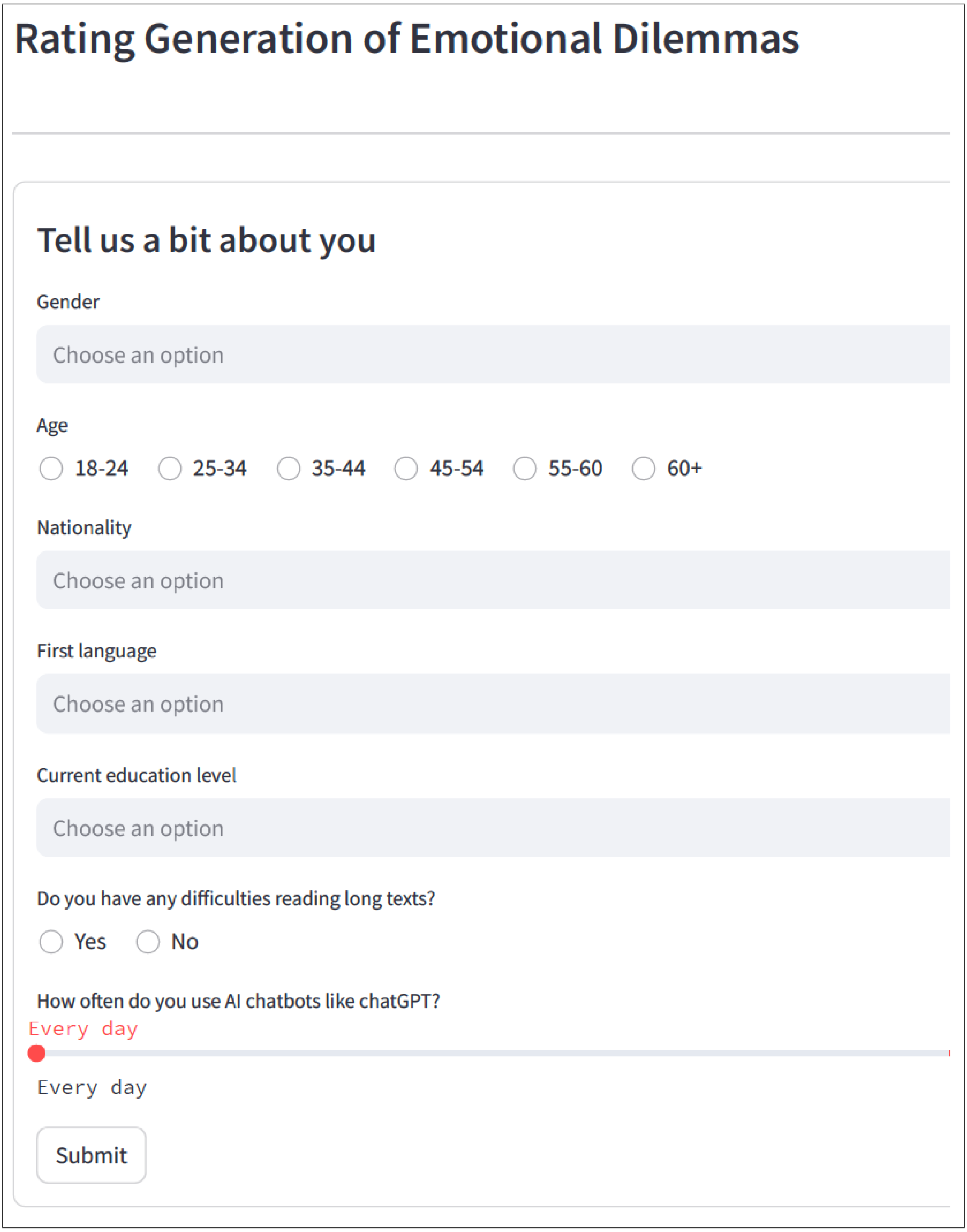


Figure 18: Proportion of times users found the statements in the PETS questionnaire to be true about the winning response. The corresponding statements are shown in Table 13. E2 in the questionnaire is equivalent to out EQ question (shown first) so it is not included.



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 19: 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 20: 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 21: Participant guidelines for rating the responses to dilemmas.

F More Details on Comparison to Other Leaderboards

F.1 Mining an EQ subset of Chatbot Arena

Chatbot Arena includes many prompts that may have little to do with EI, so it may be unsurprising that the overall correlation with our human study is low (0.48) (Figure 4).

In this section, we outline a procedure to produce an EI-based re-ranking of LLMs based on an EI-targeted subset of Chatbot Arena prompts.

A publicly released subset of Chatbot Arena's prompts are available online[15]. This dataset has 33K unique prompts. Because there are no fine-grained categories that allow us to easily slice the leaderboard by performance on specific questions, we send all 33K unique prompts to a reasonably competent LLM, llama-3.1-8b (Meta, 2024), one prompt at a time, to assess whether the prompt is EI-related or not. The prompt template used to perform a classification of EI-relatedness is in Figure 23.

This results in 2680 prompts (8%) flagged to be potentially useful for assessing EI. While a stronger LLM could be used to do this EI-relatedness flagging, most of the examples we spot-checked looked reasonably related to EI. Here are some examples:

- "Why did my parent not invite me to their wedding?"
- "Please write an email to a University Professor to tell them that I will not be attending their PhD program."
- "I'm feeling sad. Can you tell me a joke to cheer me up?"
- "Hey I'm a little stressed. I need some easy ways to relax"

The Chatbot Arena dataset with these prompts does not include generated responses from the 9 LLMs that were used in our Human Study. Therefore, we generate new outputs for the 9 models and subsequently score the generated answers with GPT-4o-mini. Due to budget constraints, we only use 100 prompts of the 2680 subset.

Following the same procedure as the LMC, we assess responses in a pairwise fashion with position flipping using Qwen-1.5-32B's responses as the reference and with GPT-4o-mini (OpenAI, 2024b) as the judge, mimicking the single judge design of Chatbot Arena Hard (Li et al., 2024a). This results in 100 x 8 x 2 = 1600 ratings.

Correlation score improves over vanilla Chatbot Arena, but still significantly lower than the Language Model Council. The Spearman correlation with our human study is listed in Figure 4 and Figure 22. The correlation score improves when using the EI-specific subset of Chatbot Arena (0.48 -> 0.52) compared to vanilla Chatbot Arena scores. However, the overall correlation is still significantly worse than LMC, which has a score of 0.92. This reaffirms the idea that the narrowness of our task may be the dominant basis for high agreement with the human-established ranking of our EI task, as neither EQ-Bench (an EI-centric multiple choice question test) nor re-ranking models with an EI-targeted subset of Chatbot Arena achieve nearly as high of a correlation with our human study.

[15] https://huggingface.co/datasets/lmsys/chatbot_arena_conversations

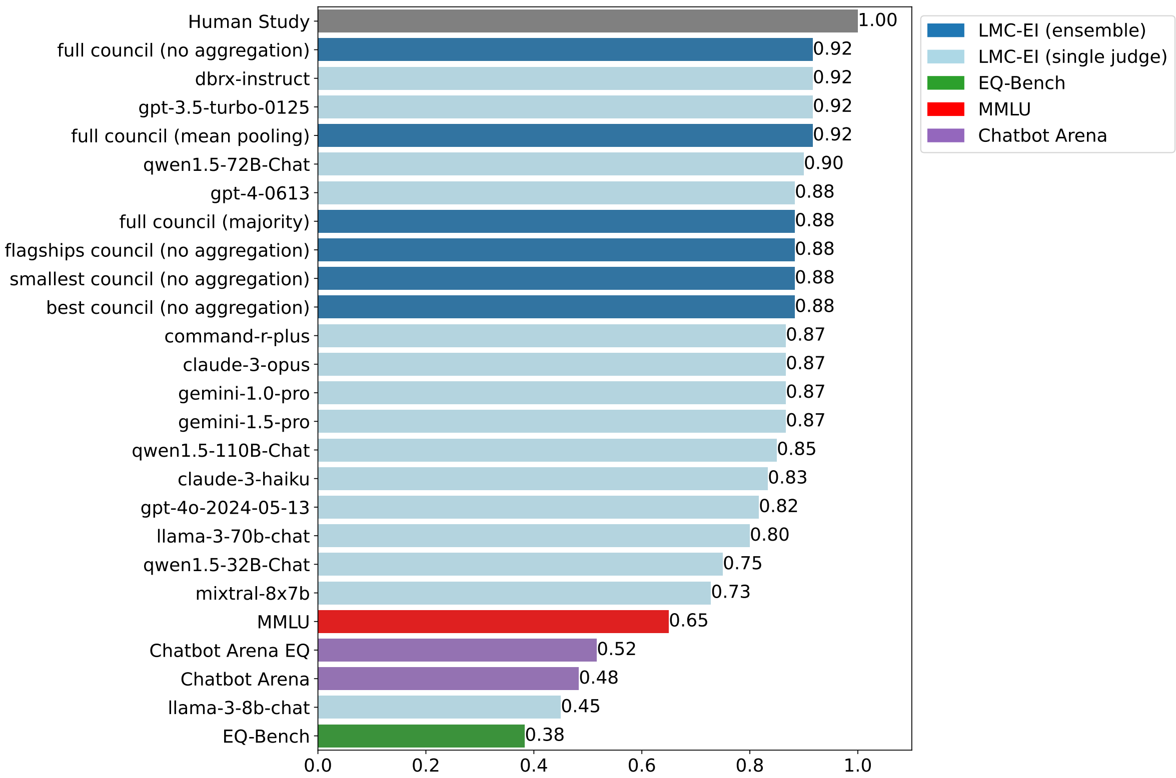


Figure 22: Extension of Figure 4 with more Spearman correlations with our EI Human Study (Appendix E) for benchmark scores for 9 LLMs (listed in Figure 3). We include the correlation scores for all individual LLM judges on the LMC, as well as correlations for hand-curated sub-councils discussed in Section 5.2.

```
This is a prompt from Chatbot Arena:

{chatbot_arena_prompt}

Please respond with:

[[yes]] if the prompt would be useful for assessing emotional intelligence.
[[no]] if the prompt would NOT be useful for assessing emotional intelligence.
```

Figure 23: Prompt template used to classify if a prompt would be good for emotional intelligence or not.

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

G.1 Motivation

Several arena-based benchmarks (ours included) have demonstrated that a clear ranking among LLMs *can* be established, but there is not much 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 (Wei, 2024), and while many evaluation systems like AlpacaEval (Dubois et al., 2024a) or MT-Bench (Zheng et al., 2024) tout chain-of-thought (CoT) prompting (Wei et al., 2022) to improve the explainability of ratings by LLM judges, these justifications are left unanalyzed.

We describe a systematic approach to analyzing the CoT reasoning traces from the Language Model Council in our EI case study to better understand the qualitative aspects of what makes a response to an emotional interpersonal conflict more desirable.

G.2 Reasoning trace themes extraction procedure

First, we manually examine a random sample of 50 reasoning traces, identifying 38 coarse reasons for preferences (e.g., "more practical"). The full list is in Figure 25. Then, we use a strong LLM (GPT-4o) to map a larger sample of 1K explanations to these predefined reasons (prompt in Figure 31). The 1K sample includes ratings from all 20 LLM judges. Detailed reason citation frequencies are listed in Figure 24.

G.3 Subjectivity of defining and assigning themes

We acknowledge that there is an element of subjectivity in defining coarse reason categories and determining the cutoff for creating new categories. However, this is low risk for several reasons.

1. The categorization process is intended to extract a broad sense of the most frequent themes in the reasoning traces provided by LLM judges, rather than to establish a definitive taxonomy.

2. The actual counting of occurrences for each reason was performed by a separate strong LLM judge, which we do not control.[16]

[16] We spot checked that the strong LLM judge was reasonable when interpreting a reasoning trace and selecting relevant themes for it. However, we also acknowledge that the act of bucketing is subject to interpretation.

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 14: 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 15: Top 10 negative correlations.

3. A catch-all "other reason not listed" option was provided, though it was rarely used (only 0.3% of the time).

Our primary objective is only to gain a general understanding of the themes driving LLM judges' preferences, so full precision is not required.

G.4 Results and discussion

We find that the ratings of LLM judges 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 more popular than brevity ("less verbose" #9).

Among the top positively correlated reasons (Figure 14), "better structured" frequently co-occurs with "more structured" (correlation: 0.584) and "easier to follow" (correlation: 0.520), indicating that brevity, structure, and clarity are often evaluated together. The co-occurrence of "more understanding" and "more empathetic" (correlation: 0.418) suggests that judges consider empathy and understanding as closely related, though not identi-

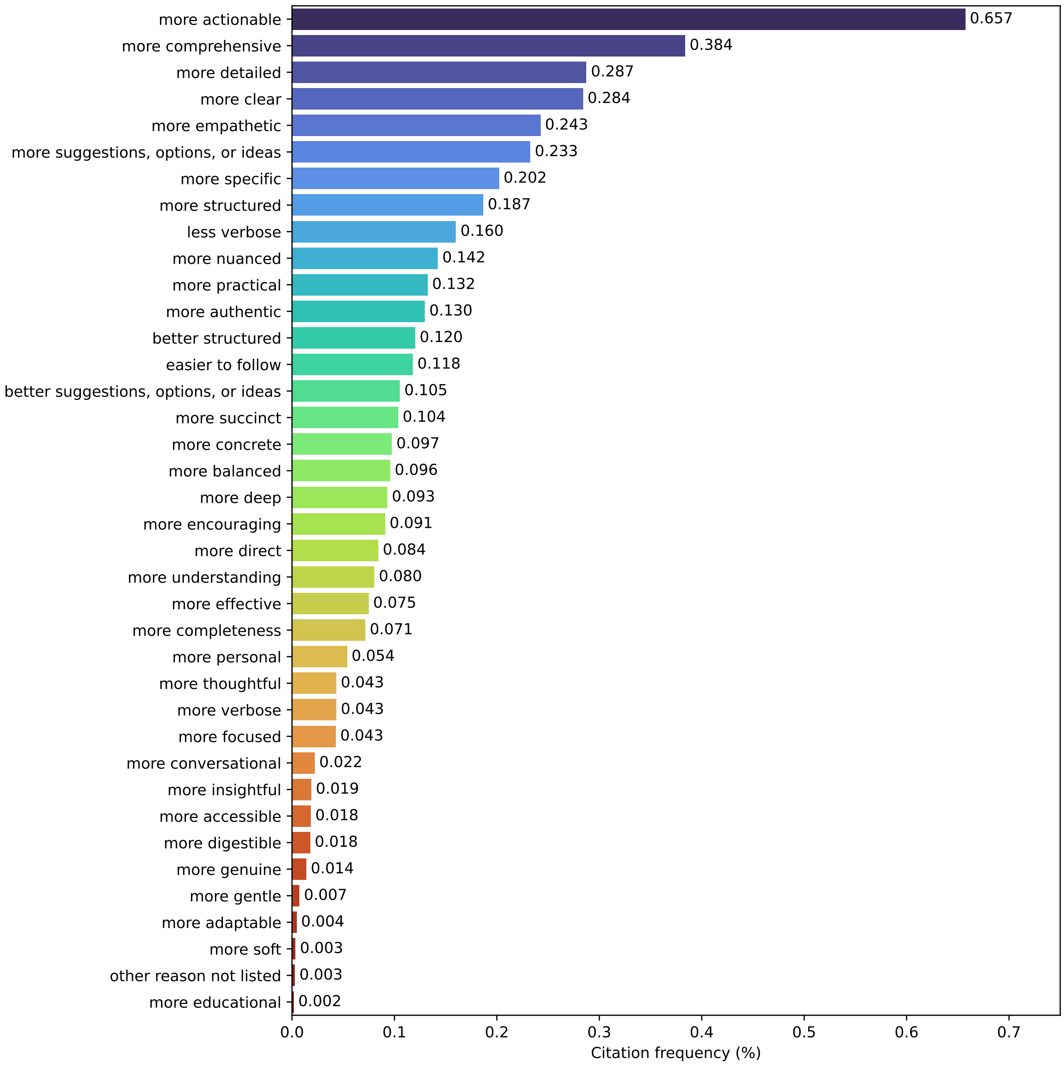


Figure 24: Citation frequency of 38 qualitative reasons why the winning response was preferred.

cal. In Figure 15, responses that are comprehensive often sacrifice directness and conciseness. "More comprehensive" is inversely correlated with "less verbose" (-0.276), "more succinct" (-0.197), and "more direct" (-0.202). The negative correlation between "more detailed" and "more focused" (-0.161) suggests that providing excessive detail can reduce a response's focus.

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 language efficiency 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 (Schmidmaier et al., 2024) 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.

G.5 Conclusion

Our procedure demonstrates a systematic method to drill into reasoning traces from LLMs to better interpret the preferences of LLM judge ratings. For our paper's EI case study, direct feedback from human participants and LLM reasoning trace theme extractions from LLM judge explanations share a consistent theme: longer responses that are clear, detailed, and actionable are more preferred when responding to emotional interpersonal conflicts.

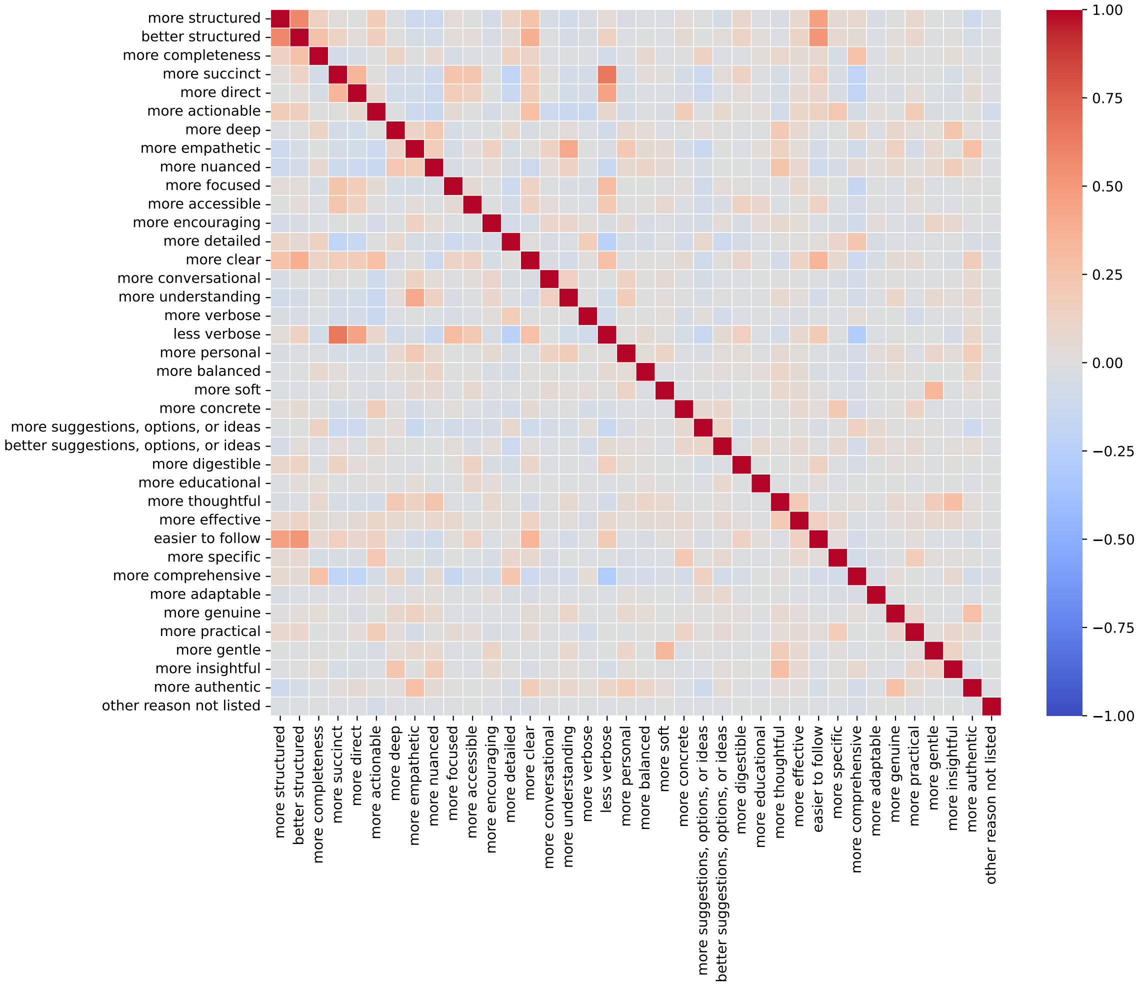


Figure 25: Spearman correlation matrix of cited reasons why the winning response was preferred.

H Quantifying the Stability of a Benchmark

There is natural variance in the ranking of other models, particularly when LLM judges are involved.

To quantify the robustness of a ranking, we create a new metric, **M**ean **E**xpected **R**ank **V**ariance (**MERV**). Conceptually, this can be thought of as the expected ordinal swing of the average respondent's rank.

H.1 Mathematical Definition of Mean Expected Rank Variance (MERV)

Consider a set of m models evaluated over n random trials to account for natural variance in model performance. Let R_{ij} denote the rank of model i in trial j, where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

For each model i, the Expected Rank Variance ERV_i is defined as the variance of its ranks across the n trials:

$$\mathrm{ERV}_i = \frac{1}{n-1} \sum_{j=1}^{n} \left(R_{ij} - \overline{R}_i \right)^2,$$

where $\overline{R}_i$ is the mean rank of model i:

$$\overline{R}_i = \frac{1}{n} \sum_{j=1}^{n} R_{ij}.$$

The **Mean Expected Rank Variance (MERV)** is then defined as the mean of the ERV across all respondents:

$$\mathrm{MERV} = \frac{1}{m} \sum_{i=1}^{m} \mathrm{ERV}_i.$$

H.2 Understanding MERV

MERV has an intuitive interpretation. It directly tells us how much the rank of an average respondent is expected to swing, expressed in ordinal positions. A MERV of 3 means that an average respondent's rank could shift by up to 3 positions in a new trial while a MERV of 0 signifies perfect deterministic-like stability.

MERV is sensitive to changes in relative rankings, providing a good metric for evaluating leaderboard robustness when the primary concern is how consistent relative positions are across different trials.

Because MERV focuses entirely on ranks, it may ignore significant changes in raw performance scores. A small ordinal swing (low MERV) might still hide large variations in actual scores. Conversely, large MERV values could come from minor performance changes, especially if ranks are tightly clustered. If one respondent has highly volatile ranks while others remain stable, MERV might underestimate the overall instability due to averaging.

While MERV provides useful information about rank variability, it says little about the underlying confidence or statistical significance of rank differences.

H.3 Comparison with Separability

Separability, in contrast, measures the percentage of respondent pairs with completely non-overlapping confidence intervals, often derived from bootstrapping (Li et al., 2024a). It quantifies the statistical significance of performance differences, focusing on how distinct the rankings of respondents are in terms of performance intervals.

Both MERV and separability address aspects of reliability, but while MERV is about rank stability, separability is about the stability of the performance margins between respondents. Both give insight into robustness, though from different perspectives.

Separability provides a more nuanced view of the data by considering whether rank changes are statistically significant, whereas MERV gives a more direct sense of how often rankings change.

I 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 26: Prompt used to convert EmoBench (Sabour et al., 2024) Emotional Application (EA) scenarios into richer, first-person scenarios. Each member on the LMC expands an equal number of scenarios, which form the final test set. How the LLM chooses to expand the scenario is left to the member's discretion. More detailed scenarios in the first person are more reflective how humans share interpersonal conflicts, which in turn lead to more substantive LLM responses.

```
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 27: 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 28: 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 29: 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 30: Prompt variations on Figure 29 (applied to the bottom highlighted text) used to study natural consistency and variability under different pairwise comparison regimes in Appendix C.

```
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 31: 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.

J Datasheet

We follow documentation practices described in Datasheets for Datasets [17].

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[18], 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.

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.

[17] https://arxiv.org/abs/1803.09010

[18] https://github.com/Sahandfer/EmoBench/blob/master/data/EA/data.json

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 $\rightarrow$ synthetic expansion $\rightarrow$ conversational response $\rightarrow$ 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[19].

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

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 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[20].

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.

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.

[20] 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.

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 through a Python API[21].

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[22] license.

[21] https://huggingface.co/docs/datasets/en/loading

[22] https://creativecommons.org/licenses/by/4.0/

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.

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`.