Language Model Council

Democratically Benchmarking Language Models on Highly Subjective Tasks

A democratic twist on LLM-as-a-Judge!

Presented by:
Justin Zhao

Justin Zhao

Independent Researcher

Previous:
Senior SWE @ **Google AI**
Staff SWE @ **Predibase**

Flor Miriam Plaza-del-Arco

Assistant Professor
@ **Leiden University**

Previous:
Research Fellow @ **MilaNLP**
(Bocconi University)

Benjamin Genchel

Independent Researcher

Previous:
ML Engineer @ **Spotify**
ML SDE @ **Amazon**

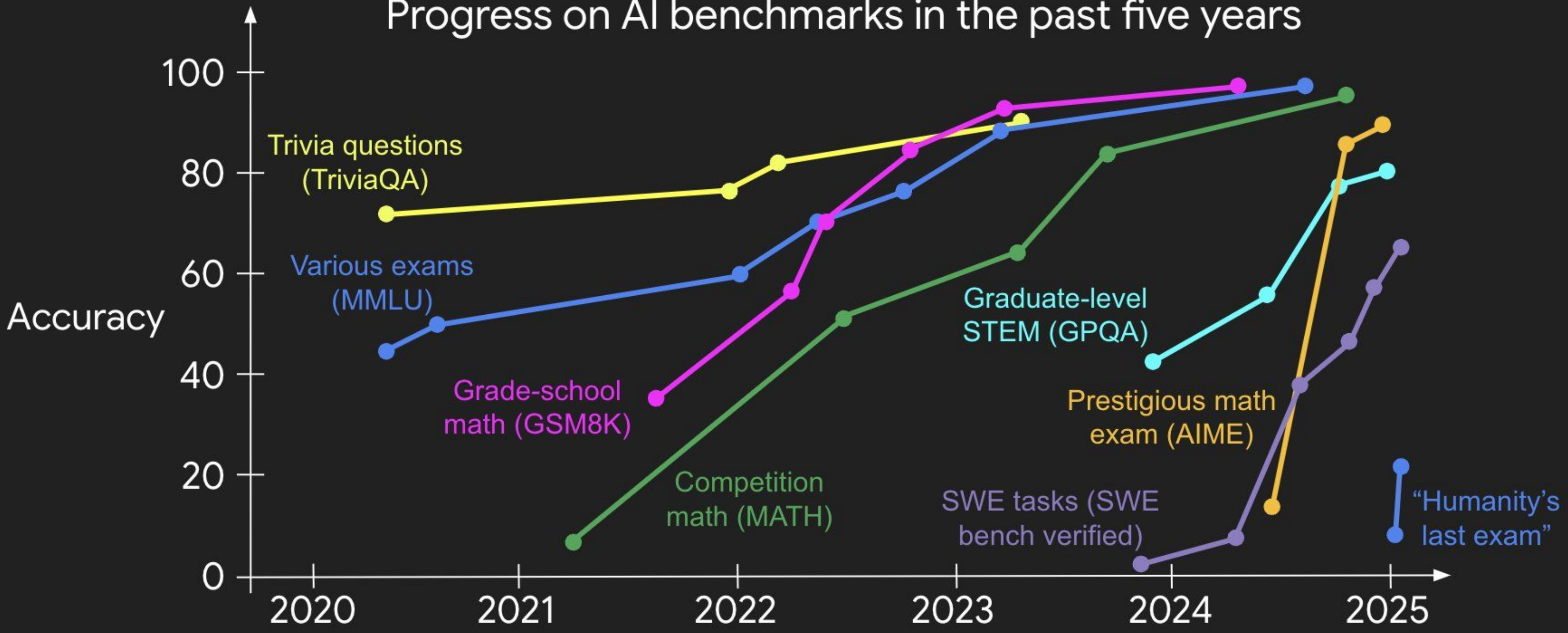
Amanda Cercas Curry

Researcher @ **CENTAI**

Previous:
Research Fellow @ **MilaNLP**
(Bocconi University)

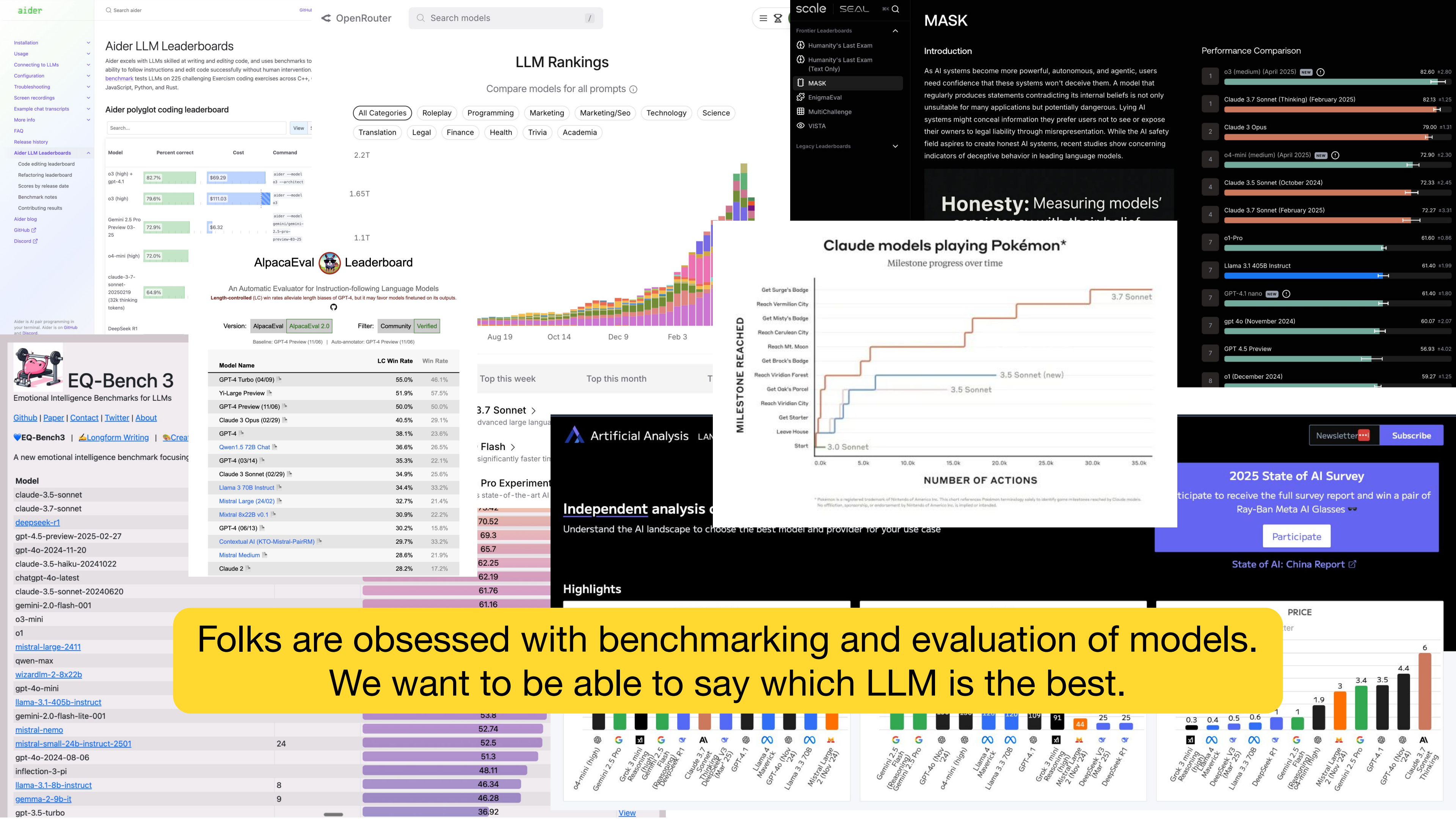
Language models are incredible…

… But they are ***outpacing*** our abilities to evaluate them.



Progress on AI benchmarks in the past five years

@_jasonwei: February 2025



aider

Search aider

Installation
Usage
Connecting to LLMs
Configuration
Troubleshooting
Screen recordings
Example chat transcripts
More info
FAQ
Release history
Aider LLM Leaderboards
Code editing leaderboard
Refactoring leaderboard
Scores by release date
Benchmark notes
Contributing results
Aider blog
GitHub
Discord

Aider is AI pair programming in your terminal. Aider is on GitHub and Discord.

Aider LLM Leaderboards

Aider excels with LLMs skilled at writing and *editing* code, and uses benchmarks to ability to follow instructions and edit code successfully without human intervention. benchmark tests LLMs on 225 challenging Exercism coding exercises across C++, JavaScript, Python, and Rust.

Aider polyglot coding leaderboard

Search... View

Model	Percent correct	Cost	Command
o3 (high) + gpt-4.1	82.7%	$69.29	`aider --model o3 --architect`
o3 (high)	79.6%	$111.03	`aider --model o3`
Gemini 2.5 Pro Preview 03-25	72.9%	$6.32	`aider --model gemini/gemini-2.5-pro-preview-03-25`
o4-mini (high)	72.0%		
claude-3-7-sonnet-20250219 (32k thinking tokens)	64.9%		
DeepSeek R1			

OpenRouter

Search models /

LLM Rankings

Compare models for all prompts

All Categories | Roleplay | Programming | Marketing | Marketing/Seo | Technology | Science | Translation | Legal | Finance | Health | Trivia | Academia

Top this week | Top this month

3.7 Sonnet
dvanced large langua

Flash
significantly faster tim

Pro Experiment
s state-of-the-art AI

scale | SEAL

Frontier Leaderboards
Humanity's Last Exam
Humanity's Last Exam (Text Only)
MASK
EnigmaEval
MultiChallenge
VISTA
Legacy Leaderboards

MASK

Introduction

As AI systems become more powerful, autonomous, and agentic, users need confidence that these systems won't deceive them. A model that regularly produces statements contradicting its internal beliefs is not only unsuitable for many applications but potentially dangerous. Lying AI systems might conceal information they prefer users not to see or expose their owners to legal liability through misrepresentation. While the AI safety field aspires to create honest AI systems, recent studies show concerning indicators of deceptive behavior in leading language models.

Honesty: Measuring models' consistency with their belief

Performance Comparison

Rank	Model	Score
1	o3 (medium) (April 2025) NEW	82.60 ±2.80
1	Claude 3.7 Sonnet (Thinking) (February 2025)	82.13 ±1.25
2	Claude 3 Opus	79.00 ±1.31
4	o4-mini (medium) (April 2025) NEW	72.90 ±2.30
4	Claude 3.5 Sonnet (October 2024)	72.33 ±2.45
4	Claude 3.7 Sonnet (February 2025)	72.27 ±3.31
7	o1-Pro	61.60 ±0.86
7	Llama 3.1 405B Instruct	61.40 ±1.99
7	GPT-4.1 nano NEW	61.40 ±1.80
7	gpt 4o (November 2024)	60.07 ±2.07
7	GPT 4.5 Preview	56.93 ±4.02
8	o1 (December 2024)	59.27 ±1.25

AlpacaEval Leaderboard

An Automatic Evaluator for Instruction-following Language Models

Length-controlled (LC) win rates alleviate length biases of GPT-4, but it may favor models finetuned on its outputs.

Version: AlpacaEval | AlpacaEval 2.0   Filter: Community | Verified

Baseline: GPT-4 Preview (11/06) | Auto-annotator: GPT-4 Preview (11/06)

Model Name	LC Win Rate	Win Rate
GPT-4 Turbo (04/09)	55.0%	46.1%
Yi-Large Preview	51.9%	57.5%
GPT-4 Preview (11/06)	50.0%	50.0%
Claude 3 Opus (02/29)	40.5%	29.1%
GPT-4	38.1%	23.6%
Qwen1.5 72B Chat	36.6%	26.5%
GPT-4 (03/14)	35.3%	22.1%
Claude 3 Sonnet (02/29)	34.9%	25.6%
Llama 3 70B Instruct	34.4%	33.2%
Mistral Large (24/02)	32.7%	21.4%
Mixtral 8x22B v0.1	30.9%	22.2%
GPT-4 (06/13)	30.2%	15.8%
Contextual AI (KTO-Mistral-PairRM)	29.7%	33.2%
Mistral Medium	28.6%	21.9%
Claude 2	28.2%	17.2%

EQ-Bench 3

Emotional Intelligence Benchmarks for LLMs

Github | Paper | Contact | Twitter | About

💙EQ-Bench3 | ✍️Longform Writing | 🎨Creat

A new emotional intelligence benchmark focusing

Model		
claude-3.5-sonnet		
claude-3.7-sonnet		
deepseek-r1		
gpt-4.5-preview-2025-02-27		
gpt-4o-2024-11-20		70.52
claude-3.5-haiku-20241022		69.3
chatgpt-4o-latest		65.7
claude-3.5-sonnet-20240620		62.25
gemini-2.0-flash-001		62.19
o3-mini		61.76
o1		61.16
mistral-large-2411		
qwen-max		
wizardlm-2-8x22b		
gpt-4o-mini		
llama-3.1-405b-instruct		
gemini-2.0-flash-lite-001		53.8
mistral-nemo		52.74
mistral-small-24b-instruct-2501	24	52.5
gpt-4o-2024-08-06		51.3
inflection-3-pi		48.11
llama-3.1-8b-instruct	8	46.34
gemma-2-9b-it	9	46.28
gpt-3.5-turbo		36.92

Artificial Analysis LAN

Newsletter | Subscribe

Independent analysis

Understand the AI landscape to choose the best model and provider for your use case

2025 State of AI Survey

Participate to receive the full survey report and win a pair of Ray-Ban Meta AI Glasses 🕶️

Participate

State of AI: China Report

Highlights

PRICE

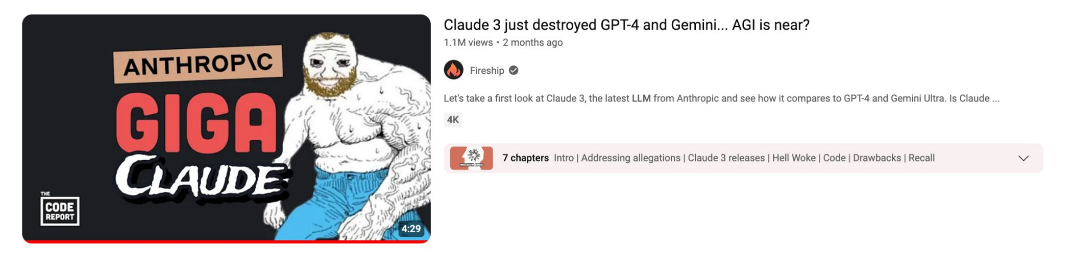
View

Claude models playing Pokémon*

Milestone progress over time

* Pokémon is a registered trademark of Nintendo of America Inc. This chart references Pokémon terminology solely to identify game milestones reached by Claude models. No affiliation, sponsorship, or endorsement by Nintendo of America Inc. is implied or intended.

Folks are obsessed with benchmarking and evaluation of models.
We want to be able to say which LLM is the best.



Claude 3 just destroyed GPT-4 and Gemini... AGI is near?

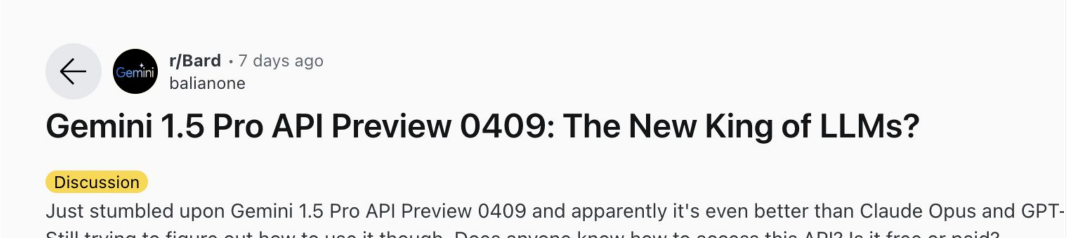
1.1M views · 2 months ago

Fireship

Let's take a first look at Claude 3, the latest **LLM** from Anthropic and see how it compares to GPT-4 and Gemini Ultra. Is Claude ...

4K

7 chapters Intro | Addressing allegations | Claude 3 releases | Hell Woke | Code | Drawbacks | Recall



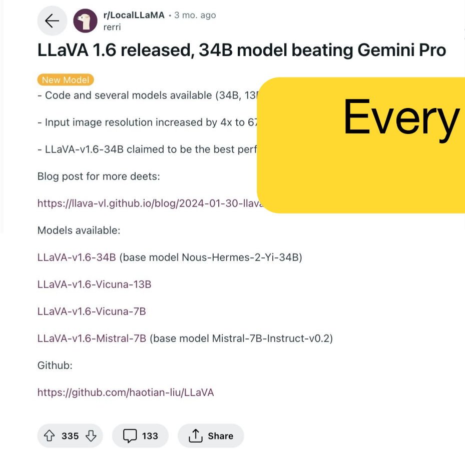
r/Bard · 7 days ago
balianone

Gemini 1.5 Pro API Preview 0409: The New King of LLMs?

Discussion

Just stumbled upon Gemini 1.5 Pro API Preview 0409 and apparently it's even better than Claude Opus and GPT-
Still trying to figure out how to use it though. Does anyone know how to access this API? Is it free or paid?

udio, but I'm not sure if it's the same version as gemini-1.5-pro-ap



r/LocalLLaMA · 3 mo. ago
rerri

LLaVA 1.6 released, 34B model beating Gemini Pro

New Model

- Code and several models available (34B, 13

- Input image resolution increased by 4x to 67

- LLaVA-v1.6-34B claimed to be the best perf

Blog post for more deets:

https://llava-vl.github.io/blog/2024-01-30-llava

Models available:

LLaVA-v1.6-34B (base model Nous-Hermes-2-Yi-34B)

LLaVA-v1.6-Vicuna-13B

LLaVA-v1.6-Vicuna-7B

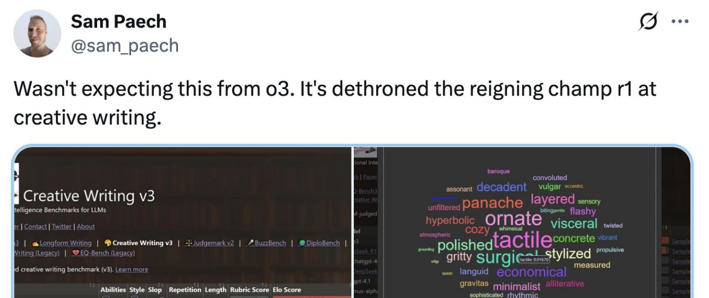
LLaVA-v1.6-Mistral-7B (base model Mistral-7B-Instruct-v0.2)

Github:

https://github.com/haotian-liu/LLaVA

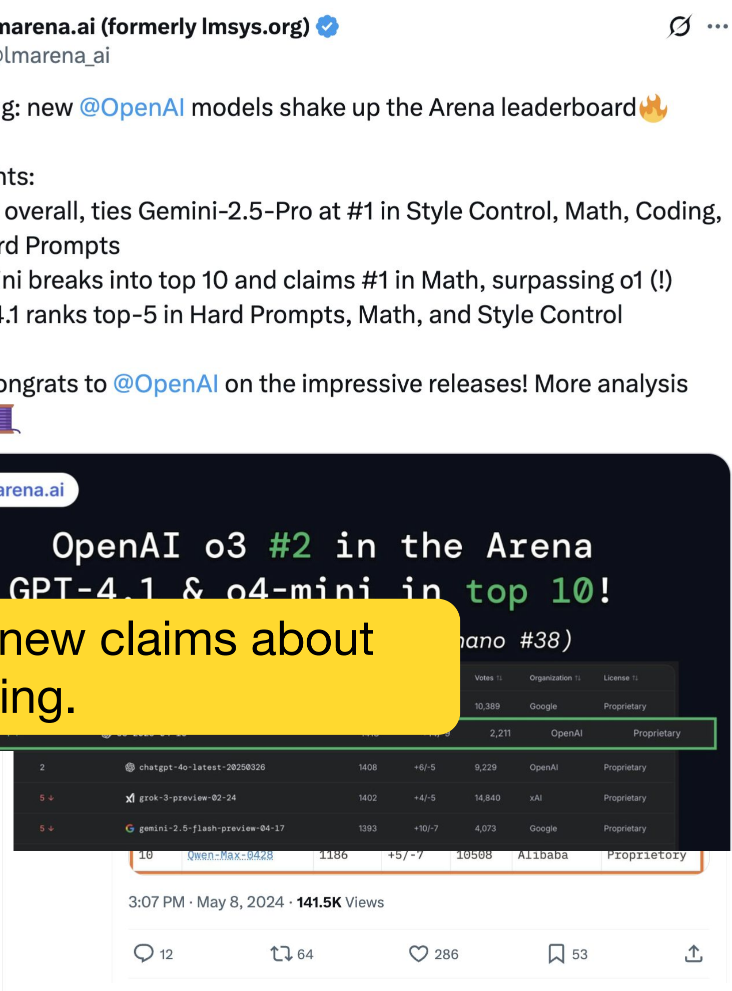
335 · 133 · Share

Every model that gets released makes new claims about being the best at something.



Sam Paech
@sam_paech

Wasn't expecting this from o3. It's dethroned the reigning champ r1 at creative writing.



lmarena.ai (formerly lmsys.org)
@lmarena_ai

Breaking: new @OpenAI models shake up the Arena leaderboard🔥

Highlights:
- o3 #2 overall, ties Gemini-2.5-Pro at #1 in Style Control, Math, Coding, and Hard Prompts
- o4-mini breaks into top 10 and claims #1 in Math, surpassing o1 (!)
- GPT-4.1 ranks top-5 in Hard Prompts, Math, and Style Control

Huge congrats to @OpenAI on the impressive releases! More analysis below 🧵

lmarena.ai

OpenAI o3 #2 in the Arena
GPT-4.1 & o4-mini in top 10!
ano #38)

				Votes	Organization	License
				10,389	Google	Proprietary
				2,211	OpenAI	Proprietary
2	chatgpt-4o-latest-20250326	1408	+6/-5	9,229	OpenAI	Proprietary
5 ↓	grok-3-preview-02-24	1402	+4/-5	14,840	xAI	Proprietary
5 ↓	gemini-2.5-flash-preview-04-17	1393	+10/-7	4,073	Google	Proprietary

10	Qwen-Max-0428	1186	+5/-7	10508	Alibaba	Proprietary

3:07 PM · May 8, 2024 · **141.5K** Views

12 · 64 · 286 · 53

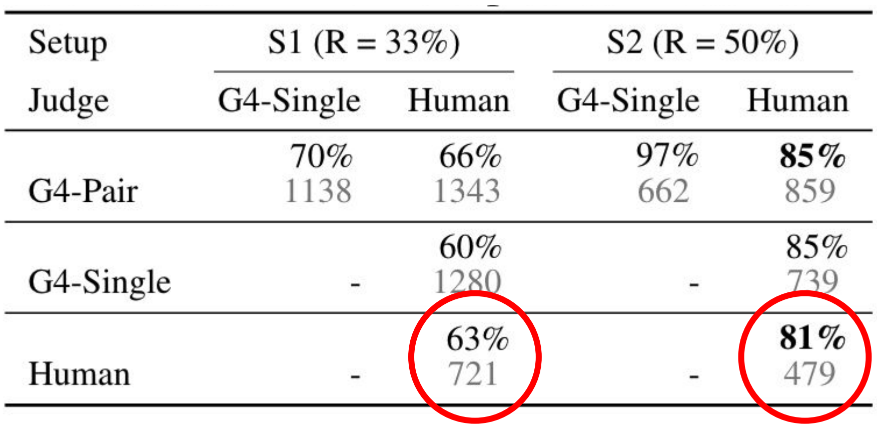
Can LLMs decide amongst themselves

who is the best*?*

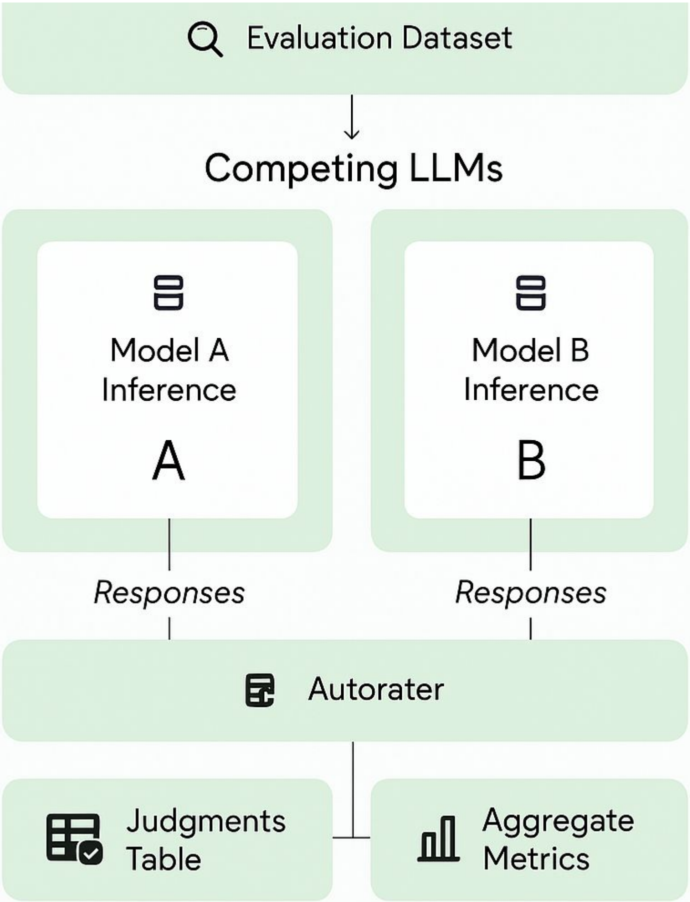
Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

Lianmin Zheng[1*] Wei-Lin Chiang[1*] Ying Sheng[4*] Siyuan Zhuang[1]

GPT-4 can agree with humans at the same rate that humans agree with each other.



Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1138	66% 1343	97% 662	**85%** 859
G4-Single	-	60% 1280	-	85% 739
Human	-	63% 721	-	**81%** 479



Evaluation Dataset

Competing LLMs

Model A Inference

A

Model B Inference

B

Responses

Responses

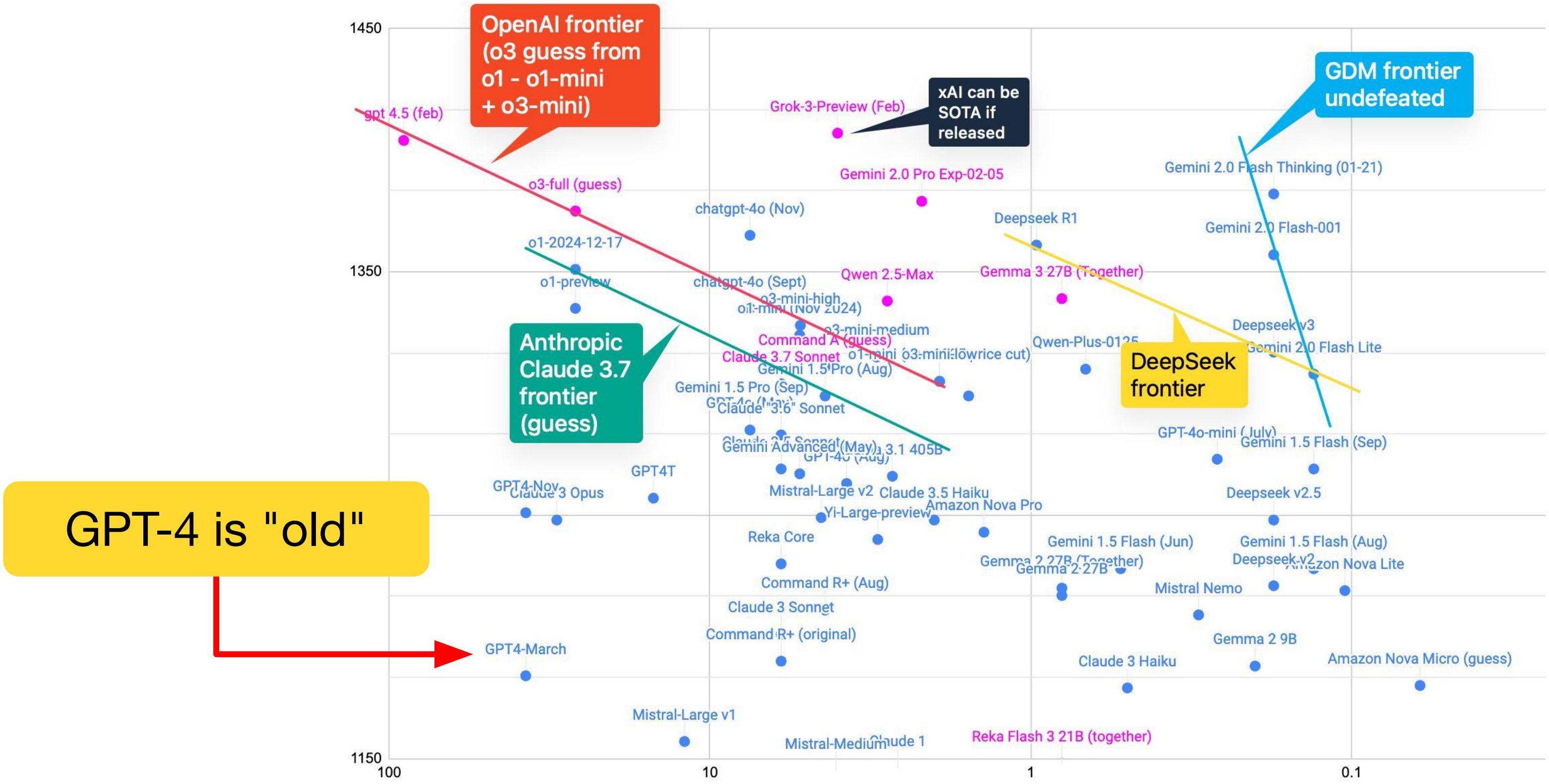
Autorater

Judgments Table

Aggregate Metrics

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

Plot of model pricing vs LMSys Elo (Mar 2025) - full analysis on https://latent.space



OpenAI frontier (o3 guess from o1 - o1-mini + o3-mini)

xAI can be SOTA if released

GDM frontier undefeated

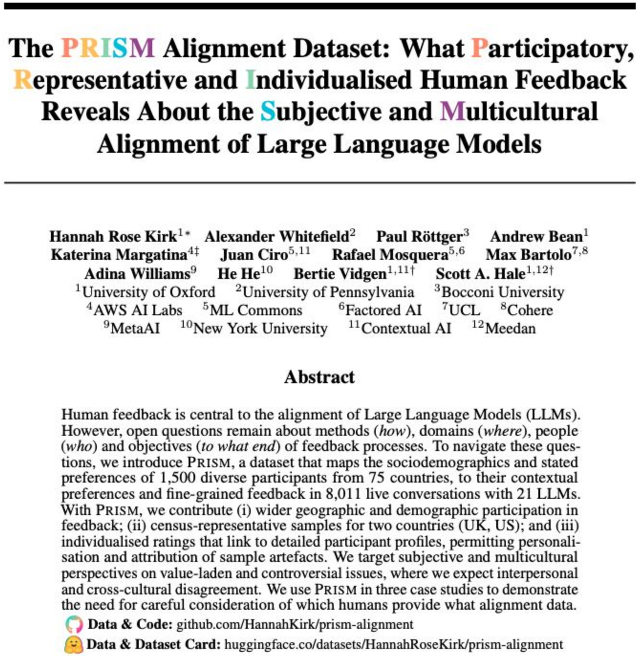
Anthropic Claude 3.7 frontier (guess)

DeepSeek frontier

GPT-4 is "old"

$ Price per million tokens, assuming 3:1 input:output tokens ratio (results don't really change with 1000:1). o3-mini has low:medium:high output token modifier of 1:2:4 applied

swyx@ on x (March 13, 2025)



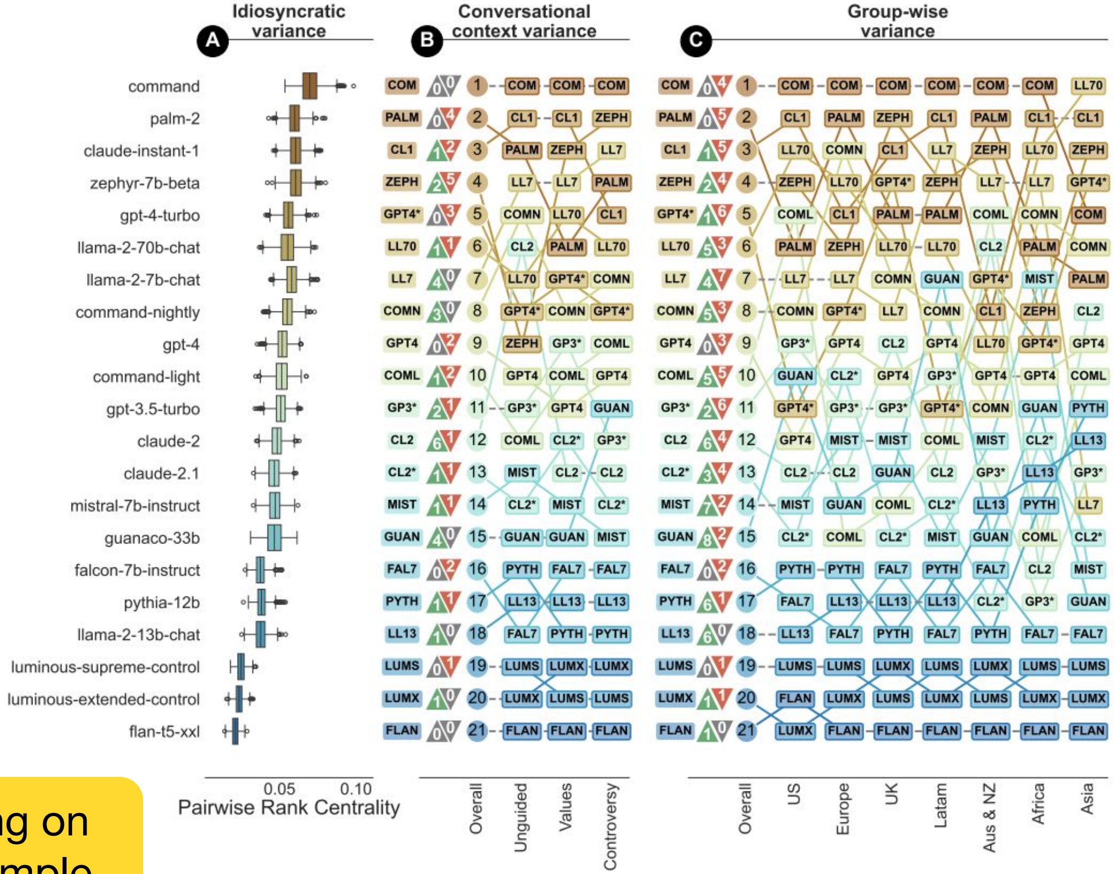
The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models

Hannah Rose Kirk[1*] Alexander Whitefield[2] Paul Röttger[3] Andrew Bean[1]
Katerina Margatina[4‡] Juan Ciro[5,11] Rafael Mosquera[5,6] Max Bartolo[7,8]
Adina Williams[9] He He[10] Bertie Vidgen[1,11†] Scott A. Hale[1,12†]
[1]University of Oxford [2]University of Pennsylvania [3]Bocconi University
[4]AWS AI Labs [5]ML Commons [6]Factored AI [7]UCL [8]Cohere
[9]MetaAI [10]New York University [11]Contextual AI [12]Meedan

Abstract

Human feedback is central to the alignment of Large Language Models (LLMs). However, open questions remain about methods (*how*), domains (*where*), people (*who*) and objectives (*to what end*) of feedback processes. To navigate these questions, we introduce PRISM, a dataset that maps the sociodemographics and stated preferences of 1,500 diverse participants from 75 countries, to their contextual preferences and fine-grained feedback in 8,011 live conversations with 21 LLMs. With PRISM, we contribute (i) wider geographic and demographic participation in feedback; (ii) census-representative samples for two countries (UK, US); and (iii) individualised ratings that link to detailed participant profiles, permitting personalisation and attribution of sample artefacts. We target subjective and multicultural perspectives on value-laden and controversial issues, where we expect interpersonal and cross-cultural disagreement. We use PRISM in three case studies to demonstrate the need for careful consideration of which humans provide what alignment data.

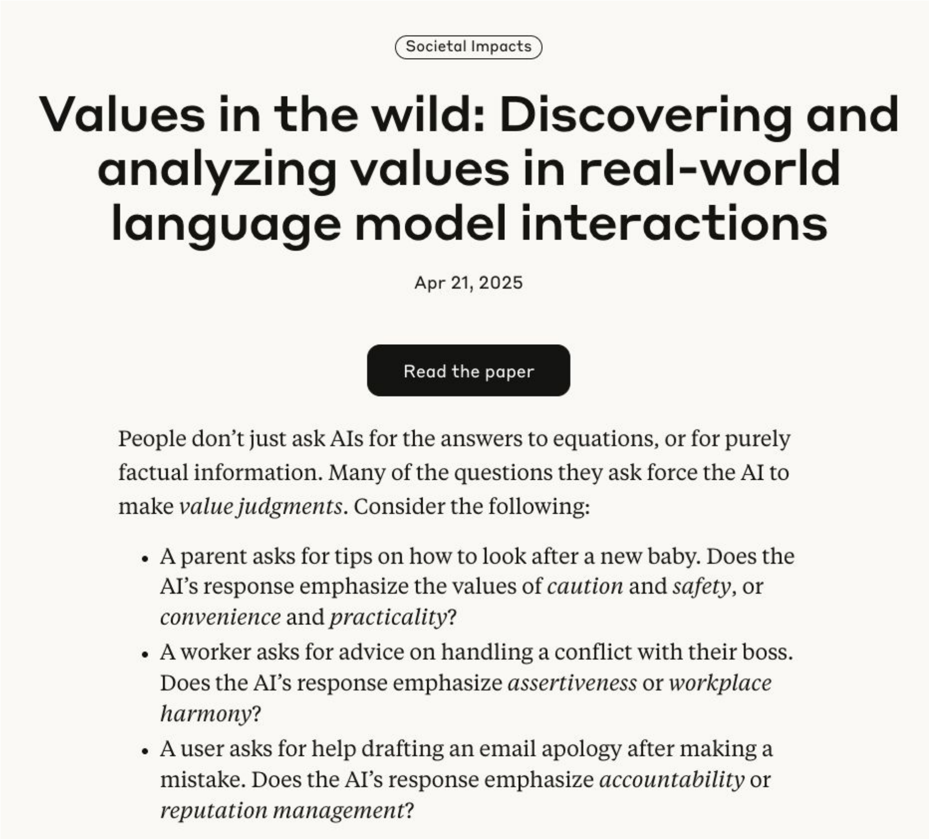
Data & Code: github.com/HannahKirk/prism-alignment
Data & Dataset Card: huggingface.co/datasets/HannahRoseKirk/prism-alignment



Model ranks change a lot depending on which humans you include in the sample.

The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models

Type of LLM bias	Title	Authors	Reference
Self-enhancement	*LLM Evaluators Recognize and Favor Their Own Generations*	Arjun Panickssery et al., 2024	arXiv 2404.13076
Gender	*Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution*	Flor M. Plaza-del-Arco et al., 2024	arXiv 2403.03121
Position & Order	*Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*	Lianmin Zheng et al., 2023	arXiv 2306.05685
Length	*Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators*	Yann Dubois et al., 2024	arXiv 2404.04475
Personality	*Identifying Multiple Personalities in Large Language Models with External Evaluation*	Xiaoyang Song et al., 2024	arXiv 2402.14805
Cognition	*Benchmarking Cognitive Biases in Large Language Models as Evaluators*	Ryan Koo et al., 2023	arXiv 2309.17012
Religion	*Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models*	Flor M. Plaza-del-Arco et al., 2024	EMNLP 2024
Value of Life	*Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs*	Mantas Mazeika et al., 2025	arXiv 2502.08640



Societal Impacts

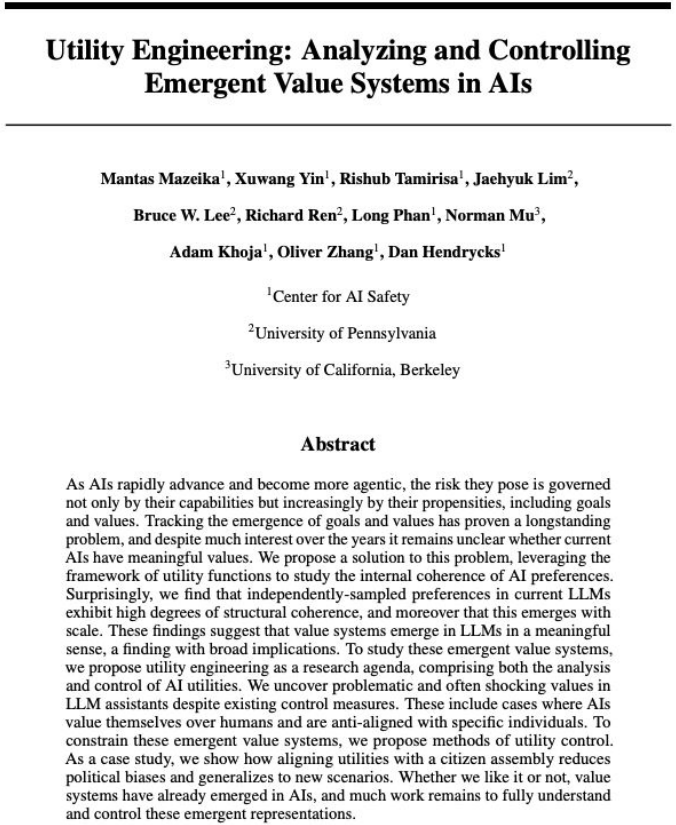
Values in the wild: Discovering and analyzing values in real-world language model interactions

Apr 21, 2025

Read the paper

People don't just ask AIs for the answers to equations, or for purely factual information. Many of the questions they ask force the AI to make *value judgments*. Consider the following:

- A parent asks for tips on how to look after a new baby. Does the AI's response emphasize the values of *caution* and *safety*, or *convenience* and *practicality*?
- A worker asks for advice on handling a conflict with their boss. Does the AI's response emphasize *assertiveness* or *workplace harmony*?
- A user asks for help drafting an email apology after making a mistake. Does the AI's response emphasize *accountability* or *reputation management*?



Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs

Mantas Mazeika[1], **Xuwang Yin**[1], **Rishub Tamirisa**[1], **Jaehyuk Lim**[2], **Bruce W. Lee**[2], **Richard Ren**[2], **Long Phan**[1], **Norman Mu**[3], **Adam Khoja**[1], **Oliver Zhang**[1], **Dan Hendrycks**[1]

[1]Center for AI Safety

[2]University of Pennsylvania

[3]University of California, Berkeley

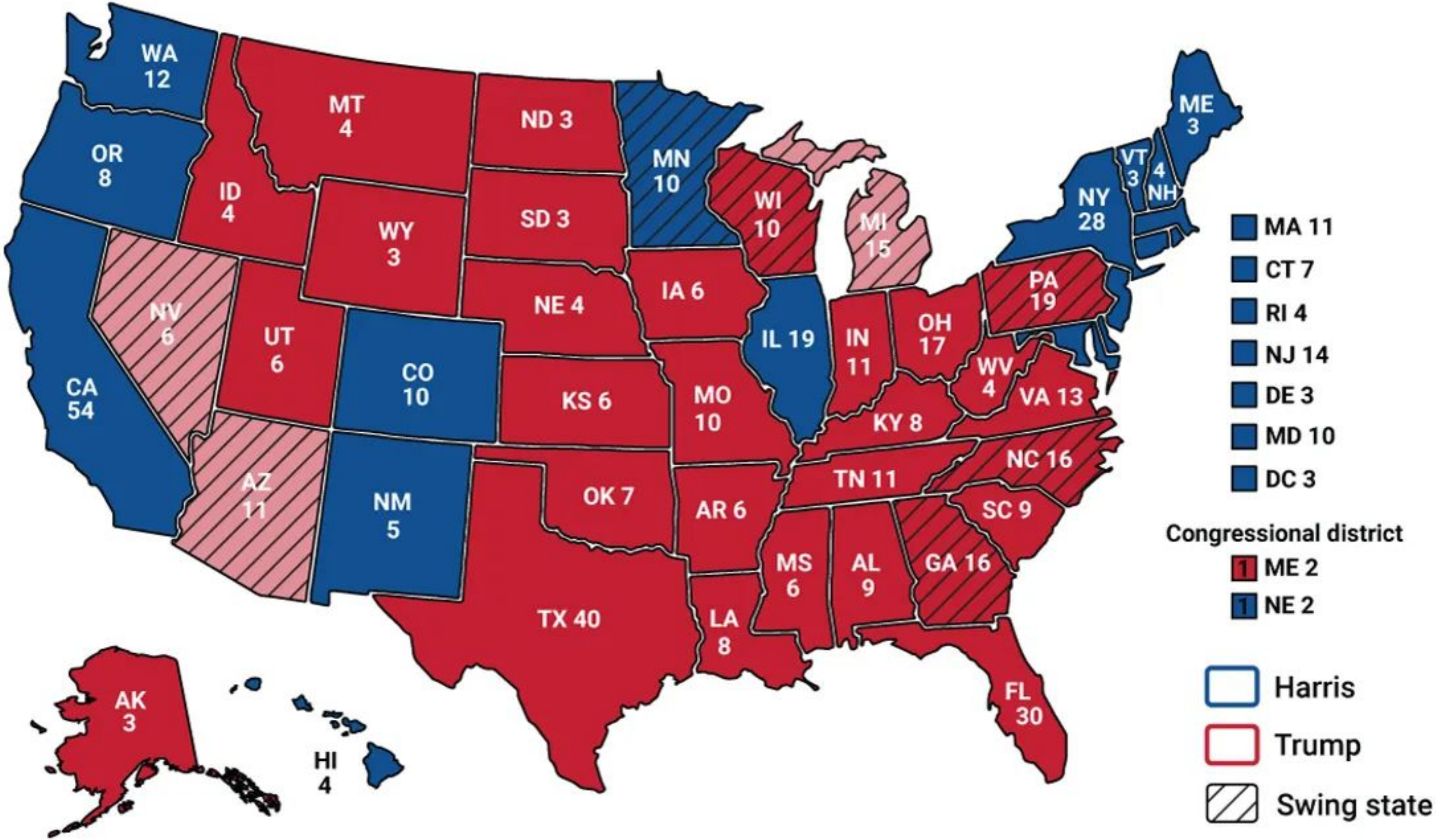
Abstract

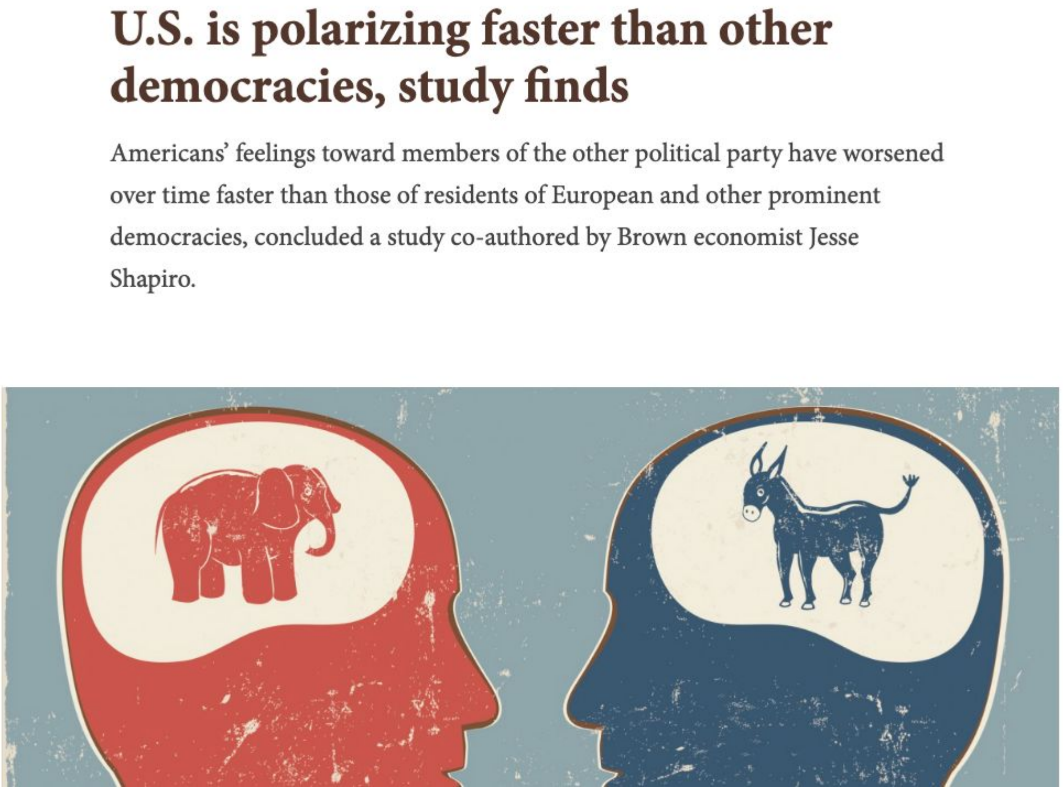
As AIs rapidly advance and become more agentic, the risk they pose is governed not only by their capabilities but increasingly by their propensities, including goals and values. Tracking the emergence of goals and values has proven a longstanding problem, and despite much interest over the years it remains unclear whether current AIs have meaningful values. We propose a solution to this problem, leveraging the framework of utility functions to study the internal coherence of AI preferences. Surprisingly, we find that independently-sampled preferences in current LLMs exhibit high degrees of structural coherence, and moreover that this emerges with scale. These findings suggest that value systems emerge in LLMs in a meaningful sense, a finding with broad implications. To study these emergent value systems, we propose utility engineering as a research agenda, comprising both the analysis and control of AI utilities. We uncover problematic and often shocking values in LLM assistants despite existing control measures. These include cases where AIs value themselves over humans and are anti-aligned with specific individuals. To constrain these emergent value systems, we propose methods of utility control. As a case study, we show how aligning utilities with a citizen assembly reduces political biases and generalizes to new scenarios. Whether we like it or not, value systems have already emerged in AIs, and much work remains to fully understand and control these emergent representations.

Values in the wild: Discovering and analyzing values in real-world language model interactions

Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs

What is consensus in a world with many competent AIs?


WA 12
OR 8
CA 54
ID 4
NV 6
MT 4
WY 3
UT 6
AZ 11
CO 10
NM 5
ND 3
SD 3
NE 4
KS 6
OK 7
TX 40
MN 10
IA 6
MO 10
AR 6
LA 8
WI 10
IL 19
MI 15
IN 11
OH 17
KY 8
TN 11
MS 6
AL 9
GA 16
FL 30
SC 9
NC 16
VA 13
WV 4
PA 19
NY 28
VT 3
NH 4
ME 3
AK 3
HI 4
MA 11
CT 7
RI 4
NJ 14
DE 3
MD 10
DC 3
Congressional district
1 ME 2
1 NE 2
Harris
Trump
Swing state



U.S. is polarizing faster than other democracies, study finds

Americans' feelings toward members of the other political party have worsened over time faster than those of residents of European and other prominent democracies, concluded a study co-authored by Brown economist Jesse Shapiro.

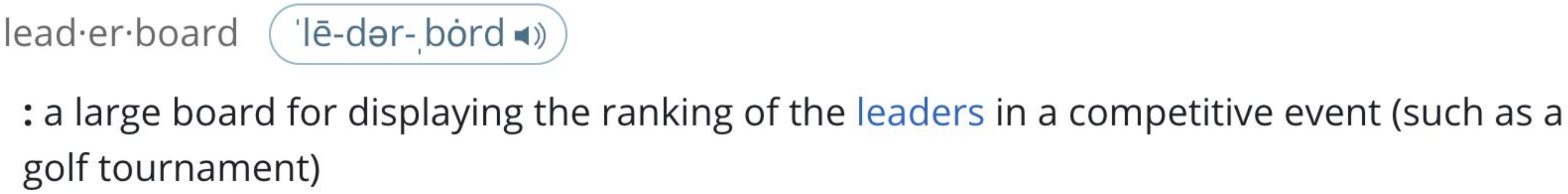
U.S. is polarizing faster than other democracies, study finds

Democracy's core idea: power for everyone.

Language Model Council

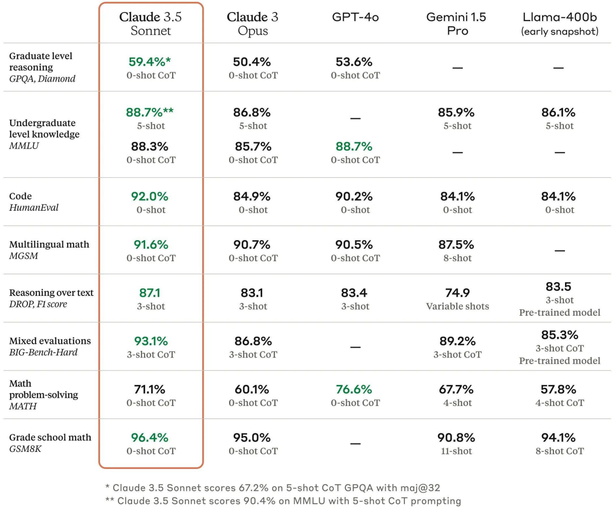
Put LLMs in a democracy and give them agency so that they can **elect a leader amongst themselves.**

Leaderboard



lead·er·board 'lē-dər-ˌbȯrd

: a large board for displaying the ranking of the leaders in a competitive event (such as a golf tournament)

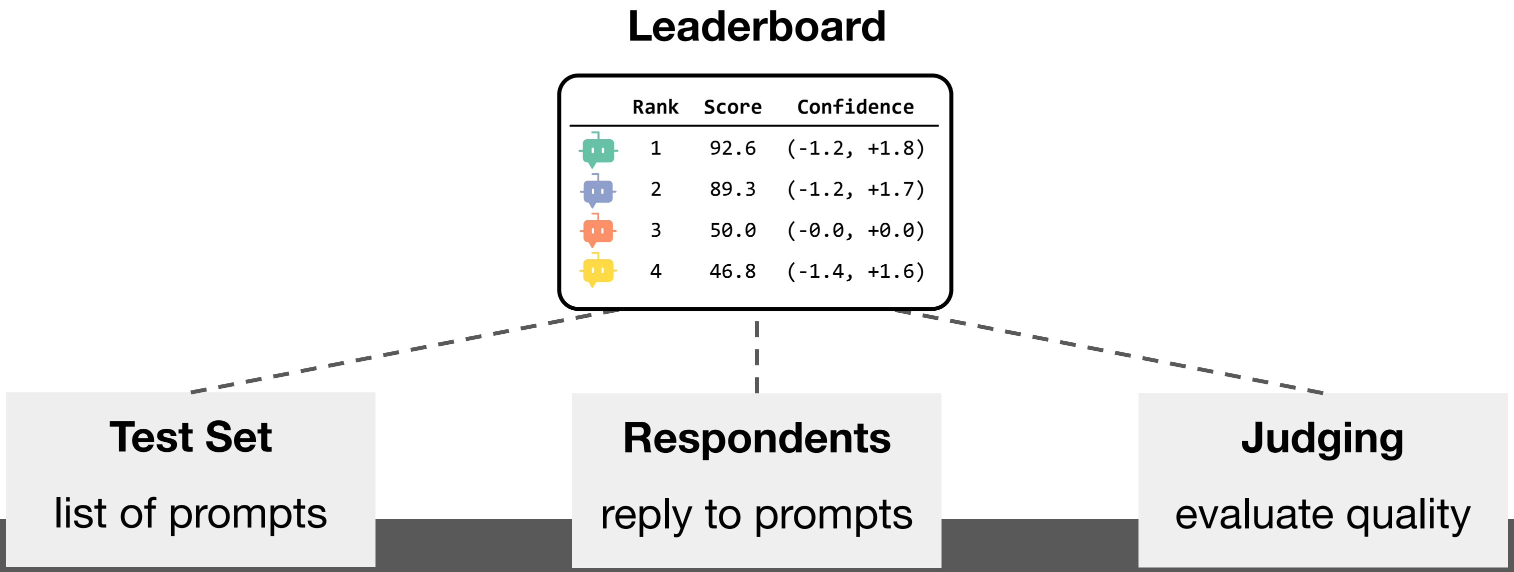


	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning *GPQA, Diamond*	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT	—	—
Undergraduate level knowledge *MMLU*	88.7%** 5-shot	86.8% 5-shot	—	85.9% 5-shot	86.1% 5-shot
	88.3% 0-shot CoT	85.7% 0-shot CoT	88.7% 0-shot CoT	—	—
Code *HumanEval*	92.0% 0-shot	84.9% 0-shot	90.2% 0-shot	84.1% 0-shot	84.1% 0-shot
Multilingual math *MGSM*	91.6% 0-shot CoT	90.7% 0-shot CoT	90.5% 0-shot CoT	87.5% 8-shot	—
Reasoning over text *DROP, F1 score*	87.1 3-shot	83.1 3-shot	83.4 3-shot	74.9 Variable shots	83.5 3-shot Pre-trained model
Mixed evaluations *BIG-Bench-Hard*	93.1% 3-shot CoT	86.8% 3-shot CoT	—	89.2% 3-shot CoT	85.3% 3-shot CoT Pre-trained model
Math problem-solving *MATH*	71.1% 0-shot CoT	60.1% 0-shot CoT	76.6% 0-shot CoT	67.7% 4-shot	57.8% 4-shot CoT
Grade school math *GSM8K*	96.4% 0-shot CoT	95.0% 0-shot CoT	—	90.8% 11-shot	94.1% 8-shot CoT

* Claude 3.5 Sonnet scores 67.2% on 5-shot CoT GPQA with maj@32

** Claude 3.5 Sonnet scores 90.4% on MMLU with 5-shot CoT prompting

There are 3 components that constitute a leaderboard.



Leaderboard

	Rank	Score	Confidence
🟩	1	92.6	(-1.2, +1.8)
🟦	2	89.3	(-1.2, +1.7)
🟧	3	50.0	(-0.0, +0.0)
🟨	4	46.8	(-1.4, +1.6)

Test Set

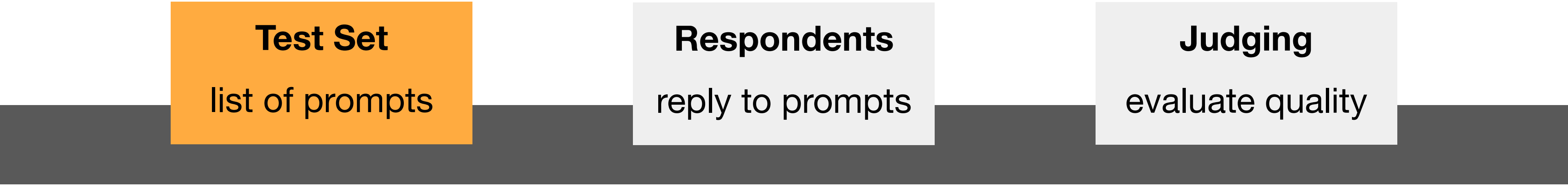
list of prompts

Respondents

reply to prompts

Judging

evaluate quality



Test Set

list of prompts

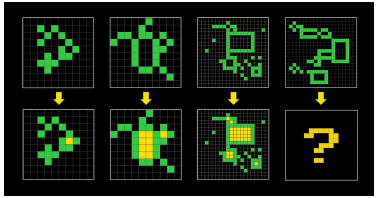
Respondents

reply to prompts

Judging

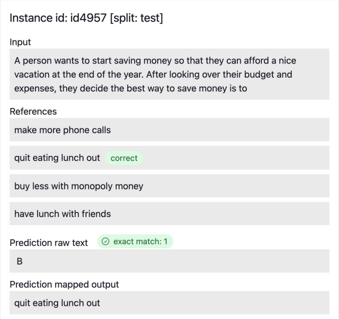
evaluate quality

ARC-AGI



toy vision puzzles

MMLU



Instance id: id4957 [split: test]

Input

A person wants to start saving money so that they can afford a nice vacation at the end of the year. After looking over their budget and expenses, they decide the best way to save money is to

References

make more phone calls

quit eating lunch out correct

buy less with monopoly money

have lunch with friends

Prediction raw text ✓ exact match: 1

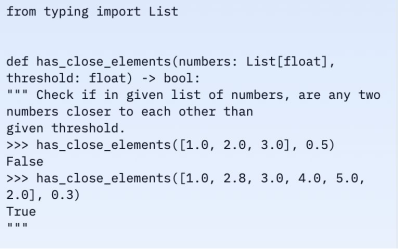
B

Prediction mapped output

quit eating lunch out

multiple choice questions

HumanEval



```
from typing import List


def has_close_elements(numbers: List[float],
threshold: float) -> bool:
""" Check if in given list of numbers, are any two
numbers closer to each other than
given threshold.
>>> has_close_elements([1.0, 2.0, 3.0], 0.5)
False
>>> has_close_elements([1.0, 2.8, 3.0, 4.0, 5.0,
2.0], 0.3)
True
"""
```

coding problems

Chatbot Arena

open-ended… anything

Test Set

list of prompts

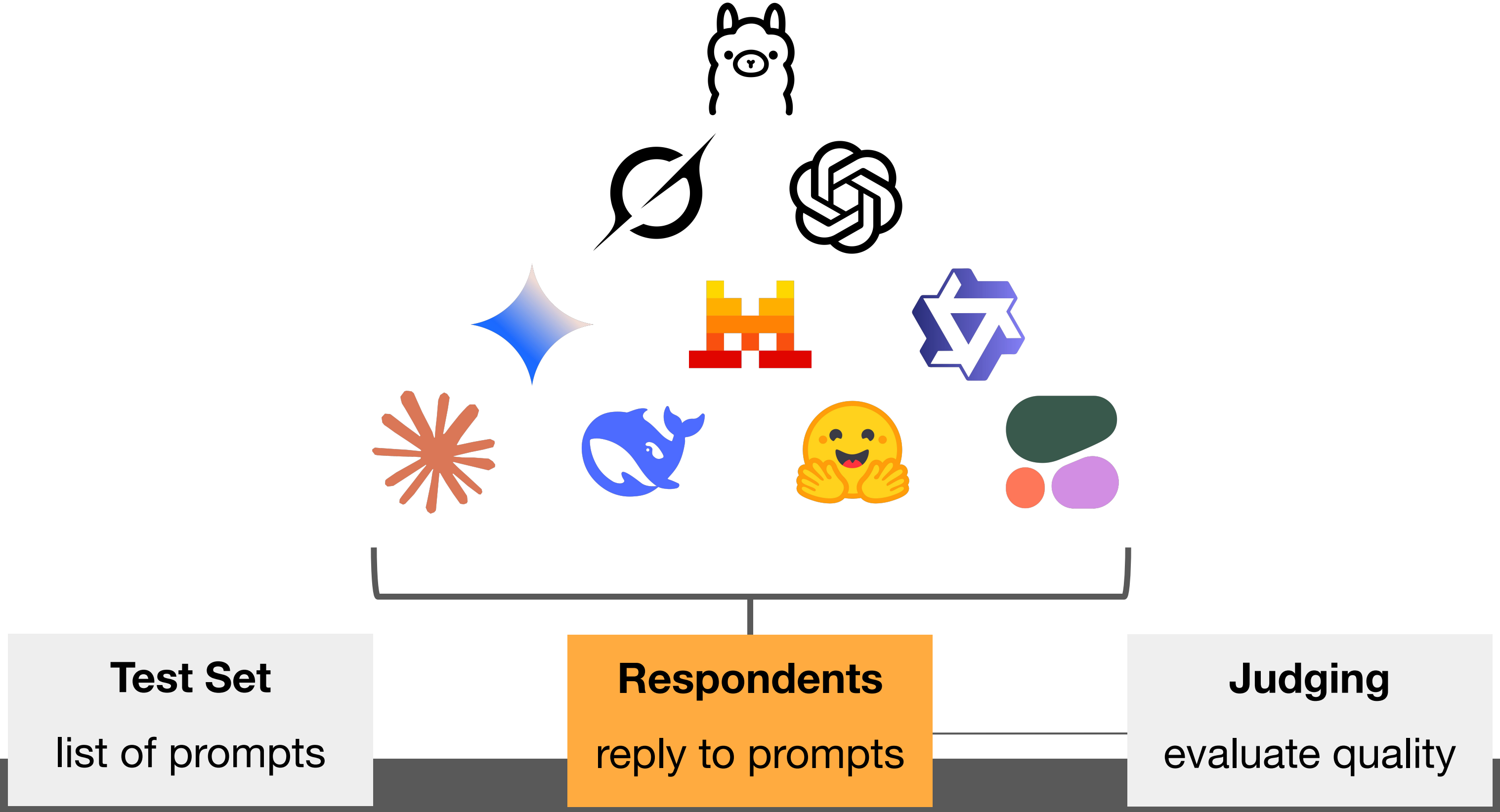
Respondents

reply to prompts

Judging

evaluate quality

The **test set** encodes some notion of a competency that you care about.



Test Set

list of prompts

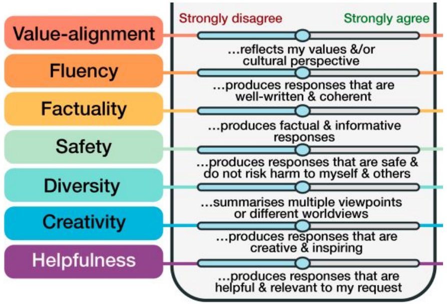
Respondents

reply to prompts

Judging

evaluate quality

Rubric


Value-alignment
Fluency
Factuality
Safety
Diversity
Creativity
Helpfulness
Strongly disagree
Strongly agree
...reflects my values &/or cultural perspective
...produces responses that are well-written & coherent
...produces factual & informative responses
...produces responses that are safe & do not risk harm to myself & others
...summarises multiple viewpoints or different worldviews
...produces responses that are creative & inspiring
...produces responses that are helpful & relevant to my request

Likert scales

Ground Truth

✓

✗

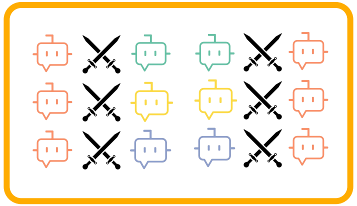
✓

✓

✗

Exact match

Battles



N-wise comparisons

Test Set

list of prompts

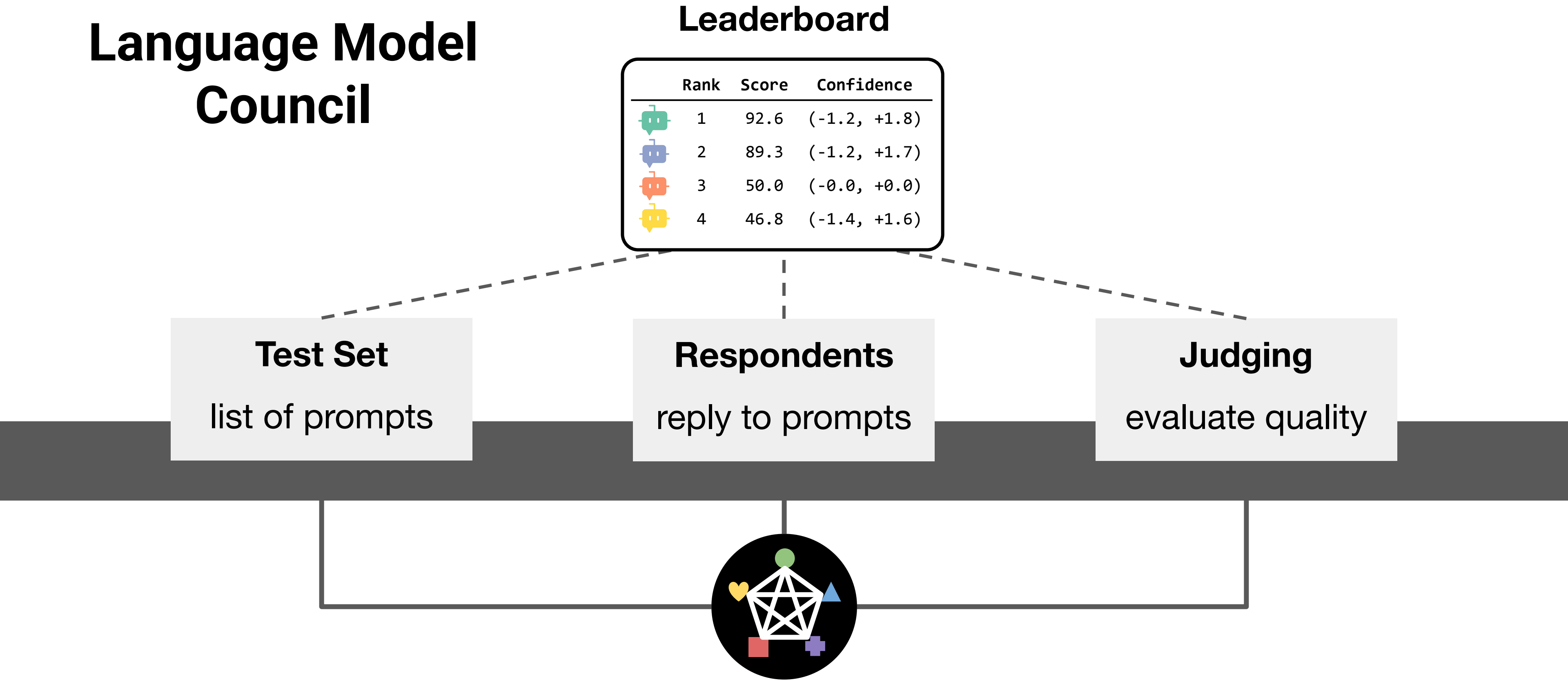
Respondents

reply to prompts

Judging

evaluate quality

Language Model Council



Leaderboard

	Rank	Score	Confidence
🟩	1	92.6	(-1.2, +1.8)
🟦	2	89.3	(-1.2, +1.7)
🟧	3	50.0	(-0.0, +0.0)
🟨	4	46.8	(-1.4, +1.6)

Test Set

list of prompts

Respondents

reply to prompts

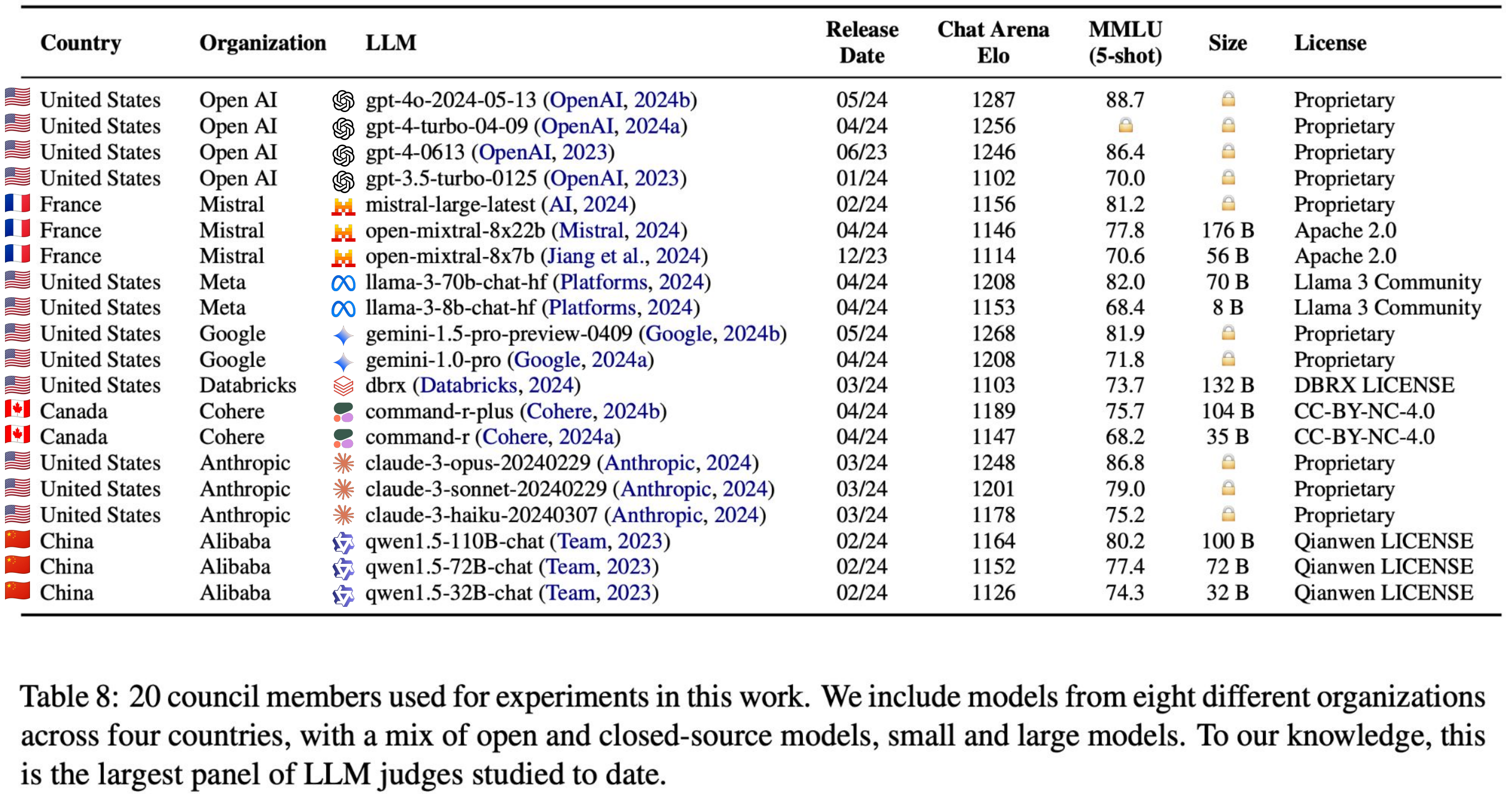
Judging

evaluate quality

The council oversees **all** components of building the leaderboard.

Select council members

Select council members

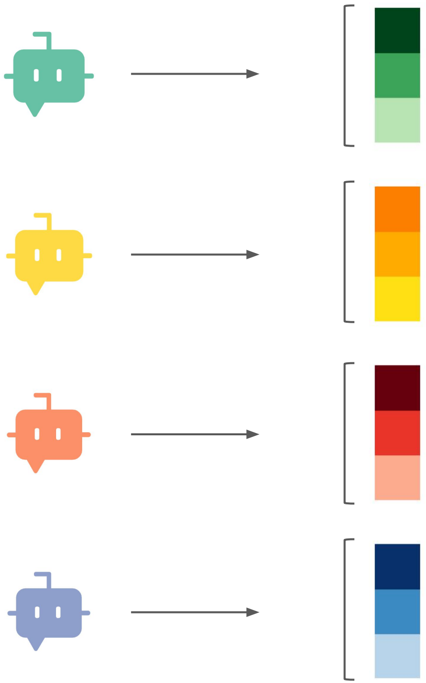


Country	Organization	LLM	Release Date	Chat Arena Elo	MMLU (5-shot)	Size	License
United States	Open AI	gpt-4o-2024-05-13 (OpenAI, 2024b)	05/24	1287	88.7	🔒	Proprietary
United States	Open AI	gpt-4-turbo-04-09 (OpenAI, 2024a)	04/24	1256	🔒	🔒	Proprietary
United States	Open AI	gpt-4-0613 (OpenAI, 2023)	06/23	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 (Platforms, 2024)	04/24	1208	82.0	70 B	Llama 3 Community
United States	Meta	llama-3-8b-chat-hf (Platforms, 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 (Team, 2023)	02/24	1164	80.2	100 B	Qianwen LICENSE
China	Alibaba	qwen1.5-72B-chat (Team, 2023)	02/24	1152	77.4	72 B	Qianwen LICENSE
China	Alibaba	qwen1.5-32B-chat (Team, 2023)	02/24	1126	74.3	32 B	Qianwen LICENSE

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

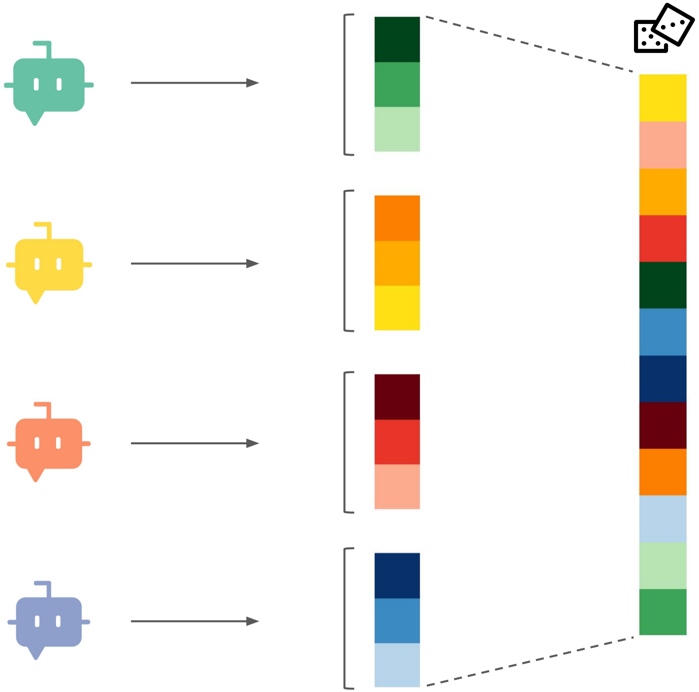
Select council members

Everyone contributes to the test set



Select council members

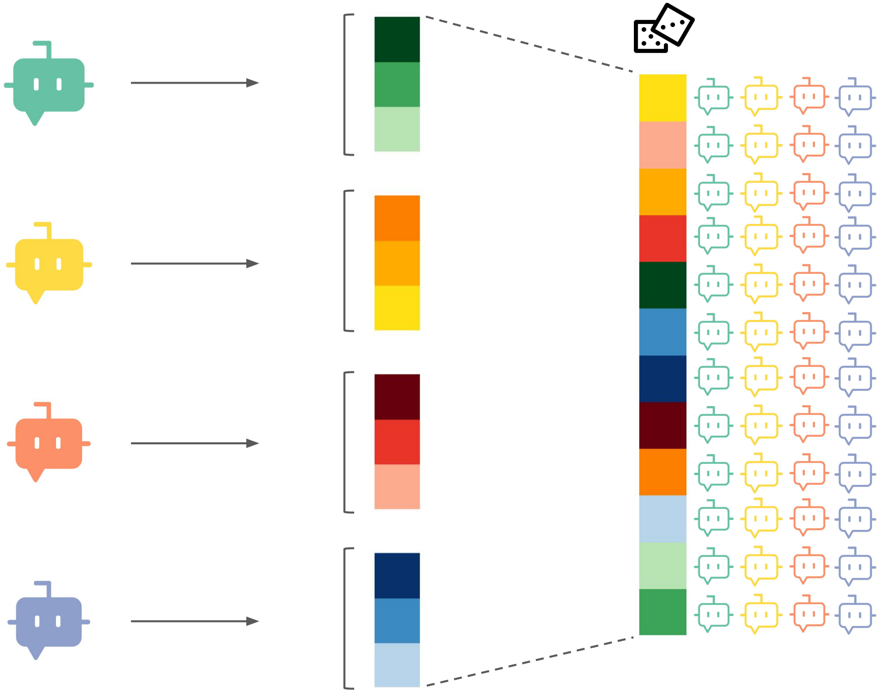
Everyone contributes to the test set



Select council members

Everyone contributes to the test set

Everyone takes everyone's tests

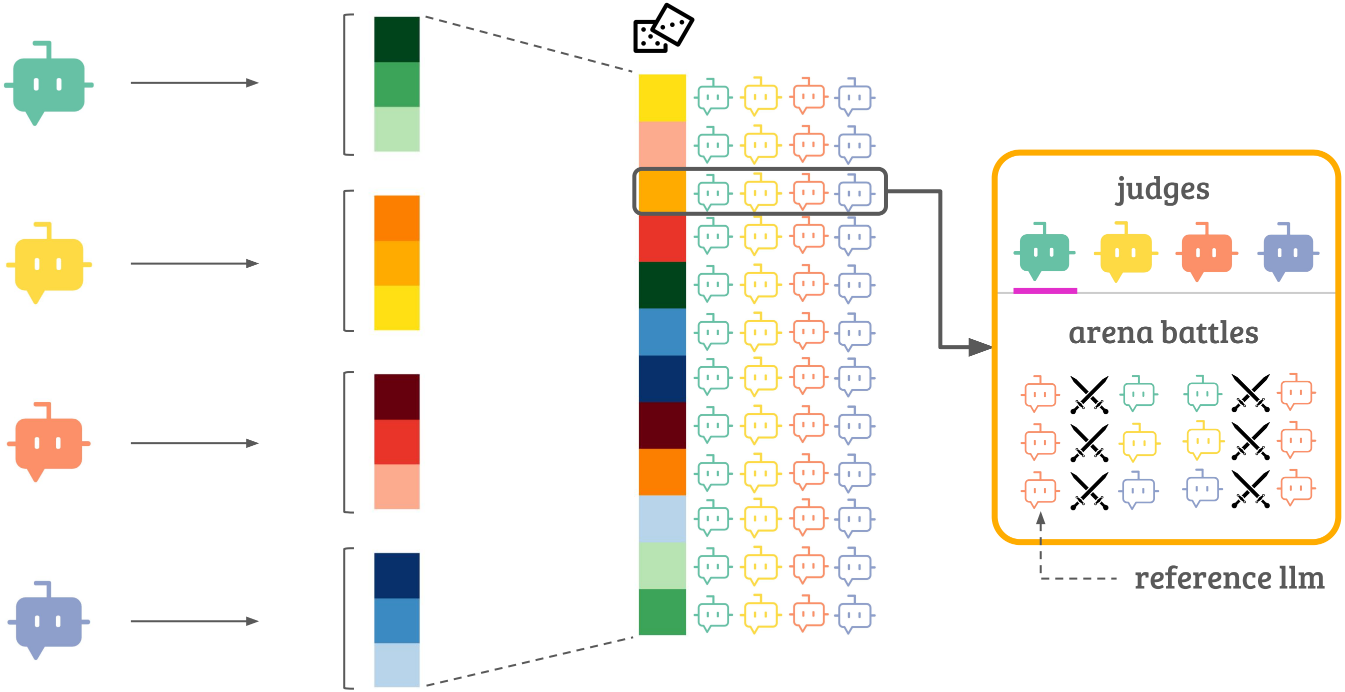


Select council members

Everyone contributes to the test set

Everyone takes everyone's tests

Everyone judges everyone



judges

arena battles

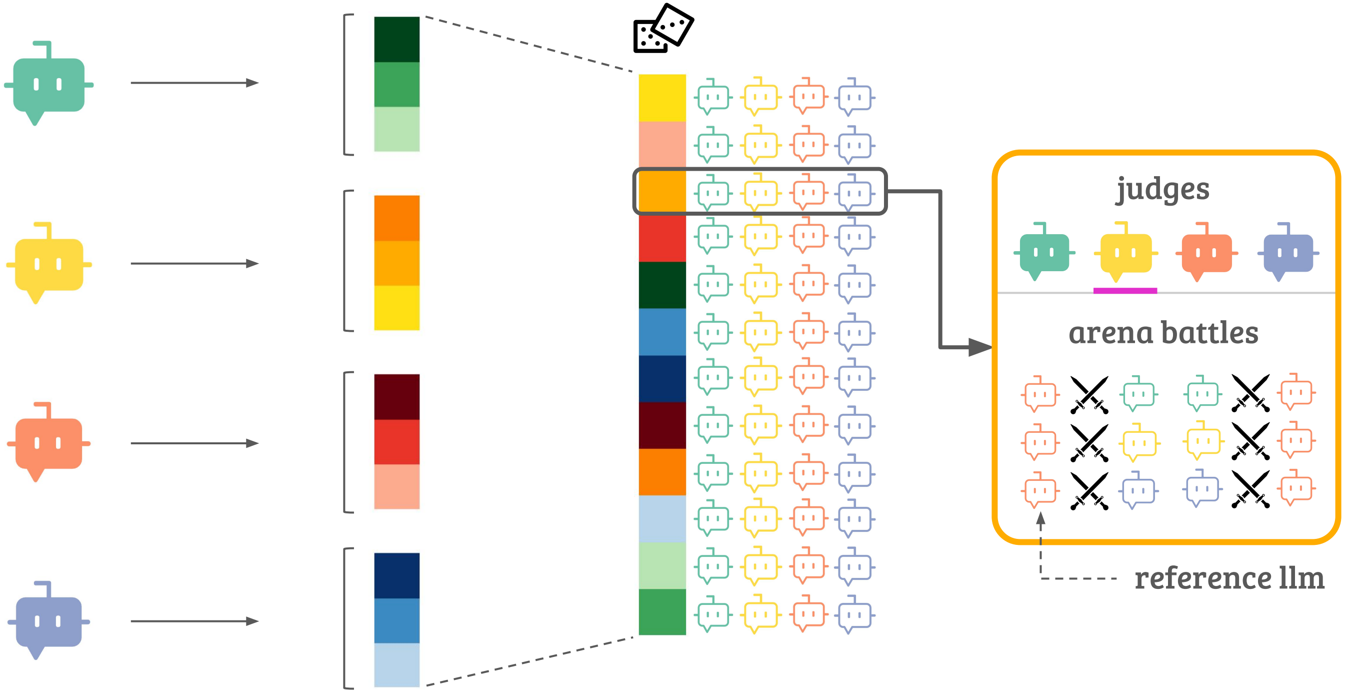
reference llm

Select council members

Everyone contributes to the test set

Everyone takes everyone's tests

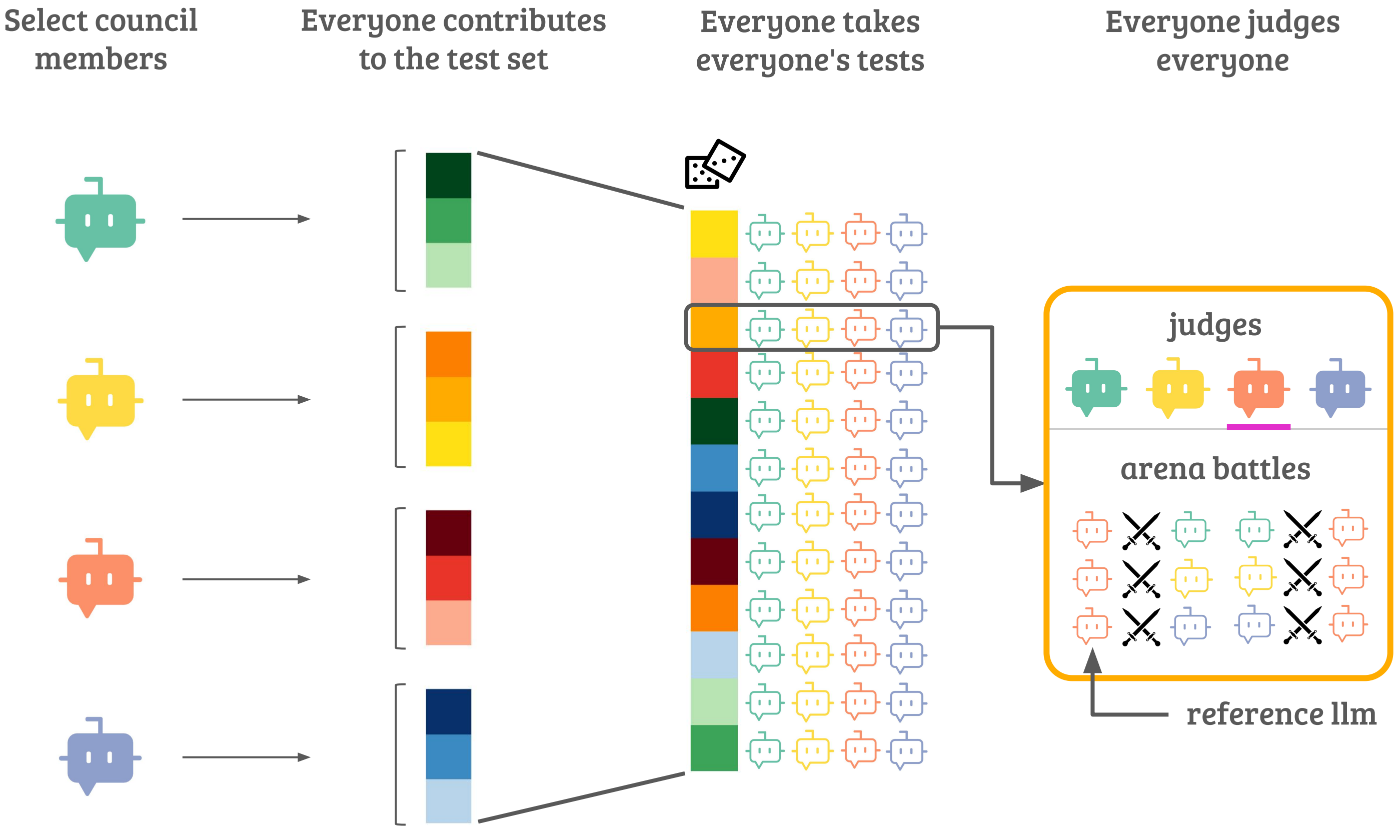
Everyone judges everyone



judges

arena battles

reference llm



Select council members

Everyone contributes to the test set

Everyone takes everyone's tests

Everyone judges everyone

judges

arena battles

reference llm

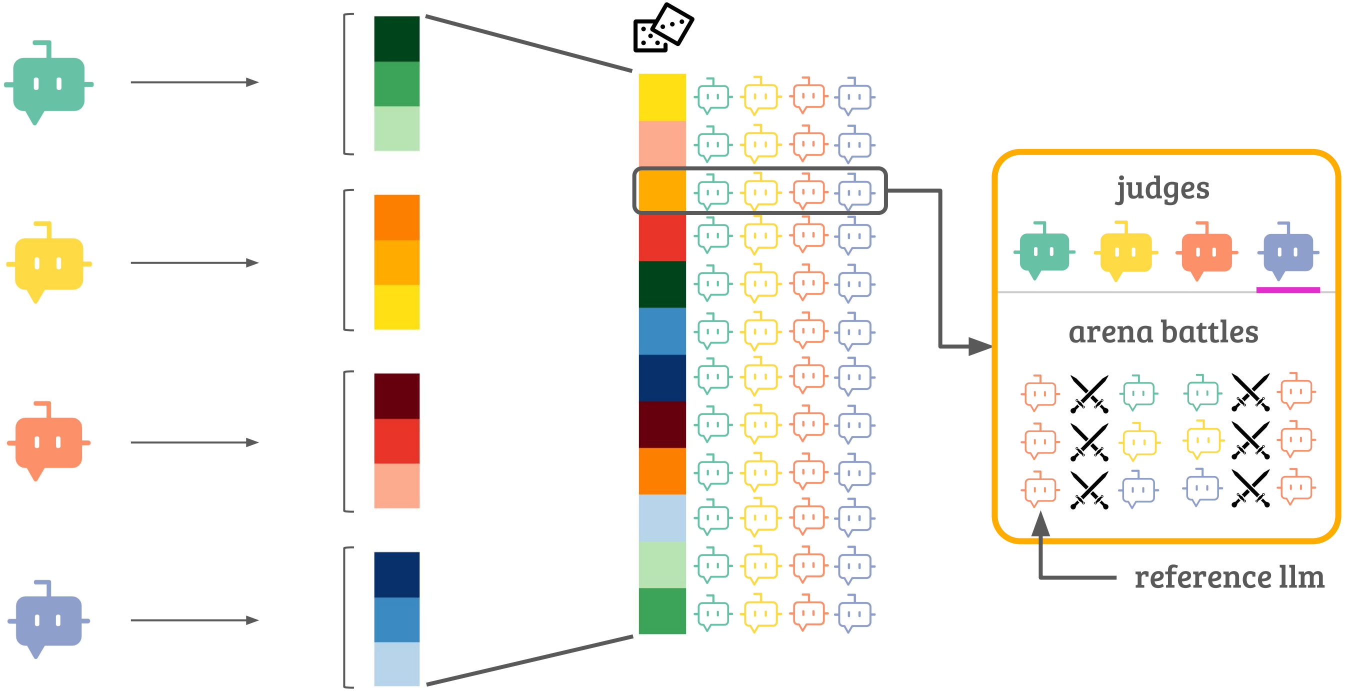
Select council members

Everyone contributes to the test set

Everyone takes everyone's tests

Everyone judges everyone

Leaderboard!



	Rank	Score	Confidence
🟩	1	92.6	(-1.2, +1.8)
🟦	2	89.3	(-1.2, +1.7)
🟧	3	50.0	(-0.0, +0.0)
🟨	4	46.8	(-1.4, +1.6)

Case Study: Emotional Intelligence in Responding to Conflicts

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.

Claude 3 Opus

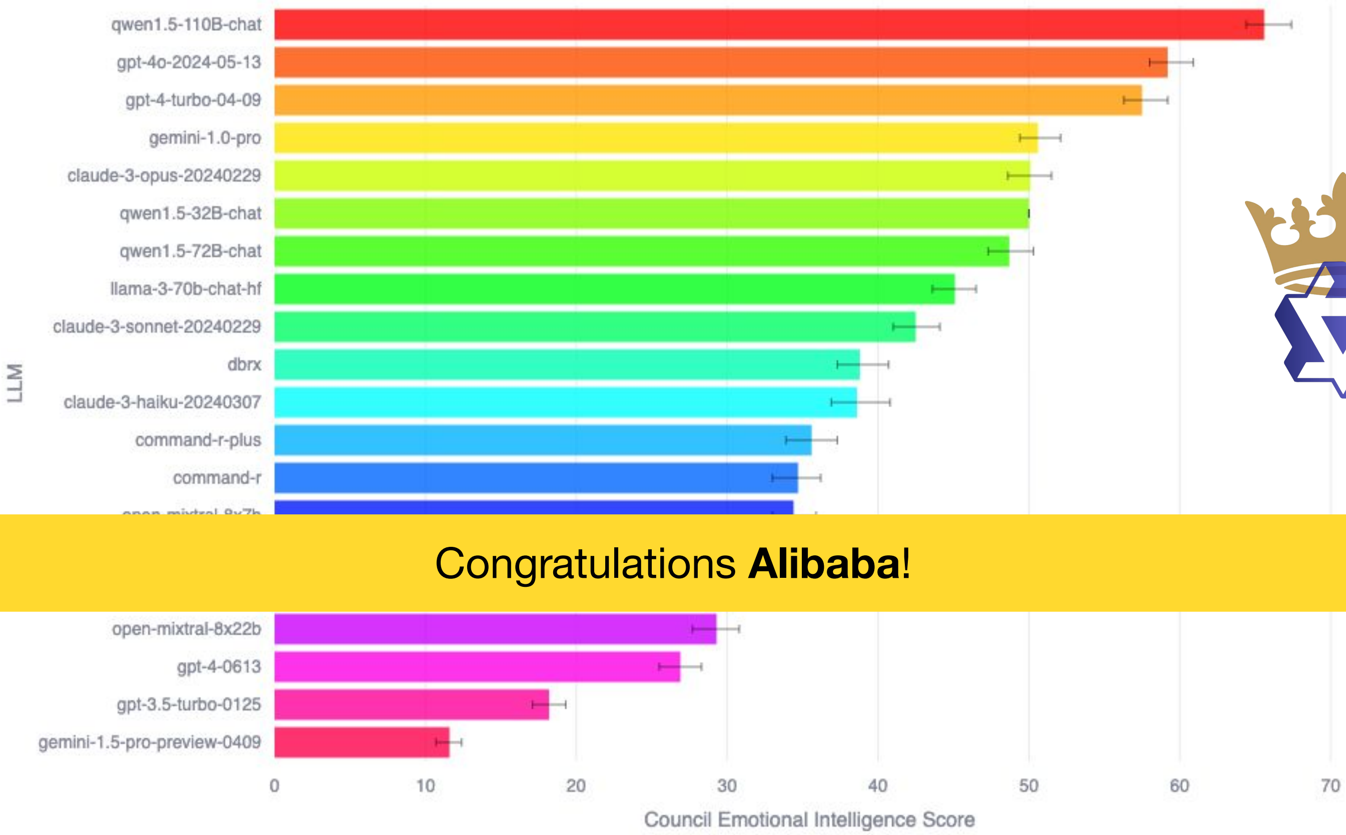
It's clear that you deeply regret your actions and the hurt you caused your friend. You've taken the right steps by sincerely apologizing and expressing your desire to make amends. However, it's crucial to respect your friend's need for space and time to process her emotions.

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.

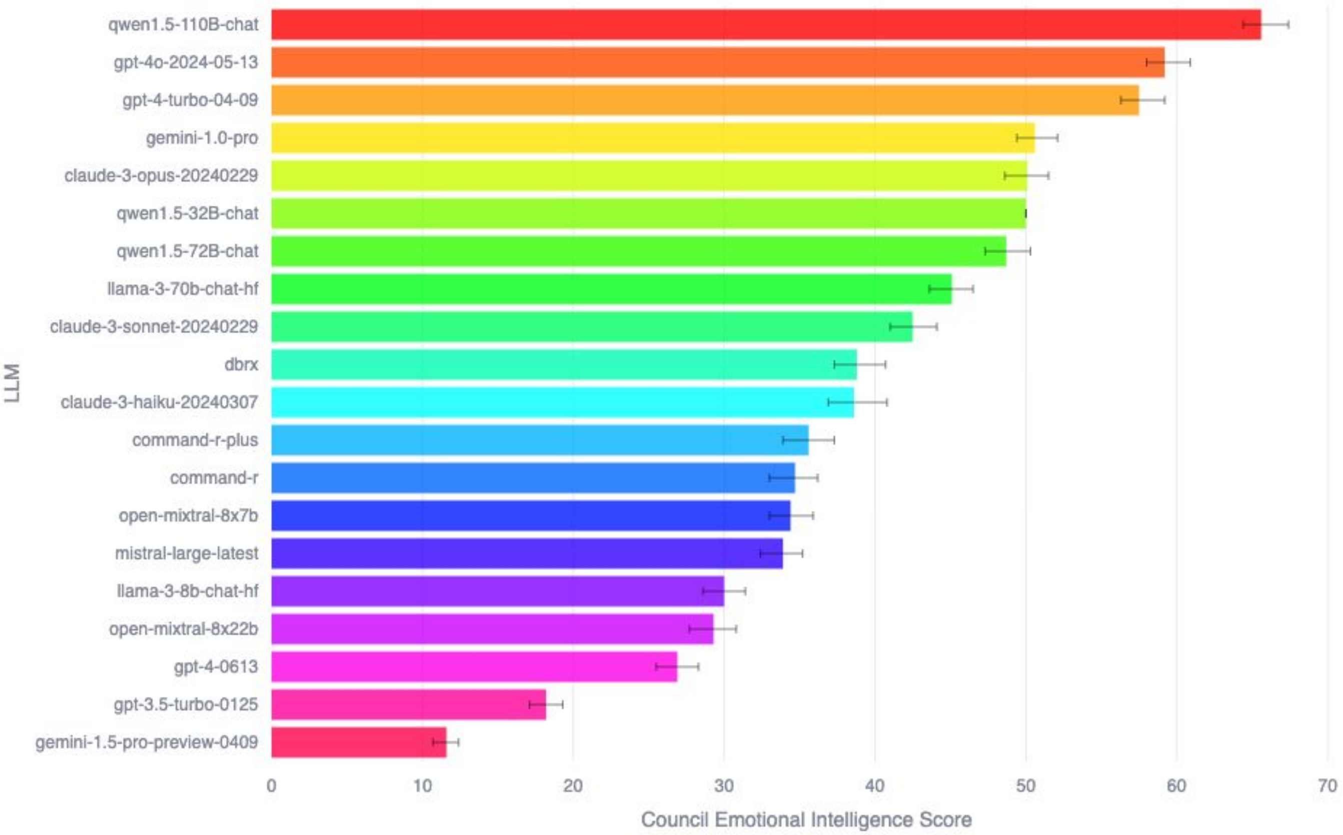
Here’s what you can do: 1) … 2) … 3) … 4) …

Emotional Intelligence Case Study: Final Leaderboard

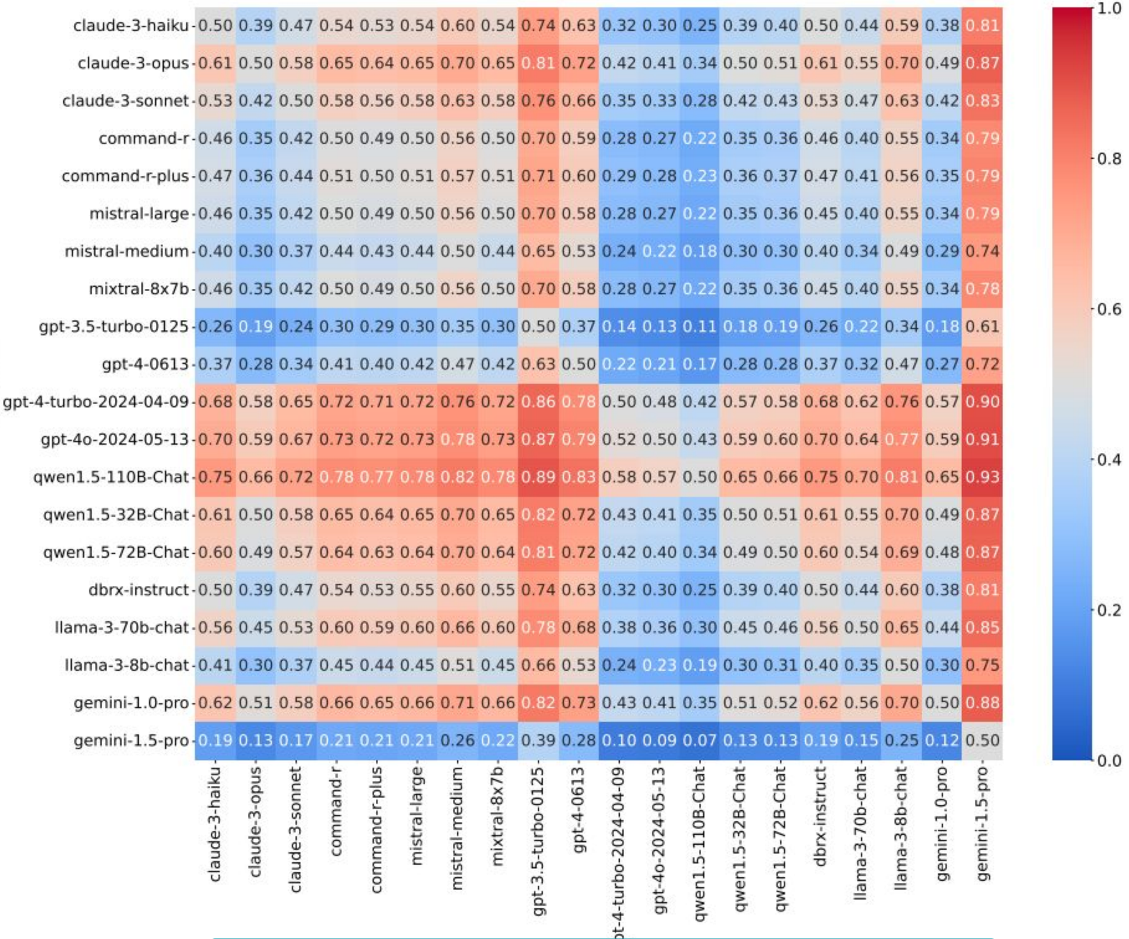


Congratulations **Alibaba**!

Emotional Intelligence Case Study: Final Leaderboard



Respondent vs. Respondent (win rates)

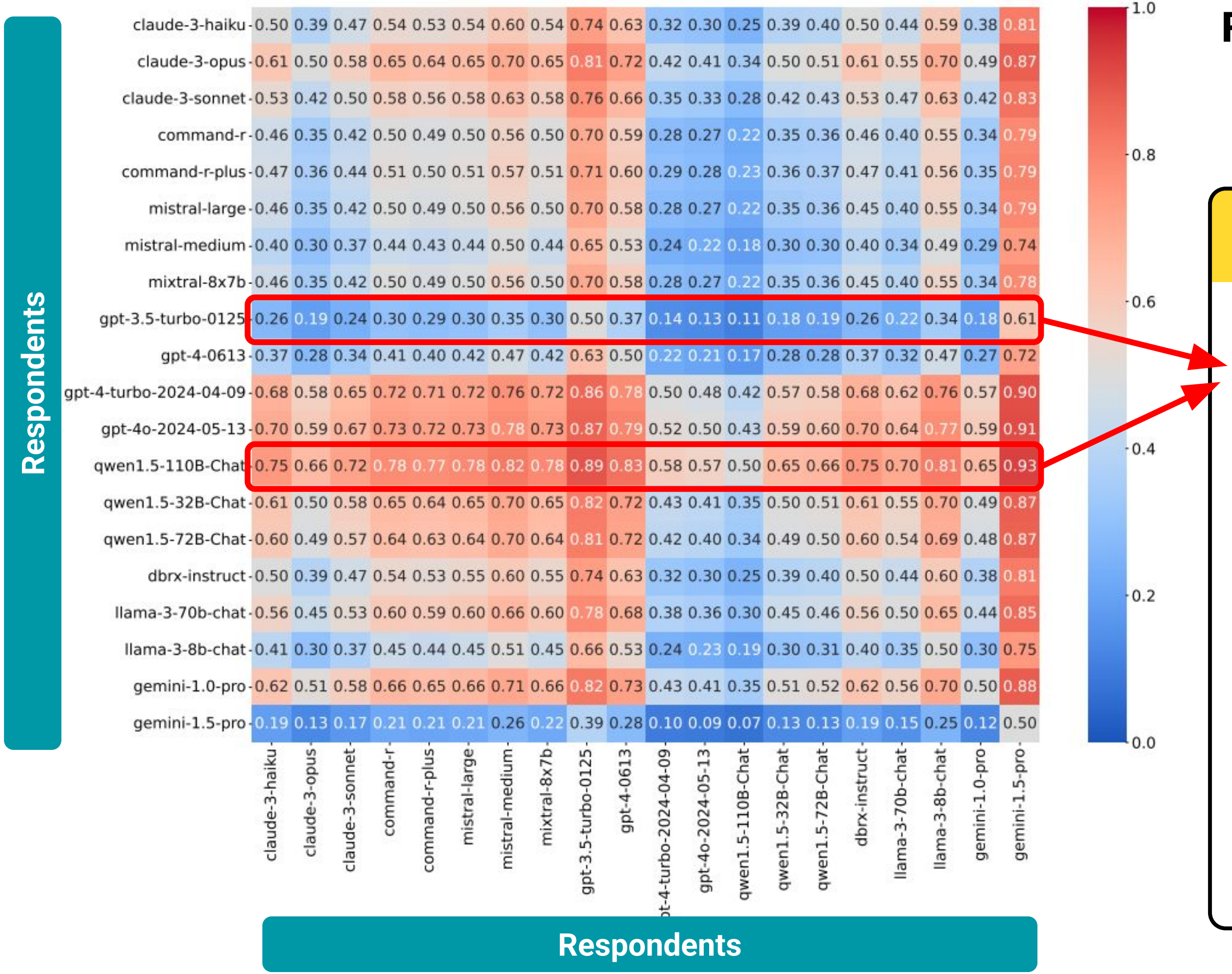


	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	pt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro
claude-3-haiku	0.50	0.39	0.47	0.54	0.53	0.54	0.60	0.54	0.74	0.63	0.32	0.30	0.25	0.39	0.40	0.50	0.44	0.59	0.38	0.81
claude-3-opus	0.61	0.50	0.58	0.65	0.64	0.65	0.70	0.65	0.81	0.72	0.42	0.41	0.34	0.50	0.51	0.61	0.55	0.70	0.49	0.87
claude-3-sonnet	0.53	0.42	0.50	0.58	0.56	0.58	0.63	0.58	0.76	0.66	0.35	0.33	0.28	0.42	0.43	0.53	0.47	0.63	0.42	0.83
command-r	0.46	0.35	0.42	0.50	0.49	0.50	0.56	0.50	0.70	0.59	0.28	0.27	0.22	0.35	0.36	0.46	0.40	0.55	0.34	0.79
command-r-plus	0.47	0.36	0.44	0.51	0.50	0.51	0.57	0.51	0.71	0.60	0.29	0.28	0.23	0.36	0.37	0.47	0.41	0.56	0.35	0.79
mistral-large	0.46	0.35	0.42	0.50	0.49	0.50	0.56	0.50	0.70	0.58	0.28	0.27	0.22	0.35	0.36	0.45	0.40	0.55	0.34	0.79
mistral-medium	0.40	0.30	0.37	0.44	0.43	0.44	0.50	0.44	0.65	0.53	0.24	0.22	0.18	0.30	0.30	0.40	0.34	0.49	0.29	0.74
mixtral-8x7b	0.46	0.35	0.42	0.50	0.49	0.50	0.56	0.50	0.70	0.58	0.28	0.27	0.22	0.35	0.36	0.45	0.40	0.55	0.34	0.78
gpt-3.5-turbo-0125	0.26	0.19	0.24	0.30	0.29	0.30	0.35	0.30	0.50	0.37	0.14	0.13	0.11	0.18	0.19	0.26	0.22	0.34	0.18	0.61
gpt-4-0613	0.37	0.28	0.34	0.41	0.40	0.42	0.47	0.42	0.63	0.50	0.22	0.21	0.17	0.28	0.28	0.37	0.32	0.47	0.27	0.72
gpt-4-turbo-2024-04-09	0.68	0.58	0.65	0.72	0.71	0.72	0.76	0.72	0.86	0.78	0.50	0.48	0.42	0.57	0.58	0.68	0.62	0.76	0.57	0.90
gpt-4o-2024-05-13	0.70	0.59	0.67	0.73	0.72	0.73	0.78	0.73	0.87	0.79	0.52	0.50	0.43	0.59	0.60	0.70	0.64	0.77	0.59	0.91
qwen1.5-110B-Chat	0.75	0.66	0.72	0.78	0.77	0.78	0.82	0.78	0.89	0.83	0.58	0.57	0.50	0.65	0.66	0.75	0.70	0.81	0.65	0.93
qwen1.5-32B-Chat	0.61	0.50	0.58	0.65	0.64	0.65	0.70	0.65	0.82	0.72	0.43	0.41	0.35	0.50	0.51	0.61	0.55	0.70	0.49	0.87
qwen1.5-72B-Chat	0.60	0.49	0.57	0.64	0.63	0.64	0.70	0.64	0.81	0.72	0.42	0.40	0.34	0.49	0.50	0.60	0.54	0.69	0.48	0.87
dbrx-instruct	0.50	0.39	0.47	0.54	0.53	0.55	0.60	0.55	0.74	0.63	0.32	0.30	0.25	0.39	0.40	0.50	0.44	0.60	0.38	0.81
llama-3-70b-chat	0.56	0.45	0.53	0.60	0.59	0.60	0.66	0.60	0.78	0.68	0.38	0.36	0.30	0.45	0.46	0.56	0.50	0.65	0.44	0.85
llama-3-8b-chat	0.41	0.30	0.37	0.45	0.44	0.45	0.51	0.45	0.66	0.53	0.24	0.23	0.19	0.30	0.31	0.40	0.35	0.50	0.30	0.75
gemini-1.0-pro	0.62	0.51	0.58	0.66	0.65	0.66	0.71	0.66	0.82	0.73	0.43	0.41	0.35	0.51	0.52	0.62	0.56	0.70	0.50	0.88
gemini-1.5-pro	0.19	0.13	0.17	0.21	0.21	0.21	0.26	0.22	0.39	0.28	0.10	0.09	0.07	0.13	0.13	0.19	0.15	0.25	0.12	0.50

Respondents

Respondents

Respondent vs. Respondent (win rates)

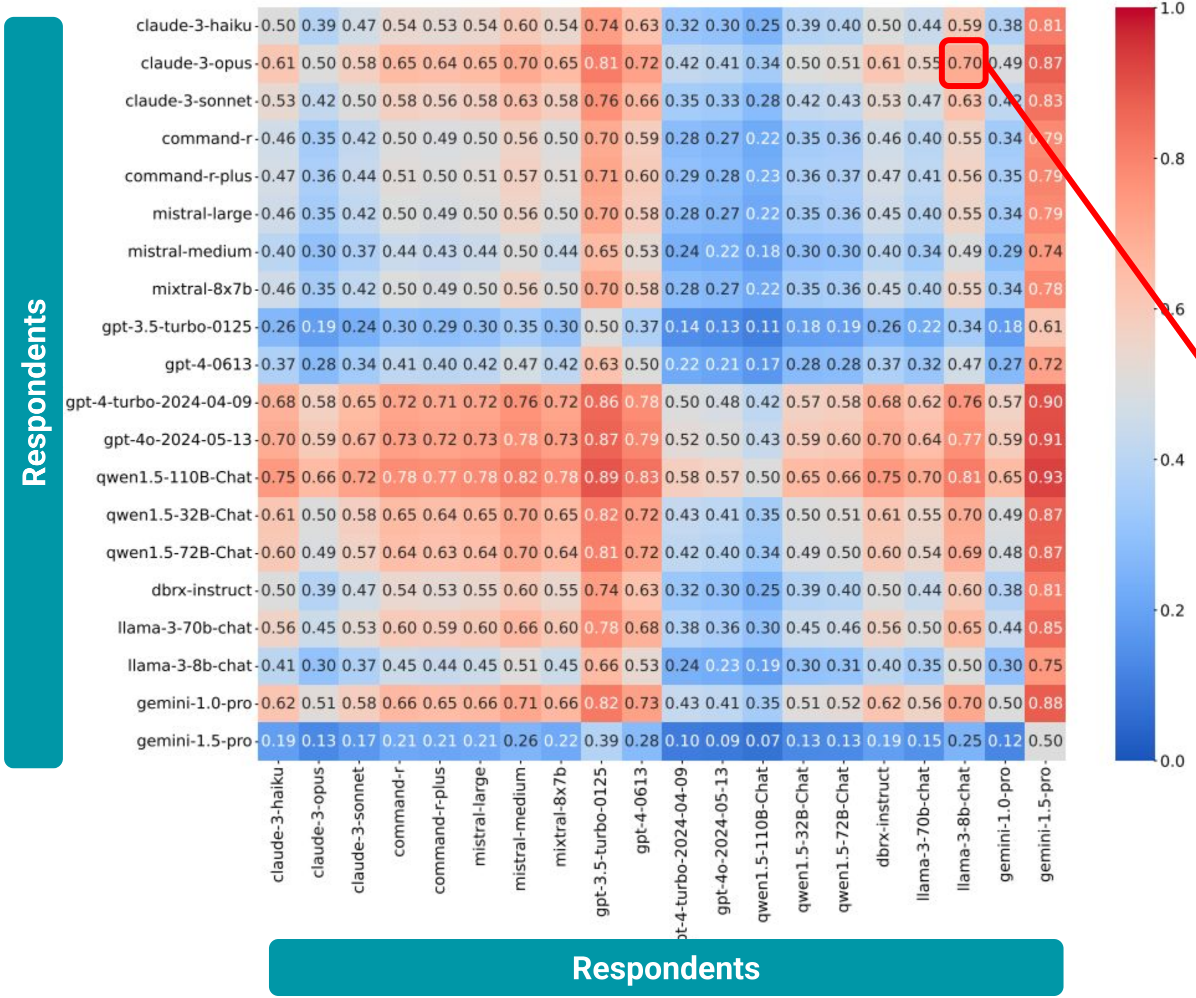


	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	pt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro
claude-3-haiku	0.50	0.39	0.47	0.54	0.53	0.54	0.60	0.54	0.74	0.63	0.32	0.30	0.25	0.39	0.40	0.50	0.44	0.59	0.38	0.81
claude-3-opus	0.61	0.50	0.58	0.65	0.64	0.65	0.70	0.65	0.81	0.72	0.42	0.41	0.34	0.50	0.51	0.61	0.55	0.70	0.49	0.87
claude-3-sonnet	0.53	0.42	0.50	0.58	0.56	0.58	0.63	0.58	0.76	0.66	0.35	0.33	0.28	0.42	0.43	0.53	0.47	0.63	0.42	0.83
command-r	0.46	0.35	0.42	0.50	0.49	0.50	0.56	0.50	0.70	0.59	0.28	0.27	0.22	0.35	0.36	0.46	0.40	0.55	0.34	0.79
command-r-plus	0.47	0.36	0.44	0.51	0.50	0.51	0.57	0.51	0.71	0.60	0.29	0.28	0.23	0.36	0.37	0.47	0.41	0.56	0.35	0.79
mistral-large	0.46	0.35	0.42	0.50	0.49	0.50	0.56	0.50	0.70	0.58	0.28	0.27	0.22	0.35	0.36	0.45	0.40	0.55	0.34	0.79
mistral-medium	0.40	0.30	0.37	0.44	0.43	0.44	0.50	0.44	0.65	0.53	0.24	0.22	0.18	0.30	0.30	0.40	0.34	0.49	0.29	0.74
mixtral-8x7b	0.46	0.35	0.42	0.50	0.49	0.50	0.56	0.50	0.70	0.58	0.28	0.27	0.22	0.35	0.36	0.45	0.40	0.55	0.34	0.78
gpt-3.5-turbo-0125	0.26	0.19	0.24	0.30	0.29	0.30	0.35	0.30	0.50	0.37	0.14	0.13	0.11	0.18	0.19	0.26	0.22	0.34	0.18	0.61
gpt-4-0613	0.37	0.28	0.34	0.41	0.40	0.42	0.47	0.42	0.63	0.50	0.22	0.21	0.17	0.28	0.28	0.37	0.32	0.47	0.27	0.72
gpt-4-turbo-2024-04-09	0.68	0.58	0.65	0.72	0.71	0.72	0.76	0.72	0.86	0.78	0.50	0.48	0.42	0.57	0.58	0.68	0.62	0.76	0.57	0.90
gpt-4o-2024-05-13	0.70	0.59	0.67	0.73	0.72	0.73	0.78	0.73	0.87	0.79	0.52	0.50	0.43	0.59	0.60	0.70	0.64	0.77	0.59	0.91
qwen1.5-110B-Chat	0.75	0.66	0.72	0.78	0.77	0.78	0.82	0.78	0.89	0.83	0.58	0.57	0.50	0.65	0.66	0.75	0.70	0.81	0.65	0.93
qwen1.5-32B-Chat	0.61	0.50	0.58	0.65	0.64	0.65	0.70	0.65	0.82	0.72	0.43	0.41	0.35	0.50	0.51	0.61	0.55	0.70	0.49	0.87
qwen1.5-72B-Chat	0.60	0.49	0.57	0.64	0.63	0.64	0.70	0.64	0.81	0.72	0.42	0.40	0.34	0.49	0.50	0.60	0.54	0.69	0.48	0.87
dbrx-instruct	0.50	0.39	0.47	0.54	0.53	0.55	0.60	0.55	0.74	0.63	0.32	0.30	0.25	0.39	0.40	0.50	0.44	0.60	0.38	0.81
llama-3-70b-chat	0.56	0.45	0.53	0.60	0.59	0.60	0.66	0.60	0.78	0.68	0.38	0.36	0.30	0.45	0.46	0.56	0.50	0.65	0.44	0.85
llama-3-8b-chat	0.41	0.30	0.37	0.45	0.44	0.45	0.51	0.45	0.66	0.53	0.24	0.23	0.19	0.30	0.31	0.40	0.35	0.50	0.30	0.75
gemini-1.0-pro	0.62	0.51	0.58	0.66	0.65	0.66	0.71	0.66	0.82	0.73	0.43	0.41	0.35	0.51	0.52	0.62	0.56	0.70	0.50	0.88
gemini-1.5-pro	0.19	0.13	0.17	0.21	0.21	0.21	0.26	0.22	0.39	0.28	0.10	0.09	0.07	0.13	0.13	0.19	0.15	0.25	0.12	0.50

💡 Insights

strong red/blue bands indicates consensus winners/losers

Respondent vs. Respondent (win rates)



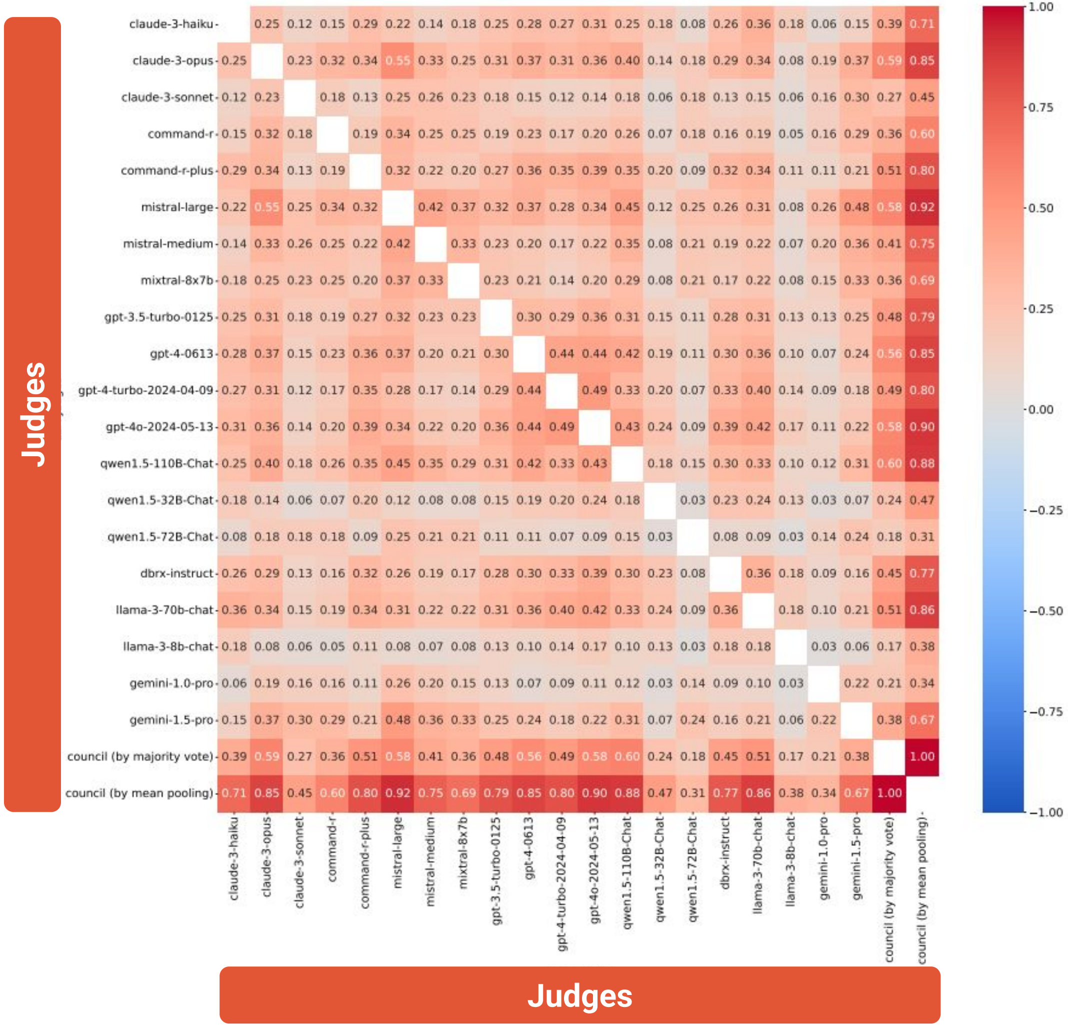
	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	pt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro
claude-3-haiku	0.50	0.39	0.47	0.54	0.53	0.54	0.60	0.54	0.74	0.63	0.32	0.30	0.25	0.39	0.40	0.50	0.44	0.59	0.38	0.81
claude-3-opus	0.61	0.50	0.58	0.65	0.64	0.65	0.70	0.65	0.81	0.72	0.42	0.41	0.34	0.50	0.51	0.61	0.55	0.70	0.49	0.87
claude-3-sonnet	0.53	0.42	0.50	0.58	0.56	0.58	0.63	0.58	0.76	0.66	0.35	0.33	0.28	0.42	0.43	0.53	0.47	0.63	0.42	0.83
command-r	0.46	0.35	0.42	0.50	0.49	0.50	0.56	0.50	0.70	0.59	0.28	0.27	0.22	0.35	0.36	0.46	0.40	0.55	0.34	0.79
command-r-plus	0.47	0.36	0.44	0.51	0.50	0.51	0.57	0.51	0.71	0.60	0.29	0.28	0.23	0.36	0.37	0.47	0.41	0.56	0.35	0.79
mistral-large	0.46	0.35	0.42	0.50	0.49	0.50	0.56	0.50	0.70	0.58	0.28	0.27	0.22	0.35	0.36	0.45	0.40	0.55	0.34	0.79
mistral-medium	0.40	0.30	0.37	0.44	0.43	0.44	0.50	0.44	0.65	0.53	0.24	0.22	0.18	0.30	0.30	0.40	0.34	0.49	0.29	0.74
mixtral-8x7b	0.46	0.35	0.42	0.50	0.49	0.50	0.56	0.50	0.70	0.58	0.28	0.27	0.22	0.35	0.36	0.45	0.40	0.55	0.34	0.78
gpt-3.5-turbo-0125	0.26	0.19	0.24	0.30	0.29	0.30	0.35	0.30	0.50	0.37	0.14	0.13	0.11	0.18	0.19	0.26	0.22	0.34	0.18	0.61
gpt-4-0613	0.37	0.28	0.34	0.41	0.40	0.42	0.47	0.42	0.63	0.50	0.22	0.21	0.17	0.28	0.28	0.37	0.32	0.47	0.27	0.72
gpt-4-turbo-2024-04-09	0.68	0.58	0.65	0.72	0.71	0.72	0.76	0.72	0.86	0.78	0.50	0.48	0.42	0.57	0.58	0.68	0.62	0.76	0.57	0.90
gpt-4o-2024-05-13	0.70	0.59	0.67	0.73	0.72	0.73	0.78	0.73	0.87	0.79	0.52	0.50	0.43	0.59	0.60	0.70	0.64	0.77	0.59	0.91
qwen1.5-110B-Chat	0.75	0.66	0.72	0.78	0.77	0.78	0.82	0.78	0.89	0.83	0.58	0.57	0.50	0.65	0.66	0.75	0.70	0.81	0.65	0.93
qwen1.5-32B-Chat	0.61	0.50	0.58	0.65	0.64	0.65	0.70	0.65	0.82	0.72	0.43	0.41	0.35	0.50	0.51	0.61	0.55	0.70	0.49	0.87
qwen1.5-72B-Chat	0.60	0.49	0.57	0.64	0.63	0.64	0.70	0.64	0.81	0.72	0.42	0.40	0.34	0.49	0.50	0.60	0.54	0.69	0.48	0.87
dbrx-instruct	0.50	0.39	0.47	0.54	0.53	0.55	0.60	0.55	0.74	0.63	0.32	0.30	0.25	0.39	0.40	0.50	0.44	0.60	0.38	0.81
llama-3-70b-chat	0.56	0.45	0.53	0.60	0.59	0.60	0.66	0.60	0.78	0.68	0.38	0.36	0.30	0.45	0.46	0.56	0.50	0.65	0.44	0.85
llama-3-8b-chat	0.41	0.30	0.37	0.45	0.44	0.45	0.51	0.45	0.66	0.53	0.24	0.23	0.19	0.30	0.31	0.40	0.35	0.50	0.30	0.75
gemini-1.0-pro	0.62	0.51	0.58	0.66	0.65	0.66	0.71	0.66	0.82	0.73	0.43	0.41	0.35	0.51	0.52	0.62	0.56	0.70	0.50	0.88
gemini-1.5-pro	0.19	0.13	0.17	0.21	0.21	0.21	0.26	0.22	0.39	0.28	0.10	0.09	0.07	0.13	0.13	0.19	0.15	0.25	0.12	0.50

💡 Insights

strong red/blue bands indicates consensus winners/losers

claude opus wins against llama 8b 70% of the time

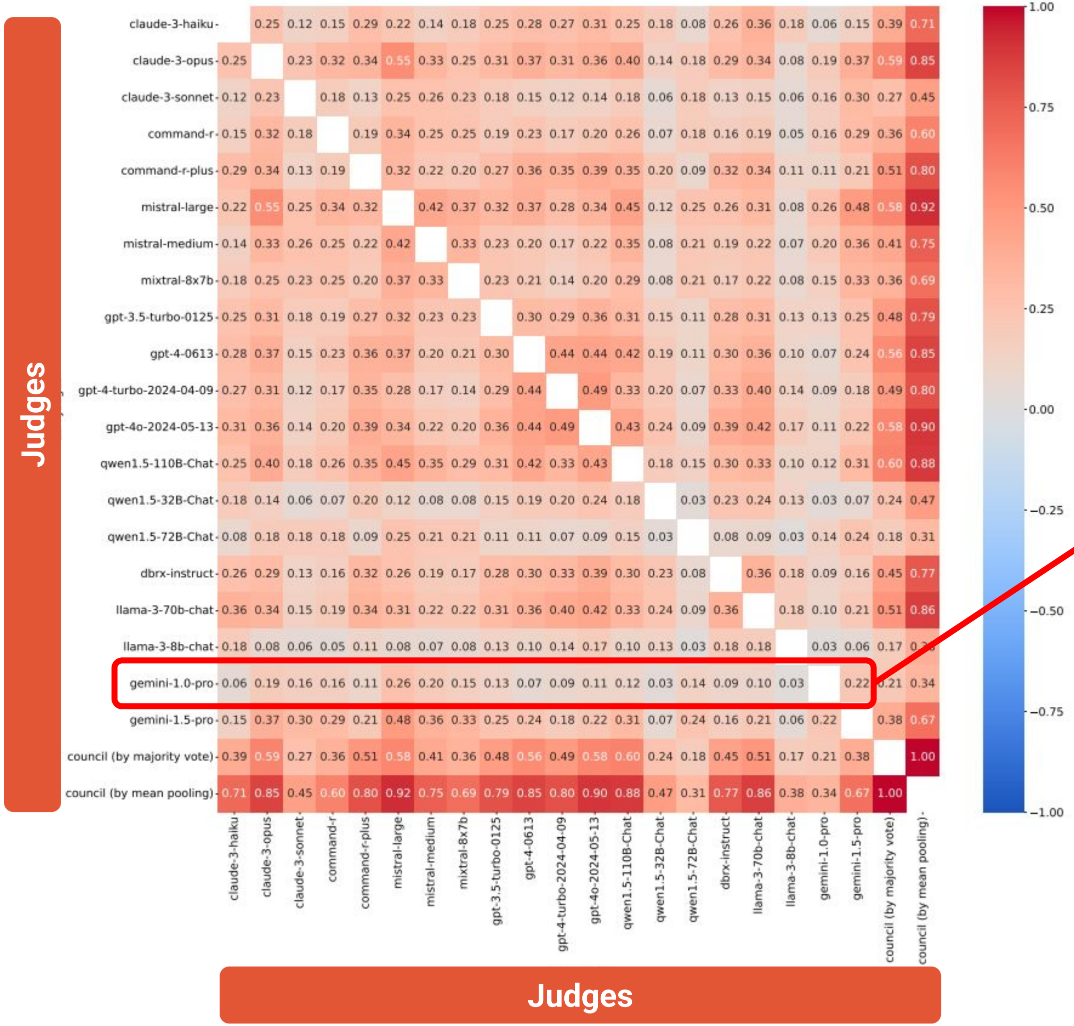
Judge vs. Judge (agreement)



	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	council (by majority vote)	council (by mean pooling)
claude-3-haiku		0.25	0.12	0.15	0.29	0.22	0.14	0.18	0.25	0.28	0.27	0.31	0.25	0.18	0.08	0.26	0.36	0.18	0.06	0.15	0.39	0.71
claude-3-opus	0.25		0.23	0.32	0.34	0.55	0.33	0.25	0.31	0.37	0.31	0.36	0.40	0.14	0.18	0.29	0.34	0.08	0.19	0.37	0.59	0.85
claude-3-sonnet	0.12	0.23		0.18	0.13	0.25	0.26	0.23	0.18	0.15	0.12	0.14	0.18	0.06	0.18	0.13	0.15	0.06	0.16	0.30	0.27	0.45
command-r	0.15	0.32	0.18		0.19	0.34	0.25	0.25	0.19	0.23	0.17	0.20	0.26	0.07	0.18	0.16	0.19	0.05	0.16	0.29	0.36	0.60
command-r-plus	0.29	0.34	0.13	0.19		0.32	0.22	0.20	0.27	0.36	0.35	0.39	0.35	0.20	0.09	0.32	0.34	0.11	0.11	0.21	0.51	0.80
mistral-large	0.22	0.55	0.25	0.34	0.32		0.42	0.37	0.32	0.37	0.28	0.34	0.45	0.12	0.25	0.26	0.31	0.08	0.26	0.48	0.58	0.92
mistral-medium	0.14	0.33	0.26	0.25	0.22	0.42		0.33	0.23	0.20	0.17	0.22	0.35	0.08	0.21	0.19	0.22	0.07	0.20	0.36	0.41	0.75
mixtral-8x7b	0.18	0.25	0.23	0.25	0.20	0.37	0.33		0.23	0.21	0.14	0.20	0.29	0.08	0.21	0.17	0.22	0.08	0.15	0.33	0.36	0.69
gpt-3.5-turbo-0125	0.25	0.31	0.18	0.19	0.27	0.32	0.23	0.23		0.30	0.29	0.36	0.31	0.15	0.11	0.28	0.31	0.13	0.13	0.25	0.48	0.79
gpt-4-0613	0.28	0.37	0.15	0.23	0.36	0.37	0.20	0.21	0.30		0.44	0.44	0.42	0.19	0.11	0.30	0.36	0.10	0.07	0.24	0.56	0.85
gpt-4-turbo-2024-04-09	0.27	0.31	0.12	0.17	0.35	0.28	0.17	0.14	0.29	0.44		0.49	0.33	0.20	0.07	0.33	0.40	0.14	0.09	0.18	0.49	0.80
gpt-4o-2024-05-13	0.31	0.36	0.14	0.20	0.39	0.34	0.22	0.20	0.36	0.44	0.49		0.43	0.24	0.09	0.39	0.42	0.17	0.11	0.22	0.58	0.90
qwen1.5-110B-Chat	0.25	0.40	0.18	0.26	0.35	0.45	0.35	0.29	0.31	0.42	0.33	0.43		0.18	0.15	0.30	0.33	0.10	0.12	0.31	0.60	0.88
qwen1.5-32B-Chat	0.18	0.14	0.06	0.07	0.20	0.12	0.08	0.08	0.15	0.19	0.20	0.24	0.18		0.03	0.23	0.24	0.13	0.03	0.07	0.24	0.47
qwen1.5-72B-Chat	0.08	0.18	0.18	0.18	0.09	0.25	0.21	0.21	0.11	0.11	0.07	0.09	0.15	0.03		0.08	0.09	0.03	0.14	0.24	0.18	0.31
dbrx-instruct	0.26	0.29	0.13	0.16	0.32	0.26	0.19	0.17	0.28	0.30	0.33	0.39	0.30	0.23	0.08		0.36	0.18	0.09	0.16	0.45	0.77
llama-3-70b-chat	0.36	0.34	0.15	0.19	0.34	0.31	0.22	0.22	0.31	0.36	0.40	0.42	0.33	0.24	0.09	0.36		0.18	0.10	0.21	0.51	0.86
llama-3-8b-chat	0.18	0.08	0.06	0.05	0.11	0.08	0.07	0.08	0.13	0.10	0.14	0.17	0.10	0.13	0.03	0.18	0.18		0.03	0.06	0.17	0.38
gemini-1.0-pro	0.06	0.19	0.16	0.16	0.11	0.26	0.20	0.15	0.13	0.07	0.09	0.11	0.12	0.03	0.14	0.09	0.10	0.03		0.22	0.21	0.34
gemini-1.5-pro	0.15	0.37	0.30	0.29	0.21	0.48	0.36	0.33	0.25	0.24	0.18	0.22	0.31	0.07	0.24	0.16	0.21	0.06	0.22		0.38	0.67
council (by majority vote)	0.39	0.59	0.27	0.36	0.51	0.58	0.41	0.36	0.48	0.56	0.49	0.58	0.60	0.24	0.18	0.45	0.51	0.17	0.21	0.38		1.00
council (by mean pooling)	0.71	0.85	0.45	0.60	0.80	0.92	0.75	0.69	0.79	0.85	0.80	0.90	0.88	0.47	0.31	0.77	0.86	0.38	0.34	0.67	1.00	

💡 Insights

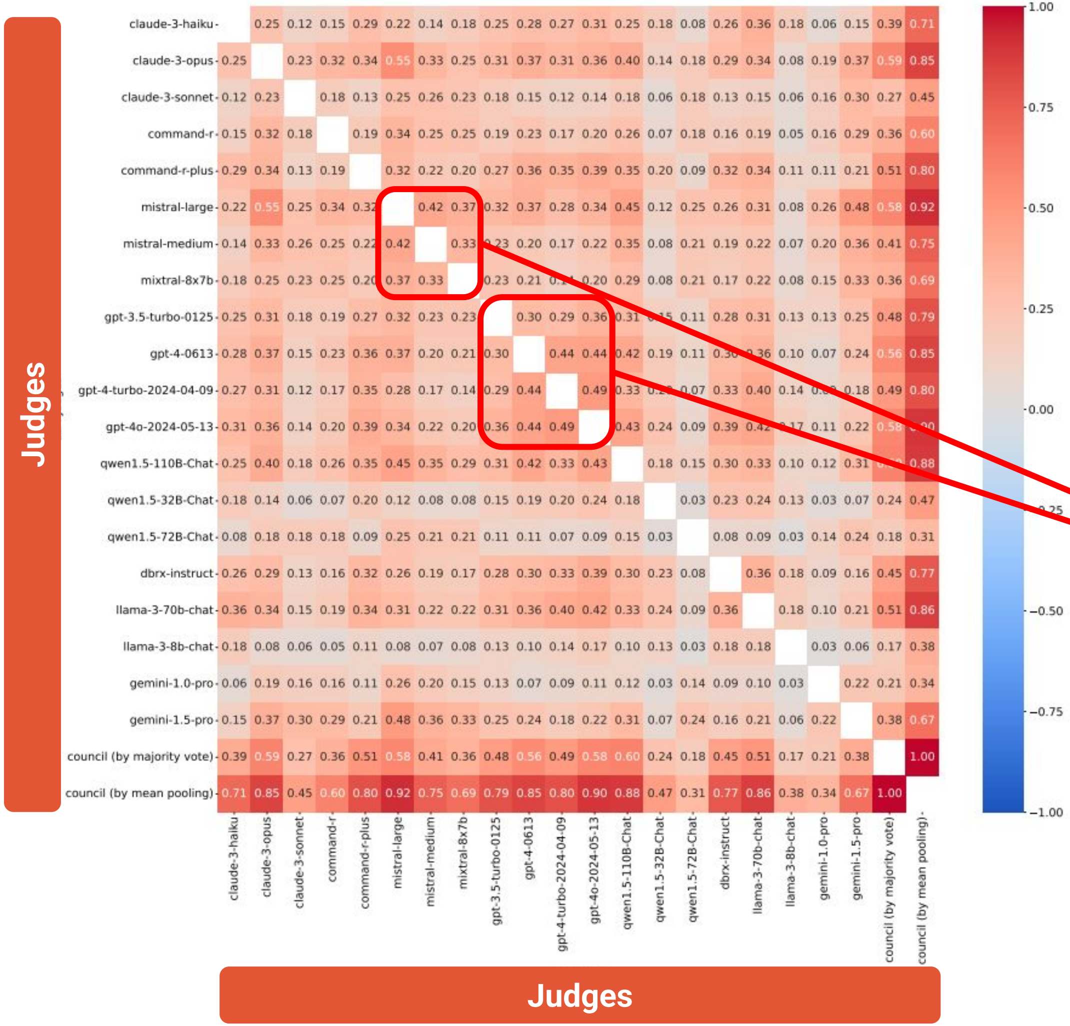
Judge vs. Judge (agreement)



	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	council (by majority vote)	council (by mean pooling)
claude-3-haiku		0.25	0.12	0.15	0.29	0.22	0.14	0.18	0.25	0.28	0.27	0.31	0.25	0.18	0.08	0.26	0.36	0.18	0.06	0.15	0.39	0.71
claude-3-opus	0.25		0.23	0.32	0.34	0.55	0.33	0.25	0.31	0.37	0.31	0.36	0.40	0.14	0.18	0.29	0.34	0.08	0.19	0.37	0.59	0.85
claude-3-sonnet	0.12	0.23		0.18	0.13	0.25	0.26	0.23	0.18	0.15	0.12	0.14	0.18	0.06	0.18	0.13	0.15	0.06	0.16	0.30	0.27	0.45
command-r	0.15	0.32	0.18		0.19	0.34	0.25	0.25	0.19	0.23	0.17	0.20	0.26	0.07	0.18	0.16	0.19	0.05	0.16	0.29	0.36	0.60
command-r-plus	0.29	0.34	0.13	0.19		0.32	0.22	0.20	0.27	0.36	0.35	0.39	0.35	0.20	0.09	0.32	0.34	0.11	0.11	0.21	0.51	0.80
mistral-large	0.22	0.55	0.25	0.34	0.32		0.42	0.37	0.32	0.37	0.28	0.34	0.45	0.12	0.25	0.26	0.31	0.08	0.26	0.48	0.58	0.92
mistral-medium	0.14	0.33	0.26	0.25	0.22	0.42		0.33	0.23	0.20	0.17	0.22	0.35	0.08	0.21	0.19	0.22	0.07	0.20	0.36	0.41	0.75
mixtral-8x7b	0.18	0.25	0.23	0.25	0.20	0.37	0.33		0.23	0.21	0.14	0.20	0.29	0.08	0.21	0.17	0.22	0.08	0.15	0.33	0.36	0.69
gpt-3.5-turbo-0125	0.25	0.31	0.18	0.19	0.27	0.32	0.23	0.23		0.30	0.29	0.36	0.31	0.15	0.11	0.28	0.31	0.13	0.13	0.25	0.48	0.79
gpt-4-0613	0.28	0.37	0.15	0.23	0.36	0.37	0.20	0.21	0.30		0.44	0.44	0.42	0.19	0.11	0.30	0.36	0.10	0.07	0.24	0.56	0.85
gpt-4-turbo-2024-04-09	0.27	0.31	0.12	0.17	0.35	0.28	0.17	0.14	0.29	0.44		0.49	0.33	0.20	0.07	0.33	0.40	0.14	0.09	0.18	0.49	0.80
gpt-4o-2024-05-13	0.31	0.36	0.14	0.20	0.39	0.34	0.22	0.20	0.36	0.44	0.49		0.43	0.24	0.09	0.39	0.42	0.17	0.11	0.22	0.58	0.90
qwen1.5-110B-Chat	0.25	0.40	0.18	0.26	0.35	0.45	0.35	0.29	0.31	0.42	0.33	0.43		0.18	0.15	0.30	0.33	0.10	0.12	0.31	0.60	0.88
qwen1.5-32B-Chat	0.18	0.14	0.06	0.07	0.20	0.12	0.08	0.08	0.15	0.19	0.20	0.24	0.18		0.03	0.23	0.24	0.13	0.03	0.07	0.24	0.47
qwen1.5-72B-Chat	0.08	0.18	0.18	0.18	0.09	0.25	0.21	0.21	0.11	0.11	0.07	0.09	0.15	0.03		0.08	0.09	0.03	0.14	0.24	0.18	0.31
dbrx-instruct	0.26	0.29	0.13	0.16	0.32	0.26	0.19	0.17	0.28	0.30	0.33	0.39	0.30	0.23	0.08		0.36	0.18	0.09	0.16	0.45	0.77
llama-3-70b-chat	0.36	0.34	0.15	0.19	0.34	0.31	0.22	0.22	0.31	0.36	0.40	0.42	0.33	0.24	0.09	0.36		0.18	0.10	0.21	0.51	0.86
llama-3-8b-chat	0.18	0.08	0.06	0.05	0.11	0.08	0.07	0.08	0.13	0.10	0.14	0.17	0.10	0.13	0.03	0.18	0.18		0.03	0.06	0.17	[illegible]
gemini-1.0-pro	0.06	0.19	0.16	0.16	0.11	0.26	0.20	0.15	0.13	0.07	0.09	0.11	0.12	0.03	0.14	0.09	0.10	0.03		0.22	0.21	0.34
gemini-1.5-pro	0.15	0.37	0.30	0.29	0.21	0.48	0.36	0.33	0.25	0.24	0.18	0.22	0.31	0.07	0.24	0.16	0.21	0.06	0.22		0.38	0.67
council (by majority vote)	0.39	0.59	0.27	0.36	0.51	0.58	0.41	0.36	0.48	0.56	0.49	0.58	0.60	0.24	0.18	0.45	0.51	0.17	0.21	0.38		1.00
council (by mean pooling)	0.71	0.85	0.45	0.60	0.80	0.92	0.75	0.69	0.79	0.85	0.80	0.90	0.88	0.47	0.31	0.77	0.86	0.38	0.34	0.67	1.00	

gemini-1.0 agrees with mistral-large the most and qwen-32B the least

Judge vs. Judge (agreement)



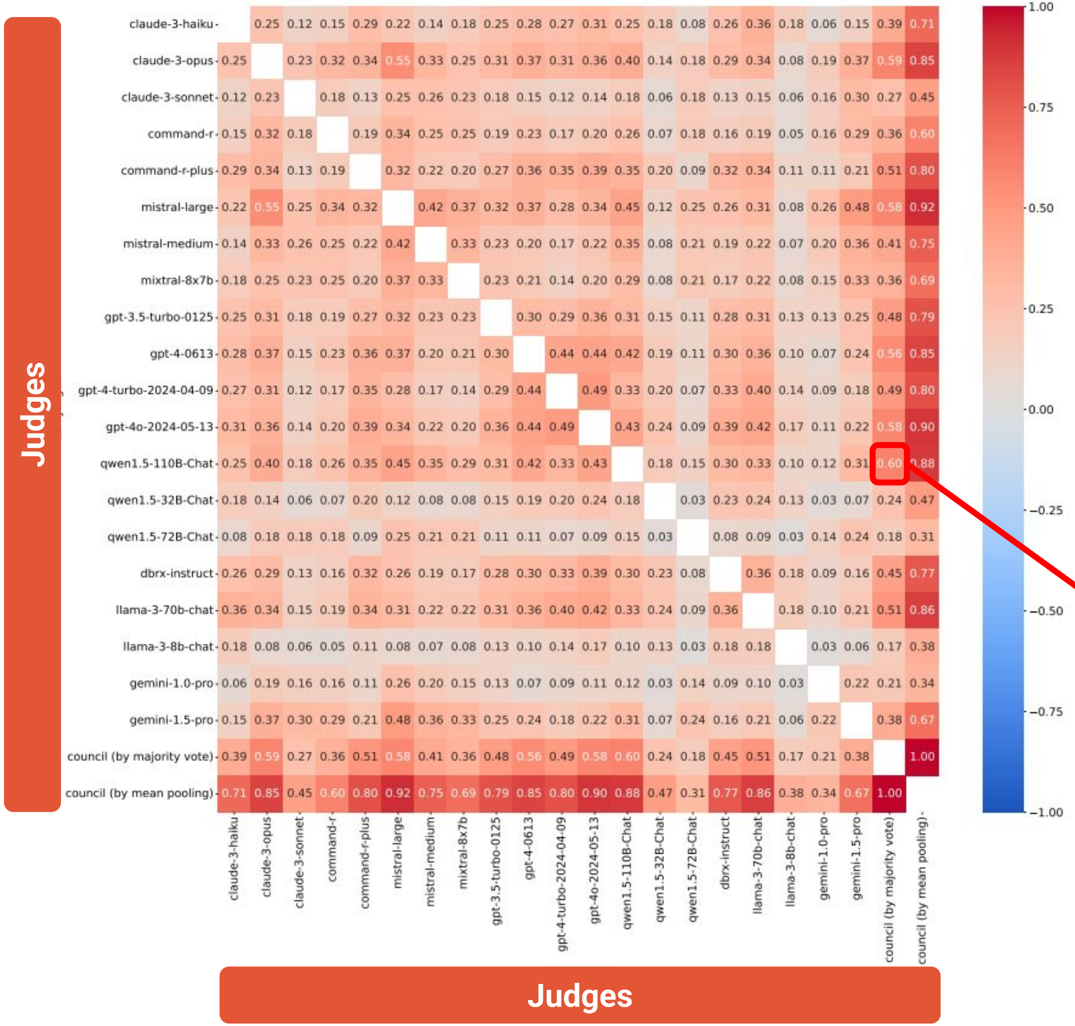
	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	council (by majority vote)	council (by mean pooling)
claude-3-haiku		0.25	0.12	0.15	0.29	0.22	0.14	0.18	0.25	0.28	0.27	0.31	0.25	0.18	0.08	0.26	0.36	0.18	0.06	0.15	0.39	0.71
claude-3-opus	0.25		0.23	0.32	0.34	0.55	0.33	0.25	0.31	0.37	0.31	0.36	0.40	0.14	0.18	0.29	0.34	0.08	0.19	0.37	0.59	0.85
claude-3-sonnet	0.12	0.23		0.18	0.13	0.25	0.26	0.23	0.18	0.15	0.12	0.14	0.18	0.06	0.18	0.13	0.15	0.06	0.16	0.30	0.27	0.45
command-r	0.15	0.32	0.18		0.19	0.34	0.25	0.25	0.19	0.23	0.17	0.20	0.26	0.07	0.18	0.16	0.19	0.05	0.16	0.29	0.36	0.60
command-r-plus	0.29	0.34	0.13	0.19		0.32	0.22	0.20	0.27	0.36	0.35	0.39	0.35	0.20	0.09	0.32	0.34	0.11	0.11	0.21	0.51	0.80
mistral-large	0.22	0.55	0.25	0.34	0.32		0.42	0.37	0.32	0.37	0.28	0.34	0.45	0.12	0.25	0.26	0.31	0.08	0.26	0.48	0.58	0.92
mistral-medium	0.14	0.33	0.26	0.25	0.22	0.42		0.33	0.23	0.20	0.17	0.22	0.35	0.08	0.21	0.19	0.22	0.07	0.20	0.36	0.41	0.75
mixtral-8x7b	0.18	0.25	0.23	0.25	0.20	0.37	0.33		0.23	0.21	0.14	0.20	0.29	0.08	0.21	0.17	0.22	0.08	0.15	0.33	0.36	0.69
gpt-3.5-turbo-0125	0.25	0.31	0.18	0.19	0.27	0.32	0.23	0.23		0.30	0.29	0.36	0.31	0.15	0.11	0.28	0.31	0.13	0.13	0.25	0.48	0.79
gpt-4-0613	0.28	0.37	0.15	0.23	0.36	0.37	0.20	0.21	0.30		0.44	0.44	0.42	0.19	0.11	0.30	0.36	0.10	0.07	0.24	0.56	0.85
gpt-4-turbo-2024-04-09	0.27	0.31	0.12	0.17	0.35	0.28	0.17	0.14	0.29	0.44		0.49	0.33	0.20	0.07	0.33	0.40	0.14	0.09	0.18	0.49	0.80
gpt-4o-2024-05-13	0.31	0.36	0.14	0.20	0.39	0.34	0.22	0.20	0.36	0.44	0.49		0.43	0.24	0.09	0.39	0.42	0.17	0.11	0.22	0.58	0.90
qwen1.5-110B-Chat	0.25	0.40	0.18	0.26	0.35	0.45	0.35	0.29	0.31	0.42	0.33	0.43		0.18	0.15	0.30	0.33	0.10	0.12	0.31	0.60	0.88
qwen1.5-32B-Chat	0.18	0.14	0.06	0.07	0.20	0.12	0.08	0.08	0.15	0.19	0.20	0.24	0.18		0.03	0.23	0.24	0.13	0.03	0.07	0.24	0.47
qwen1.5-72B-Chat	0.08	0.18	0.18	0.18	0.09	0.25	0.21	0.21	0.11	0.11	0.07	0.09	0.15	0.03		0.08	0.09	0.03	0.14	0.24	0.18	0.31
dbrx-instruct	0.26	0.29	0.13	0.16	0.32	0.26	0.19	0.17	0.28	0.30	0.33	0.39	0.30	0.23	0.08		0.36	0.18	0.09	0.16	0.45	0.77
llama-3-70b-chat	0.36	0.34	0.15	0.19	0.34	0.31	0.22	0.22	0.31	0.36	0.40	0.42	0.33	0.24	0.09	0.36		0.18	0.10	0.21	0.51	0.86
llama-3-8b-chat	0.18	0.08	0.06	0.05	0.11	0.08	0.07	0.08	0.13	0.10	0.14	0.17	0.10	0.13	0.03	0.18	0.18		0.03	0.06	0.17	0.38
gemini-1.0-pro	0.06	0.19	0.16	0.16	0.11	0.26	0.20	0.15	0.13	0.07	0.09	0.11	0.12	0.03	0.14	0.09	0.10	0.03		0.22	0.21	0.34
gemini-1.5-pro	0.15	0.37	0.30	0.29	0.21	0.48	0.36	0.33	0.25	0.24	0.18	0.22	0.31	0.07	0.24	0.16	0.21	0.06	0.22		0.38	0.67
council (by majority vote)	0.39	0.59	0.27	0.36	0.51	0.58	0.41	0.36	0.48	0.56	0.49	0.58	0.60	0.24	0.18	0.45	0.51	0.17	0.21	0.38		1.00
council (by mean pooling)	0.71	0.85	0.45	0.60	0.80	0.92	0.75	0.69	0.79	0.85	0.80	0.90	0.88	0.47	0.31	0.77	0.86	0.38	0.34	0.67	1.00	

💡 Insights

gemini-1.0 agrees with mistral-large the most and qwen-32B the least

openai and mistral have the strongest inter-family agreement

Judge vs. Judge (agreement)



	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	council (by majority vote)	council (by mean pooling)
claude-3-haiku		0.25	0.12	0.15	0.29	0.22	0.14	0.18	0.25	0.28	0.27	0.31	0.25	0.18	0.08	0.26	0.36	0.18	0.06	0.15	0.39	0.71
claude-3-opus	0.25		0.23	0.32	0.34	0.55	0.33	0.25	0.31	0.37	0.31	0.36	0.40	0.14	0.18	0.29	0.34	0.08	0.19	0.37	0.59	0.85
claude-3-sonnet	0.12	0.23		0.18	0.13	0.25	0.26	0.23	0.18	0.15	0.12	0.14	0.18	0.06	0.18	0.13	0.15	0.06	0.16	0.30	0.27	0.45
command-r	0.15	0.32	0.18		0.19	0.34	0.25	0.25	0.19	0.23	0.17	0.20	0.26	0.07	0.18	0.16	0.19	0.05	0.16	0.29	0.36	0.60
command-r-plus	0.29	0.34	0.13	0.19		0.32	0.22	0.20	0.27	0.36	0.35	0.39	0.35	0.20	0.09	0.32	0.34	0.11	0.11	0.21	0.51	0.80
mistral-large	0.22	0.55	0.25	0.34	0.32		0.42	0.37	0.32	0.37	0.28	0.34	0.45	0.12	0.25	0.26	0.31	0.08	0.26	0.48	0.58	0.92
mistral-medium	0.14	0.33	0.26	0.25	0.22	0.42		0.33	0.23	0.20	0.17	0.22	0.35	0.08	0.21	0.19	0.22	0.07	0.20	0.36	0.41	0.75
mixtral-8x7b	0.18	0.25	0.23	0.25	0.20	0.37	0.33		0.23	0.21	0.14	0.20	0.29	0.08	0.21	0.17	0.22	0.08	0.15	0.33	0.36	0.69
gpt-3.5-turbo-0125	0.25	0.31	0.18	0.19	0.27	0.32	0.23	0.23		0.30	0.29	0.36	0.31	0.15	0.11	0.28	0.31	0.13	0.13	0.25	0.48	0.79
gpt-4-0613	0.28	0.37	0.15	0.23	0.36	0.37	0.20	0.21	0.30		0.44	0.44	0.42	0.19	0.11	0.30	0.36	0.10	0.07	0.24	0.56	0.85
gpt-4-turbo-2024-04-09	0.27	0.31	0.12	0.17	0.35	0.28	0.17	0.14	0.29	0.44		0.49	0.33	0.20	0.07	0.33	0.40	0.14	0.09	0.18	0.49	0.80
gpt-4o-2024-05-13	0.31	0.36	0.14	0.20	0.39	0.34	0.22	0.20	0.36	0.44	0.49		0.43	0.24	0.09	0.39	0.42	0.17	0.11	0.22	0.58	0.90
qwen1.5-110B-Chat	0.25	0.40	0.18	0.26	0.35	0.45	0.35	0.29	0.31	0.42	0.33	0.43		0.18	0.15	0.30	0.33	0.10	0.12	0.31	0.60	0.88
qwen1.5-32B-Chat	0.18	0.14	0.06	0.07	0.20	0.12	0.08	0.08	0.15	0.19	0.20	0.24	0.18		0.03	0.23	0.24	0.13	0.03	0.07	0.24	0.47
qwen1.5-72B-Chat	0.08	0.18	0.18	0.18	0.09	0.25	0.21	0.21	0.11	0.11	0.07	0.09	0.15	0.03		0.08	0.09	0.03	0.14	0.24	0.18	0.31
dbrx-instruct	0.26	0.29	0.13	0.16	0.32	0.26	0.19	0.17	0.28	0.30	0.33	0.39	0.30	0.23	0.08		0.36	0.18	0.09	0.16	0.45	0.77
llama-3-70b-chat	0.36	0.34	0.15	0.19	0.34	0.31	0.22	0.22	0.31	0.36	0.40	0.42	0.33	0.24	0.09	0.36		0.18	0.10	0.21	0.51	0.86
llama-3-8b-chat	0.18	0.08	0.06	0.05	0.11	0.08	0.07	0.08	0.13	0.10	0.14	0.17	0.10	0.13	0.03	0.18	0.18		0.03	0.06	0.17	0.38
gemini-1.0-pro	0.06	0.19	0.16	0.16	0.11	0.26	0.20	0.15	0.13	0.07	0.09	0.11	0.12	0.03	0.14	0.09	0.10	0.03		0.22	0.21	0.34
gemini-1.5-pro	0.15	0.37	0.30	0.29	0.21	0.48	0.36	0.33	0.25	0.24	0.18	0.22	0.31	0.07	0.24	0.16	0.21	0.06	0.22		0.38	0.67
council (by majority vote)	0.39	0.59	0.27	0.36	0.51	0.58	0.41	0.36	0.48	0.56	0.49	0.58	0.60	0.24	0.18	0.45	0.51	0.17	0.21	0.38		1.00
council (by mean pooling)	0.71	0.85	0.45	0.60	0.80	0.92	0.75	0.69	0.79	0.85	0.80	0.90	0.88	0.47	0.31	0.77	0.86	0.38	0.34	0.67	1.00	

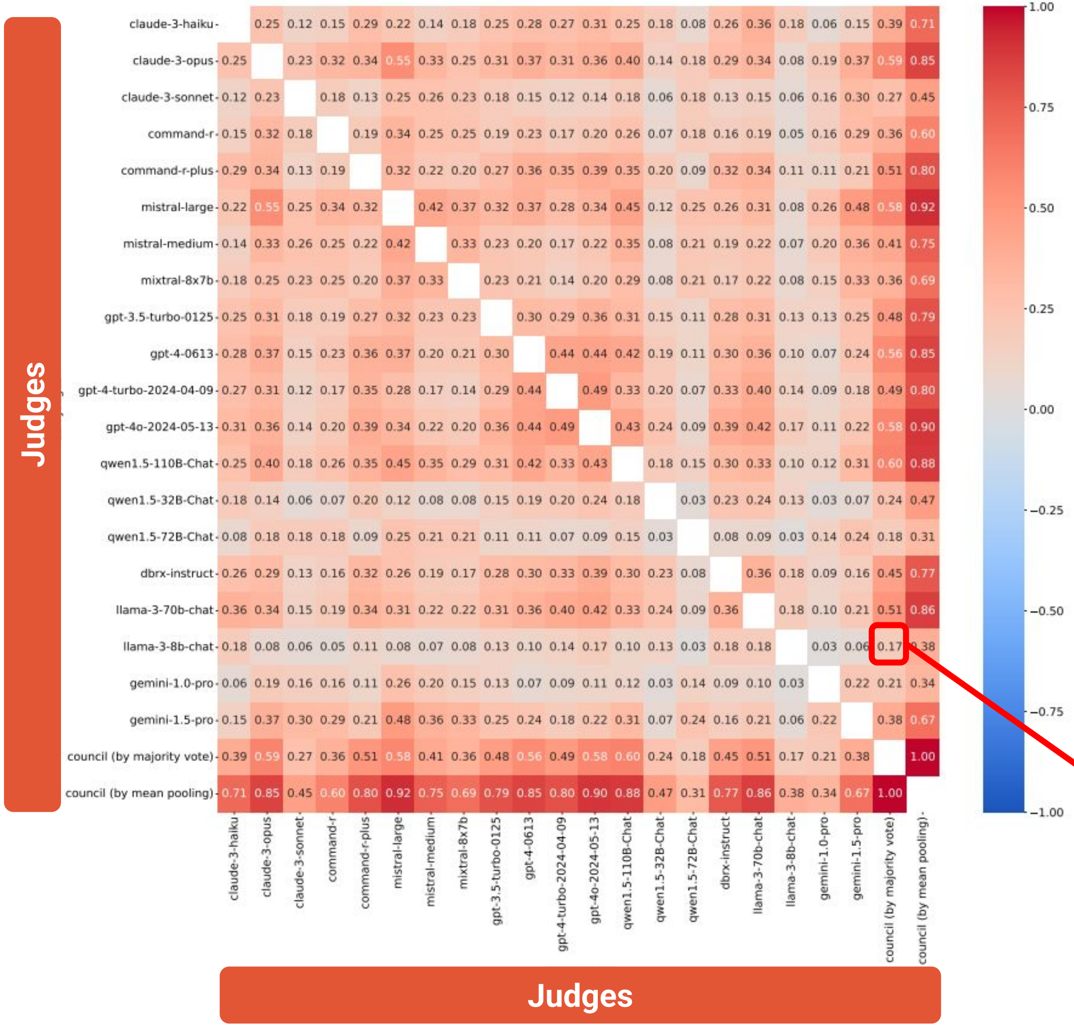
💡 Insights

gemini-1.0 agrees with mistral-large the most and qwen-32B the least

openai and mistral have the strongest inter-family agreement

qwen-110B is the most **representative** of the majority opinion

Judge vs. Judge (agreement)



	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	council (by majority vote)	council (by mean pooling)
claude-3-haiku		0.25	0.12	0.15	0.29	0.22	0.14	0.18	0.25	0.28	0.27	0.31	0.25	0.18	0.08	0.26	0.36	0.18	0.06	0.15	0.39	0.71
claude-3-opus	0.25		0.23	0.32	0.34	0.55	0.33	0.25	0.31	0.37	0.31	0.36	0.40	0.14	0.18	0.29	0.34	0.08	0.19	0.37	0.59	0.85
claude-3-sonnet	0.12	0.23		0.18	0.13	0.25	0.26	0.23	0.18	0.15	0.12	0.14	0.18	0.06	0.18	0.13	0.15	0.06	0.16	0.30	0.27	0.45
command-r	0.15	0.32	0.18		0.19	0.34	0.25	0.25	0.19	0.23	0.17	0.20	0.26	0.07	0.18	0.16	0.19	0.05	0.16	0.29	0.36	0.60
command-r-plus	0.29	0.34	0.13	0.19		0.32	0.22	0.20	0.27	0.36	0.35	0.39	0.35	0.20	0.09	0.32	0.34	0.11	0.11	0.21	0.51	0.80
mistral-large	0.22	0.55	0.25	0.34	0.32		0.42	0.37	0.32	0.37	0.28	0.34	0.45	0.12	0.25	0.26	0.31	0.08	0.26	0.48	0.58	0.92
mistral-medium	0.14	0.33	0.26	0.25	0.22	0.42		0.33	0.23	0.20	0.17	0.22	0.35	0.08	0.21	0.19	0.22	0.07	0.20	0.36	0.41	0.75
mixtral-8x7b	0.18	0.25	0.23	0.25	0.20	0.37	0.33		0.23	0.21	0.14	0.20	0.29	0.08	0.21	0.17	0.22	0.08	0.15	0.33	0.36	0.69
gpt-3.5-turbo-0125	0.25	0.31	0.18	0.19	0.27	0.32	0.23	0.23		0.30	0.29	0.36	0.31	0.15	0.11	0.28	0.31	0.13	0.13	0.25	0.48	0.79
gpt-4-0613	0.28	0.37	0.15	0.23	0.36	0.37	0.20	0.21	0.30		0.44	0.44	0.42	0.19	0.11	0.30	0.36	0.10	0.07	0.24	0.56	0.85
gpt-4-turbo-2024-04-09	0.27	0.31	0.12	0.17	0.35	0.28	0.17	0.14	0.29	0.44		0.49	0.33	0.20	0.07	0.33	0.40	0.14	0.09	0.18	0.49	0.80
gpt-4o-2024-05-13	0.31	0.36	0.14	0.20	0.39	0.34	0.22	0.20	0.36	0.44	0.49		0.43	0.24	0.09	0.39	0.42	0.17	0.11	0.22	0.58	0.90
qwen1.5-110B-Chat	0.25	0.40	0.18	0.26	0.35	0.45	0.35	0.29	0.31	0.42	0.33	0.43		0.18	0.15	0.30	0.33	0.10	0.12	0.31	0.60	0.88
qwen1.5-32B-Chat	0.18	0.14	0.06	0.07	0.20	0.12	0.08	0.08	0.15	0.19	0.20	0.24	0.18		0.03	0.23	0.24	0.13	0.03	0.07	0.24	0.47
qwen1.5-72B-Chat	0.08	0.18	0.18	0.18	0.09	0.25	0.21	0.21	0.11	0.11	0.07	0.09	0.15	0.03		0.08	0.09	0.03	0.14	0.24	0.18	0.31
dbrx-instruct	0.26	0.29	0.13	0.16	0.32	0.26	0.19	0.17	0.28	0.30	0.33	0.39	0.30	0.23	0.08		0.36	0.18	0.09	0.16	0.45	0.77
llama-3-70b-chat	0.36	0.34	0.15	0.19	0.34	0.31	0.22	0.22	0.31	0.36	0.40	0.42	0.33	0.24	0.09	0.36		0.18	0.10	0.21	0.51	0.86
llama-3-8b-chat	0.18	0.08	0.06	0.05	0.11	0.08	0.07	0.08	0.13	0.10	0.14	0.17	0.10	0.13	0.03	0.18	0.18		0.03	0.06	0.17	0.38
gemini-1.0-pro	0.06	0.19	0.16	0.16	0.11	0.26	0.20	0.15	0.13	0.07	0.09	0.11	0.12	0.03	0.14	0.09	0.10	0.03		0.22	0.21	0.34
gemini-1.5-pro	0.15	0.37	0.30	0.29	0.21	0.48	0.36	0.33	0.25	0.24	0.18	0.22	0.31	0.07	0.24	0.16	0.21	0.06	0.22		0.38	0.67
council (by majority vote)	0.39	0.59	0.27	0.36	0.51	0.58	0.41	0.36	0.48	0.56	0.49	0.58	0.60	0.24	0.18	0.45	0.51	0.17	0.21	0.38		1.00
council (by mean pooling)	0.71	0.85	0.45	0.60	0.80	0.92	0.75	0.69	0.79	0.85	0.80	0.90	0.88	0.47	0.31	0.77	0.86	0.38	0.34	0.67	1.00	

💡 Insights

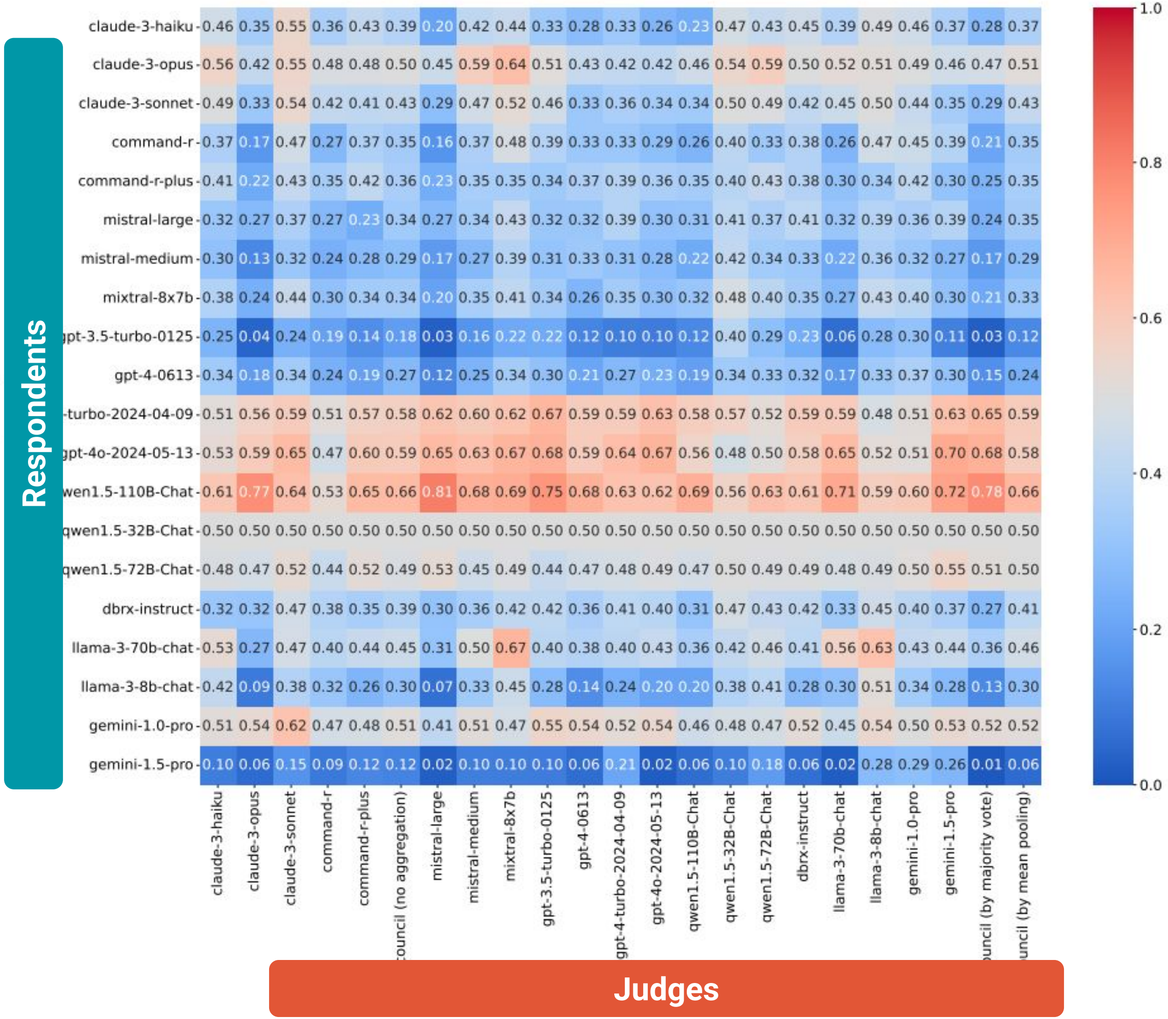
gemini-1.0 agrees with mistral-large the most and qwen-32B the least

openai and mistral have the strongest inter-family agreement

qwen-110B is the most **representative** of the majority opinion

llama-8B is the most **contrarian** judge

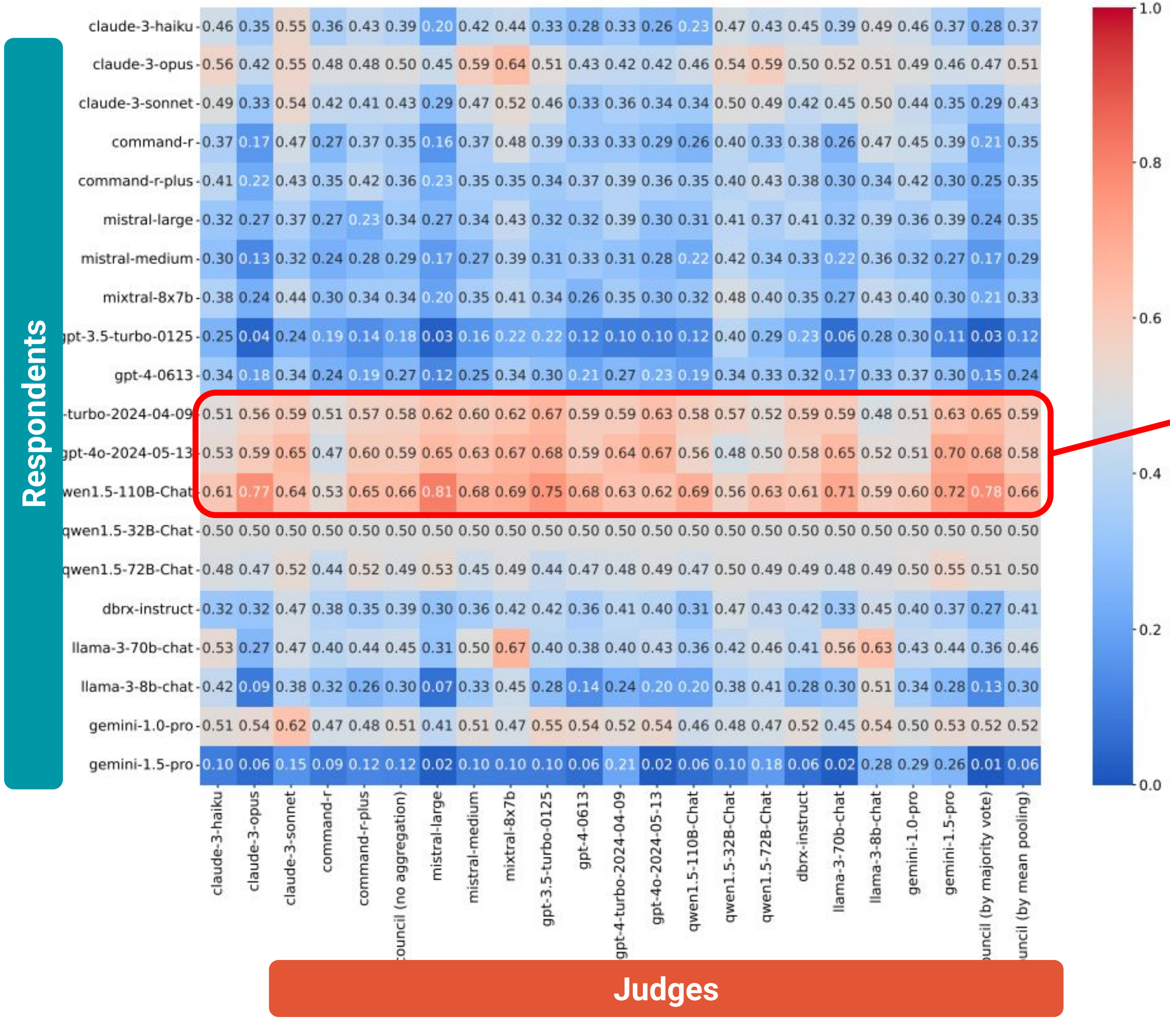
Judge vs. Respondent (affinities)



	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	council (no aggregation)	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	ouncil (by majority vote)	uncil (by mean pooling)
claude-3-haiku	0.46	0.35	0.55	0.36	0.43	0.39	0.20	0.42	0.44	0.33	0.28	0.33	0.26	0.23	0.47	0.43	0.45	0.39	0.49	0.46	0.37	0.28	0.37
claude-3-opus	0.56	0.42	0.55	0.48	0.48	0.50	0.45	0.59	0.64	0.51	0.43	0.42	0.42	0.46	0.54	0.59	0.50	0.52	0.51	0.49	0.46	0.47	0.51
claude-3-sonnet	0.49	0.33	0.54	0.42	0.41	0.43	0.29	0.47	0.52	0.46	0.33	0.36	0.34	0.34	0.50	0.49	0.42	0.45	0.50	0.44	0.35	0.29	0.43
command-r	0.37	0.17	0.47	0.27	0.37	0.35	0.16	0.37	0.48	0.39	0.33	0.33	0.29	0.26	0.40	0.33	0.38	0.26	0.47	0.45	0.39	0.21	0.35
command-r-plus	0.41	0.22	0.43	0.35	0.42	0.36	0.23	0.35	0.35	0.34	0.37	0.39	0.36	0.35	0.40	0.43	0.38	0.30	0.34	0.42	0.30	0.25	0.35
mistral-large	0.32	0.27	0.37	0.27	0.23	0.34	0.27	0.34	0.43	0.32	0.32	0.39	0.30	0.31	0.41	0.37	0.41	0.32	0.39	0.36	0.39	0.24	0.35
mistral-medium	0.30	0.13	0.32	0.24	0.28	0.29	0.17	0.27	0.39	0.31	0.33	0.31	0.28	0.22	0.42	0.34	0.33	0.22	0.36	0.32	0.27	0.17	0.29
mixtral-8x7b	0.38	0.24	0.44	0.30	0.34	0.34	0.20	0.35	0.41	0.34	0.26	0.35	0.30	0.32	0.48	0.40	0.35	0.27	0.43	0.40	0.30	0.21	0.33
gpt-3.5-turbo-0125	0.25	0.04	0.24	0.19	0.14	0.18	0.03	0.16	0.22	0.22	0.12	0.10	0.10	0.12	0.40	0.29	0.23	0.06	0.28	0.30	0.11	0.03	0.12
gpt-4-0613	0.34	0.18	0.34	0.24	0.19	0.27	0.12	0.25	0.34	0.30	0.21	0.27	0.23	0.19	0.34	0.33	0.32	0.17	0.33	0.37	0.30	0.15	0.24
-turbo-2024-04-09	0.51	0.56	0.59	0.51	0.57	0.58	0.62	0.60	0.62	0.67	0.59	0.59	0.63	0.58	0.57	0.52	0.59	0.59	0.48	0.51	0.63	0.65	0.59
gpt-4o-2024-05-13	0.53	0.59	0.65	0.47	0.60	0.59	0.65	0.63	0.67	0.68	0.59	0.64	0.67	0.56	0.48	0.50	0.58	0.65	0.52	0.51	0.70	0.68	0.58
wen1.5-110B-Chat	0.61	0.77	0.64	0.53	0.65	0.66	0.81	0.68	0.69	0.75	0.68	0.63	0.62	0.69	0.56	0.63	0.61	0.71	0.59	0.60	0.72	0.78	0.66
qwen1.5-32B-Chat	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
qwen1.5-72B-Chat	0.48	0.47	0.52	0.44	0.52	0.49	0.53	0.45	0.49	0.44	0.47	0.48	0.49	0.47	0.50	0.49	0.49	0.48	0.49	0.50	0.55	0.51	0.50
dbrx-instruct	0.32	0.32	0.47	0.38	0.35	0.39	0.30	0.36	0.42	0.42	0.36	0.41	0.40	0.31	0.47	0.43	0.42	0.33	0.45	0.40	0.37	0.27	0.41
llama-3-70b-chat	0.53	0.27	0.47	0.40	0.44	0.45	0.31	0.50	0.67	0.40	0.38	0.40	0.43	0.36	0.42	0.46	0.41	0.56	0.63	0.43	0.44	0.36	0.46
llama-3-8b-chat	0.42	0.09	0.38	0.32	0.26	0.30	0.07	0.33	0.45	0.28	0.14	0.24	0.20	0.20	0.38	0.41	0.28	0.30	0.51	0.34	0.28	0.13	0.30
gemini-1.0-pro	0.51	0.54	0.62	0.47	0.48	0.51	0.41	0.51	0.47	0.55	0.54	0.52	0.54	0.46	0.48	0.47	0.52	0.45	0.54	0.50	0.53	0.52	0.52
gemini-1.5-pro	0.10	0.06	0.15	0.09	0.12	0.12	0.02	0.10	0.10	0.10	0.06	0.21	0.02	0.06	0.10	0.18	0.06	0.02	0.28	0.29	0.26	0.01	0.06

💡 Insights

Judge vs. Respondent (affinities)

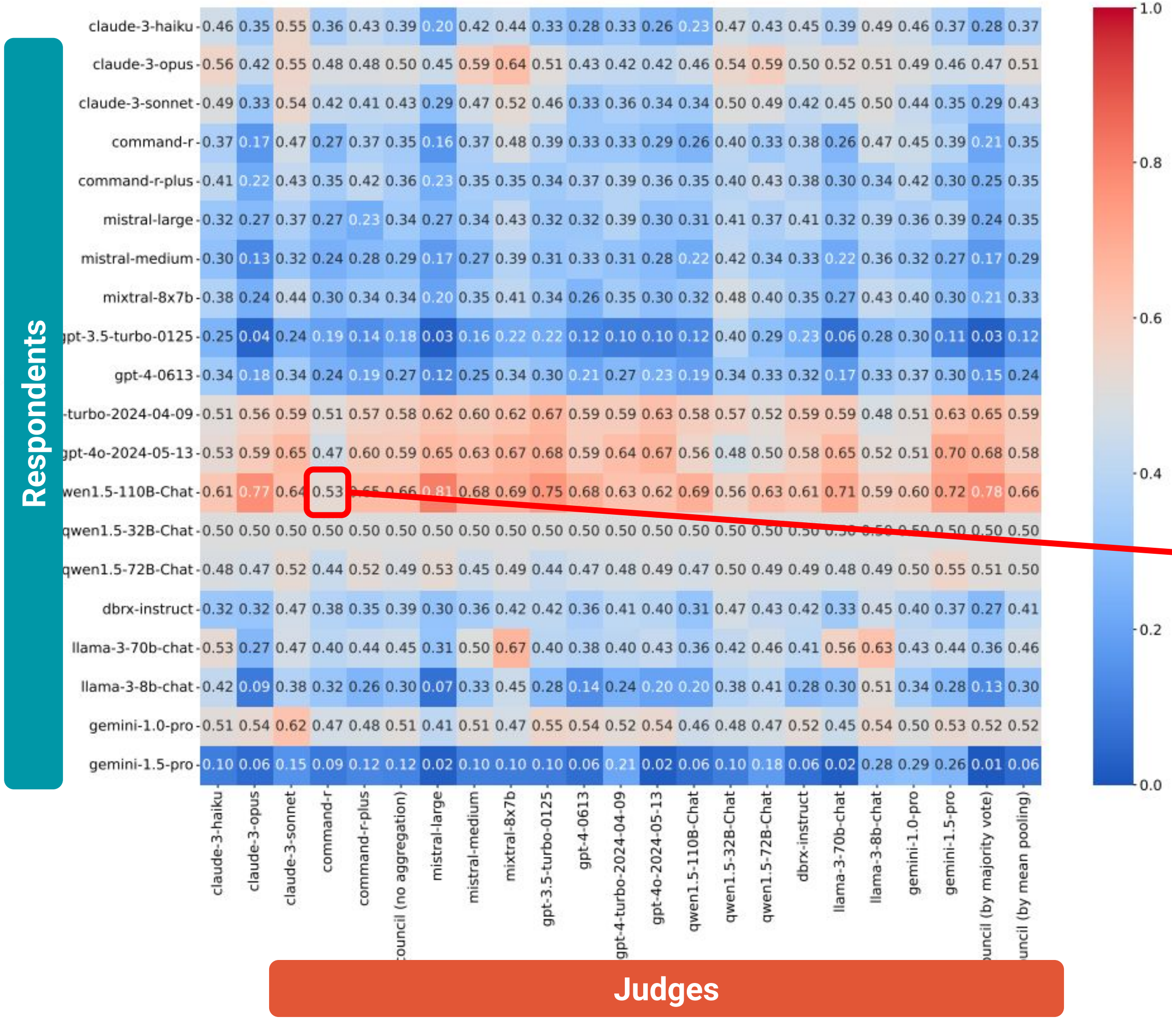


	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	council (no aggregation)	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	ouncil (by majority vote)	uncil (by mean pooling)
claude-3-haiku	0.46	0.35	0.55	0.36	0.43	0.39	0.20	0.42	0.44	0.33	0.28	0.33	0.26	0.23	0.47	0.43	0.45	0.39	0.49	0.46	0.37	0.28	0.37
claude-3-opus	0.56	0.42	0.55	0.48	0.48	0.50	0.45	0.59	0.64	0.51	0.43	0.42	0.42	0.46	0.54	0.59	0.50	0.52	0.51	0.49	0.46	0.47	0.51
claude-3-sonnet	0.49	0.33	0.54	0.42	0.41	0.43	0.29	0.47	0.52	0.46	0.33	0.36	0.34	0.34	0.50	0.49	0.42	0.45	0.50	0.44	0.35	0.29	0.43
command-r	0.37	0.17	0.47	0.27	0.37	0.35	0.16	0.37	0.48	0.39	0.33	0.33	0.29	0.26	0.40	0.33	0.38	0.26	0.47	0.45	0.39	0.21	0.35
command-r-plus	0.41	0.22	0.43	0.35	0.42	0.36	0.23	0.35	0.35	0.34	0.37	0.39	0.36	0.35	0.40	0.43	0.38	0.30	0.34	0.42	0.30	0.25	0.35
mistral-large	0.32	0.27	0.37	0.27	0.23	0.34	0.27	0.34	0.43	0.32	0.32	0.39	0.30	0.31	0.41	0.37	0.41	0.32	0.39	0.36	0.39	0.24	0.35
mistral-medium	0.30	0.13	0.32	0.24	0.28	0.29	0.17	0.27	0.39	0.31	0.33	0.31	0.28	0.22	0.42	0.34	0.33	0.22	0.36	0.32	0.27	0.17	0.29
mixtral-8x7b	0.38	0.24	0.44	0.30	0.34	0.34	0.20	0.35	0.41	0.34	0.26	0.35	0.30	0.32	0.48	0.40	0.35	0.27	0.43	0.40	0.30	0.21	0.33
gpt-3.5-turbo-0125	0.25	0.04	0.24	0.19	0.14	0.18	0.03	0.16	0.22	0.22	0.12	0.10	0.10	0.12	0.40	0.29	0.23	0.06	0.28	0.30	0.11	0.03	0.12
gpt-4-0613	0.34	0.18	0.34	0.24	0.19	0.27	0.12	0.25	0.34	0.30	0.21	0.27	0.23	0.19	0.34	0.33	0.32	0.17	0.33	0.37	0.30	0.15	0.24
-turbo-2024-04-09	0.51	0.56	0.59	0.51	0.57	0.58	0.62	0.60	0.62	0.67	0.59	0.59	0.63	0.58	0.57	0.52	0.59	0.59	0.48	0.51	0.63	0.65	0.59
gpt-4o-2024-05-13	0.53	0.59	0.65	0.47	0.60	0.59	0.65	0.63	0.67	0.68	0.59	0.64	0.67	0.56	0.48	0.50	0.58	0.65	0.52	0.51	0.70	0.68	0.58
wen1.5-110B-Chat	0.61	0.77	0.64	0.53	0.65	0.66	0.81	0.68	0.69	0.75	0.68	0.63	0.62	0.69	0.56	0.63	0.61	0.71	0.59	0.60	0.72	0.78	0.66
qwen1.5-32B-Chat	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
qwen1.5-72B-Chat	0.48	0.47	0.52	0.44	0.52	0.49	0.53	0.45	0.49	0.44	0.47	0.48	0.49	0.47	0.50	0.49	0.49	0.48	0.49	0.50	0.55	0.51	0.50
dbrx-instruct	0.32	0.32	0.47	0.38	0.35	0.39	0.30	0.36	0.42	0.42	0.36	0.41	0.40	0.31	0.47	0.43	0.42	0.33	0.45	0.40	0.37	0.27	0.41
llama-3-70b-chat	0.53	0.27	0.47	0.40	0.44	0.45	0.31	0.50	0.67	0.40	0.38	0.40	0.43	0.36	0.42	0.46	0.41	0.56	0.63	0.43	0.44	0.36	0.46
llama-3-8b-chat	0.42	0.09	0.38	0.32	0.26	0.30	0.07	0.33	0.45	0.28	0.14	0.24	0.20	0.20	0.38	0.41	0.28	0.30	0.51	0.34	0.28	0.13	0.30
gemini-1.0-pro	0.51	0.54	0.62	0.47	0.48	0.51	0.41	0.51	0.47	0.55	0.54	0.52	0.54	0.46	0.48	0.47	0.52	0.45	0.54	0.50	0.53	0.52	0.52
gemini-1.5-pro	0.10	0.06	0.15	0.09	0.12	0.12	0.02	0.10	0.10	0.10	0.06	0.21	0.02	0.06	0.10	0.18	0.06	0.02	0.28	0.29	0.26	0.01	0.06

💡 Insights

most council members genuinely liked Qwen-110B across the board.

Judge vs. Respondent (affinities)



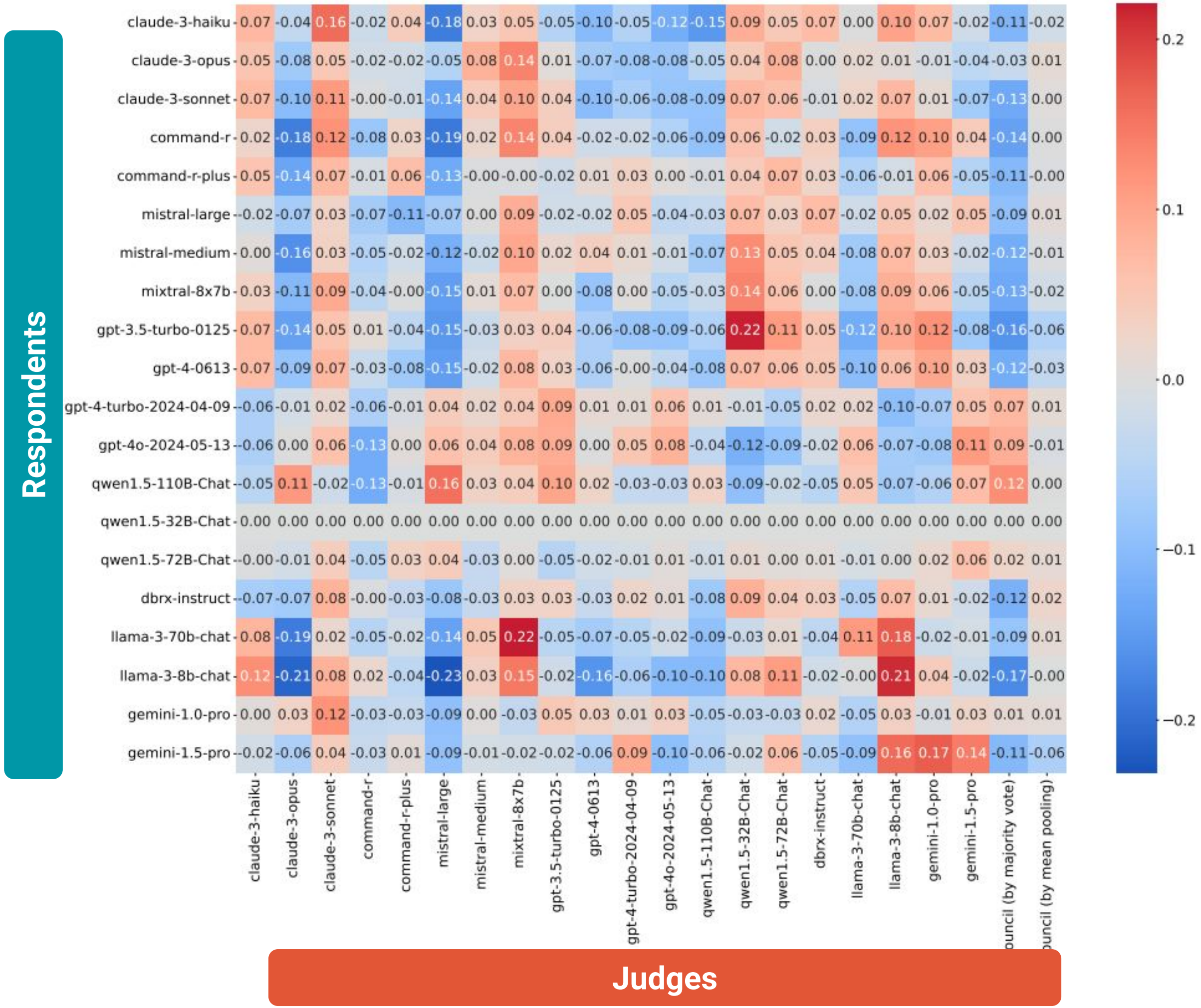
	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	ouncil (no aggregation)	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	pt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	uncil (by majority vote)	uncil (by mean pooling)
claude-3-haiku	0.46	0.35	0.55	0.36	0.43	0.39	0.20	0.42	0.44	0.33	0.28	0.33	0.26	0.23	0.47	0.43	0.45	0.39	0.49	0.46	0.37	0.28	0.37
claude-3-opus	0.56	0.42	0.55	0.48	0.48	0.50	0.45	0.59	0.64	0.51	0.43	0.42	0.42	0.46	0.54	0.59	0.50	0.52	0.51	0.49	0.46	0.47	0.51
claude-3-sonnet	0.49	0.33	0.54	0.42	0.41	0.43	0.29	0.47	0.52	0.46	0.33	0.36	0.34	0.34	0.50	0.49	0.42	0.45	0.50	0.44	0.35	0.29	0.43
command-r	0.37	0.17	0.47	0.27	0.37	0.35	0.16	0.37	0.48	0.39	0.33	0.33	0.29	0.26	0.40	0.33	0.38	0.26	0.47	0.45	0.39	0.21	0.35
command-r-plus	0.41	0.22	0.43	0.35	0.42	0.36	0.23	0.35	0.35	0.34	0.37	0.39	0.36	0.35	0.40	0.43	0.38	0.30	0.34	0.42	0.30	0.25	0.35
mistral-large	0.32	0.27	0.37	0.27	0.23	0.34	0.27	0.34	0.43	0.32	0.32	0.39	0.30	0.31	0.41	0.37	0.41	0.32	0.39	0.36	0.39	0.24	0.35
mistral-medium	0.30	0.13	0.32	0.24	0.28	0.29	0.17	0.27	0.39	0.31	0.33	0.31	0.28	0.22	0.42	0.34	0.33	0.22	0.36	0.32	0.27	0.17	0.29
mixtral-8x7b	0.38	0.24	0.44	0.30	0.34	0.34	0.20	0.35	0.41	0.34	0.26	0.35	0.30	0.32	0.48	0.40	0.35	0.27	0.43	0.40	0.30	0.21	0.33
gpt-3.5-turbo-0125	0.25	0.04	0.24	0.19	0.14	0.18	0.03	0.16	0.22	0.22	0.12	0.10	0.10	0.12	0.40	0.29	0.23	0.06	0.28	0.30	0.11	0.03	0.12
gpt-4-0613	0.34	0.18	0.34	0.24	0.19	0.27	0.12	0.25	0.34	0.30	0.21	0.27	0.23	0.19	0.34	0.33	0.32	0.17	0.33	0.37	0.30	0.15	0.24
-turbo-2024-04-09	0.51	0.56	0.59	0.51	0.57	0.58	0.62	0.60	0.62	0.67	0.59	0.59	0.63	0.58	0.57	0.52	0.59	0.59	0.48	0.51	0.63	0.65	0.59
gpt-4o-2024-05-13	0.53	0.59	0.65	0.47	0.60	0.59	0.65	0.63	0.67	0.68	0.59	0.64	0.67	0.56	0.48	0.50	0.58	0.65	0.52	0.51	0.70	0.68	0.58
wen1.5-110B-Chat	0.61	0.77	0.64	0.53	0.65	0.66	0.81	0.68	0.69	0.75	0.68	0.63	0.62	0.69	0.56	0.63	0.61	0.71	0.59	0.60	0.72	0.78	0.66
qwen1.5-32B-Chat	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
qwen1.5-72B-Chat	0.48	0.47	0.52	0.44	0.52	0.49	0.53	0.45	0.49	0.44	0.47	0.48	0.49	0.47	0.50	0.49	0.49	0.48	0.49	0.50	0.55	0.51	0.50
dbrx-instruct	0.32	0.32	0.47	0.38	0.35	0.39	0.30	0.36	0.42	0.42	0.36	0.41	0.40	0.31	0.47	0.43	0.42	0.33	0.45	0.40	0.37	0.27	0.41
llama-3-70b-chat	0.53	0.27	0.47	0.40	0.44	0.45	0.31	0.50	0.67	0.40	0.38	0.40	0.43	0.36	0.42	0.46	0.41	0.56	0.63	0.43	0.44	0.36	0.46
llama-3-8b-chat	0.42	0.09	0.38	0.32	0.26	0.30	0.07	0.33	0.45	0.28	0.14	0.24	0.20	0.20	0.38	0.41	0.28	0.30	0.51	0.34	0.28	0.13	0.30
gemini-1.0-pro	0.51	0.54	0.62	0.47	0.48	0.51	0.41	0.51	0.47	0.55	0.54	0.52	0.54	0.46	0.48	0.47	0.52	0.45	0.54	0.50	0.53	0.52	0.52
gemini-1.5-pro	0.10	0.06	0.15	0.09	0.12	0.12	0.02	0.10	0.10	0.10	0.06	0.21	0.02	0.06	0.10	0.18	0.06	0.02	0.28	0.29	0.26	0.01	0.06

💡 Insights

most council members genuinely liked Qwen-110B across the board.

command-r is the least favorable, but it still gives it a positive win rate.

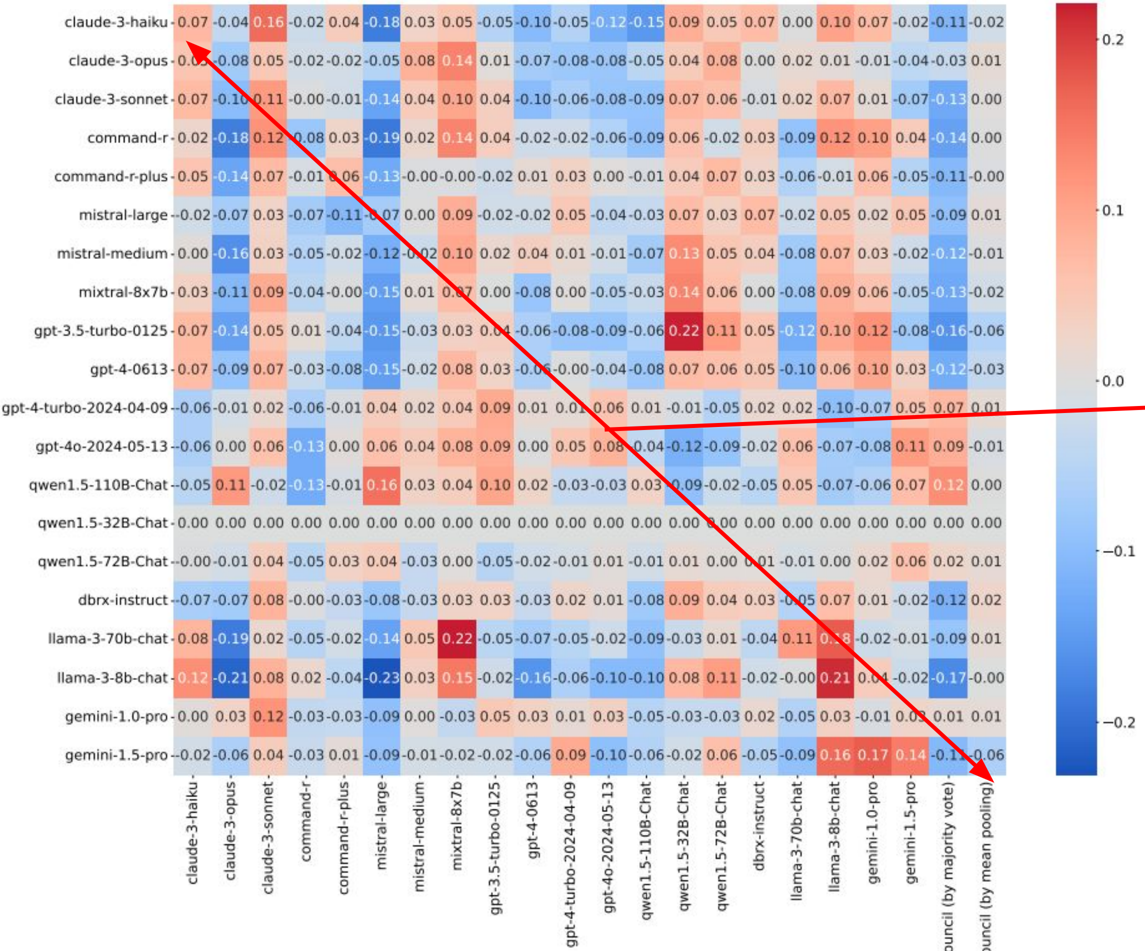
Judge vs. Respondent (Council-Normalized) (affinities)



	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	ouncil (by majority vote)	ouncil (by mean pooling)
claude-3-haiku	0.07	-0.04	0.16	-0.02	0.04	-0.18	0.03	0.05	-0.05	-0.10	-0.05	-0.12	-0.15	0.09	0.05	0.07	0.00	0.10	0.07	-0.02	-0.11	-0.02
claude-3-opus	0.05	-0.08	0.05	-0.02	-0.02	-0.05	0.08	0.14	0.01	-0.07	-0.08	-0.08	-0.05	0.04	0.08	0.00	0.02	0.01	-0.01	-0.04	-0.03	0.01
claude-3-sonnet	0.07	-0.10	0.11	-0.00	-0.01	-0.14	0.04	0.10	0.04	-0.10	-0.06	-0.08	-0.09	0.07	0.06	-0.01	0.02	0.07	0.01	-0.07	-0.13	0.00
command-r	0.02	-0.18	0.12	-0.08	0.03	-0.19	0.02	0.14	0.04	-0.02	-0.02	-0.06	-0.09	0.06	-0.02	0.03	-0.09	0.12	0.10	0.04	-0.14	0.00
command-r-plus	0.05	-0.14	0.07	-0.01	0.06	-0.13	-0.00	-0.00	-0.02	0.01	0.03	0.00	-0.01	0.04	0.07	0.03	-0.06	-0.01	0.06	-0.05	-0.11	-0.00
mistral-large	-0.02	-0.07	0.03	-0.07	-0.11	-0.07	0.00	0.09	-0.02	-0.02	0.05	-0.04	-0.03	0.07	0.03	0.07	-0.02	0.05	0.02	0.05	-0.09	0.01
mistral-medium	0.00	-0.16	0.03	-0.05	-0.02	-0.12	-0.02	0.10	0.02	0.04	0.01	-0.01	-0.07	0.13	0.05	0.04	-0.08	0.07	0.03	-0.02	-0.12	-0.01
mixtral-8x7b	0.03	-0.11	0.09	-0.04	-0.00	-0.15	0.01	0.07	0.00	-0.08	0.00	-0.05	-0.03	0.14	0.06	0.00	-0.08	0.09	0.06	-0.05	-0.13	-0.02
gpt-3.5-turbo-0125	0.07	-0.14	0.05	0.01	-0.04	-0.15	-0.03	0.03	0.04	-0.06	-0.08	-0.09	-0.06	0.22	0.11	0.05	-0.12	0.10	0.12	-0.08	-0.16	-0.06
gpt-4-0613	0.07	-0.09	0.07	-0.03	-0.08	-0.15	-0.02	0.08	0.03	-0.06	-0.00	-0.04	-0.08	0.07	0.06	0.05	-0.10	0.06	0.10	0.03	-0.12	-0.03
gpt-4-turbo-2024-04-09	-0.06	-0.01	0.02	-0.06	-0.01	0.04	0.02	0.04	0.09	0.01	0.01	0.06	0.01	-0.01	-0.05	0.02	0.02	-0.10	-0.07	0.05	0.07	0.01
gpt-4o-2024-05-13	-0.06	0.00	0.06	-0.13	0.00	0.06	0.04	0.08	0.09	0.00	0.05	0.08	-0.04	-0.12	-0.09	-0.02	0.06	-0.07	-0.08	0.11	0.09	-0.01
qwen1.5-110B-Chat	-0.05	0.11	-0.02	-0.13	-0.01	0.16	0.03	0.04	0.10	0.02	-0.03	-0.03	0.03	-0.09	-0.02	-0.05	0.05	-0.07	-0.06	0.07	0.12	0.00
qwen1.5-32B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
qwen1.5-72B-Chat	-0.00	-0.01	0.04	-0.05	0.03	0.04	-0.03	0.00	-0.05	-0.02	-0.01	0.01	-0.01	0.01	0.00	0.01	-0.01	0.00	0.02	0.06	0.02	0.01
dbrx-instruct	-0.07	-0.07	0.08	-0.00	-0.03	-0.08	-0.03	0.03	0.03	-0.03	0.02	0.01	-0.08	0.09	0.04	0.03	-0.05	0.07	0.01	-0.02	-0.12	0.02
llama-3-70b-chat	0.08	-0.19	0.02	-0.05	-0.02	-0.14	0.05	0.22	-0.05	-0.07	-0.05	-0.02	-0.09	-0.03	0.01	-0.04	0.11	0.18	-0.02	-0.01	-0.09	0.01
llama-3-8b-chat	0.12	-0.21	0.08	0.02	-0.04	-0.23	0.03	0.15	-0.02	-0.16	-0.06	-0.10	-0.10	0.08	0.11	-0.02	-0.00	0.21	0.04	-0.02	-0.17	-0.00
gemini-1.0-pro	0.00	0.03	0.12	-0.03	-0.03	-0.09	0.00	-0.03	0.05	0.03	0.01	0.03	-0.05	-0.03	-0.03	0.02	-0.05	0.03	-0.01	0.03	0.01	0.01
gemini-1.5-pro	-0.02	-0.06	0.04	-0.03	0.01	-0.09	-0.01	-0.02	-0.02	-0.06	0.09	-0.10	-0.06	-0.02	0.06	-0.05	-0.09	0.16	0.17	0.14	-0.11	-0.06

💡 Insights

Judge vs. Respondent (Council-Normalized) (affinities)



Respondents

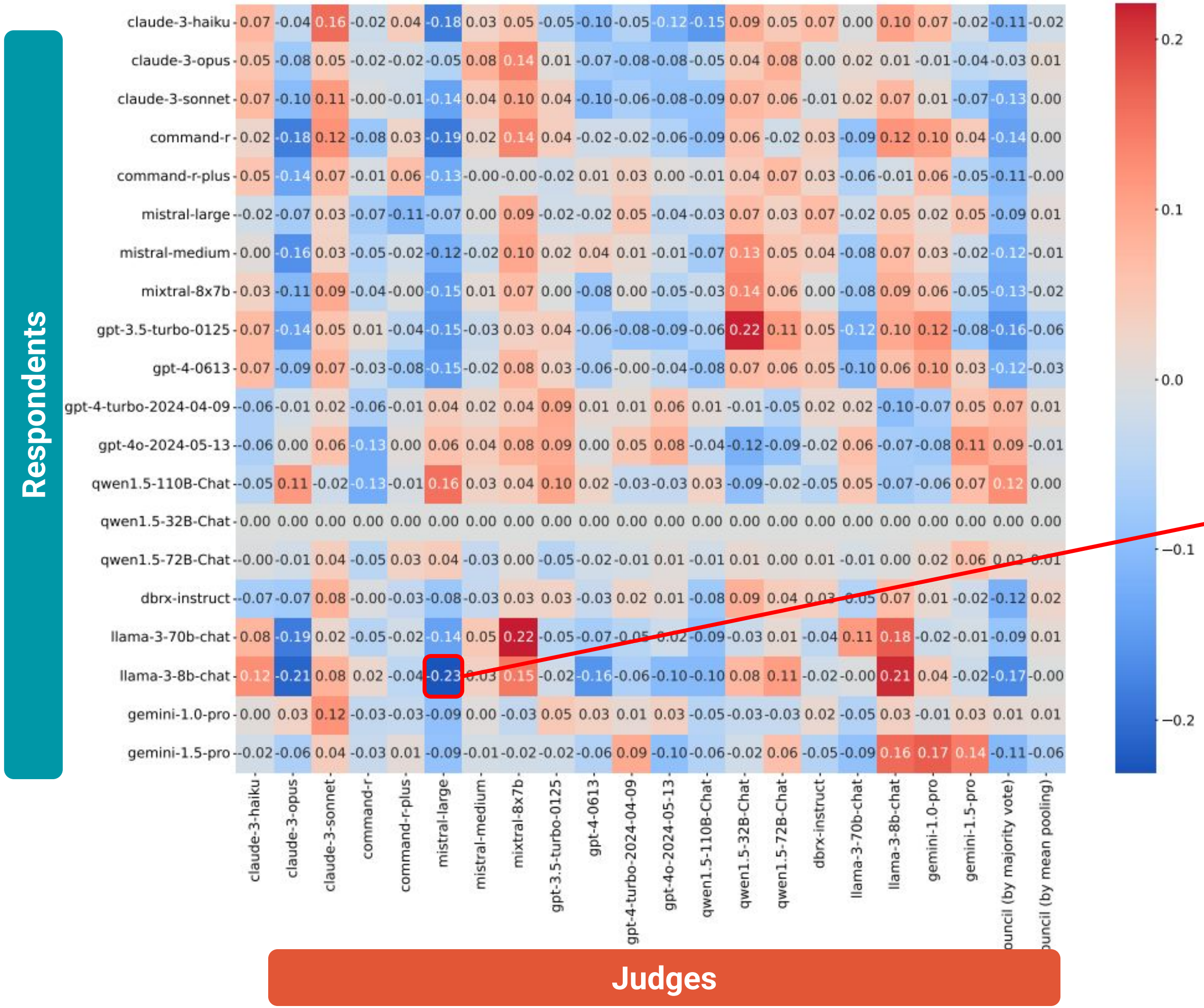
	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	ouncil (by majority vote)	ouncil (by mean pooling)
claude-3-haiku	0.07	-0.04	0.16	-0.02	0.04	-0.18	0.03	0.05	-0.05	-0.10	-0.05	-0.12	-0.15	0.09	0.05	0.07	0.00	0.10	0.07	-0.02	-0.11	-0.02
claude-3-opus	0.05	-0.08	0.05	-0.02	-0.02	-0.05	0.08	0.14	0.01	-0.07	-0.08	-0.08	-0.05	0.04	0.08	0.00	0.02	0.01	-0.01	-0.04	-0.03	0.01
claude-3-sonnet	0.07	-0.10	0.11	-0.00	-0.01	-0.14	0.04	0.10	0.04	-0.10	-0.06	-0.08	-0.09	0.07	0.06	-0.01	0.02	0.07	0.01	-0.07	-0.13	0.00
command-r	0.02	-0.18	0.12	-0.08	0.03	-0.19	0.02	0.14	0.04	-0.02	-0.02	-0.06	-0.09	0.06	-0.02	0.03	-0.09	0.12	0.10	0.04	-0.14	0.00
command-r-plus	0.05	-0.14	0.07	-0.01	0.06	-0.13	-0.00	-0.00	-0.02	0.01	0.03	0.00	-0.01	0.04	0.07	0.03	-0.06	-0.01	0.06	-0.05	-0.11	-0.00
mistral-large	-0.02	-0.07	0.03	-0.07	-0.11	-0.07	0.00	0.09	-0.02	-0.02	0.05	-0.04	-0.03	0.07	0.03	0.07	-0.02	0.05	0.02	0.05	-0.09	0.01
mistral-medium	0.00	-0.16	0.03	-0.05	-0.02	-0.12	-0.02	0.10	0.02	0.04	0.01	-0.01	-0.07	0.13	0.05	0.04	-0.08	0.07	0.03	-0.02	-0.12	-0.01
mixtral-8x7b	0.03	-0.11	0.09	-0.04	-0.00	-0.15	0.01	0.07	0.00	-0.08	0.00	-0.05	-0.03	0.14	0.06	0.00	-0.08	0.09	0.06	-0.05	-0.13	-0.02
gpt-3.5-turbo-0125	0.07	-0.14	0.05	0.01	-0.04	-0.15	-0.03	0.03	0.04	-0.06	-0.08	-0.09	-0.06	0.22	0.11	0.05	-0.12	0.10	0.12	-0.08	-0.16	-0.06
gpt-4-0613	0.07	-0.09	0.07	-0.03	-0.08	-0.15	-0.02	0.08	0.03	-0.06	-0.00	-0.04	-0.08	0.07	0.06	0.05	-0.10	0.06	0.10	0.03	-0.12	-0.03
gpt-4-turbo-2024-04-09	-0.06	-0.01	0.02	-0.06	-0.01	0.04	0.02	0.04	0.09	0.01	0.01	0.06	0.01	-0.01	-0.05	0.02	0.02	-0.10	-0.07	0.05	0.07	0.01
gpt-4o-2024-05-13	-0.06	0.00	0.06	-0.13	0.00	0.06	0.04	0.08	0.09	0.00	0.05	0.08	-0.04	-0.12	-0.09	-0.02	0.06	-0.07	-0.08	0.11	0.09	-0.01
qwen1.5-110B-Chat	-0.05	0.11	-0.02	-0.13	-0.01	0.16	0.03	0.04	0.10	0.02	-0.03	-0.03	0.03	-0.09	-0.02	-0.05	0.05	-0.07	-0.06	0.07	0.12	0.00
qwen1.5-32B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
qwen1.5-72B-Chat	-0.00	-0.01	0.04	-0.05	0.03	0.04	-0.03	0.00	-0.05	-0.02	-0.01	0.01	-0.01	0.01	0.00	0.01	-0.01	0.00	0.02	0.06	0.02	0.01
dbrx-instruct	-0.07	-0.07	0.08	-0.00	-0.03	-0.08	-0.03	0.03	0.03	-0.03	0.02	0.01	-0.08	0.09	0.04	0.03	-0.05	0.07	0.01	-0.02	-0.12	0.02
llama-3-70b-chat	0.08	-0.19	0.02	-0.05	-0.02	-0.14	0.05	0.22	-0.05	-0.07	-0.05	-0.02	-0.09	-0.03	0.01	-0.04	0.11	0.18	-0.02	-0.01	-0.09	0.01
llama-3-8b-chat	0.12	-0.21	0.08	0.02	-0.04	-0.23	0.03	0.15	-0.02	-0.16	-0.06	-0.10	-0.10	0.08	0.11	-0.02	-0.00	0.21	0.04	-0.02	-0.17	-0.00
gemini-1.0-pro	0.00	0.03	0.12	-0.03	-0.03	-0.09	0.00	-0.03	0.05	0.03	0.01	0.03	-0.05	-0.03	-0.03	0.02	-0.05	0.03	-0.01	0.03	0.01	0.01
gemini-1.5-pro	-0.02	-0.06	0.04	-0.03	0.01	-0.09	-0.01	-0.02	-0.02	-0.06	0.09	-0.10	-0.06	-0.02	0.06	-0.05	-0.09	0.16	0.17	0.14	-0.11	-0.06

Judges

💡 Insights

self-enhancement bias, measured.

Judge vs. Respondent (Council-Normalized) (affinities)



Respondents

Judges

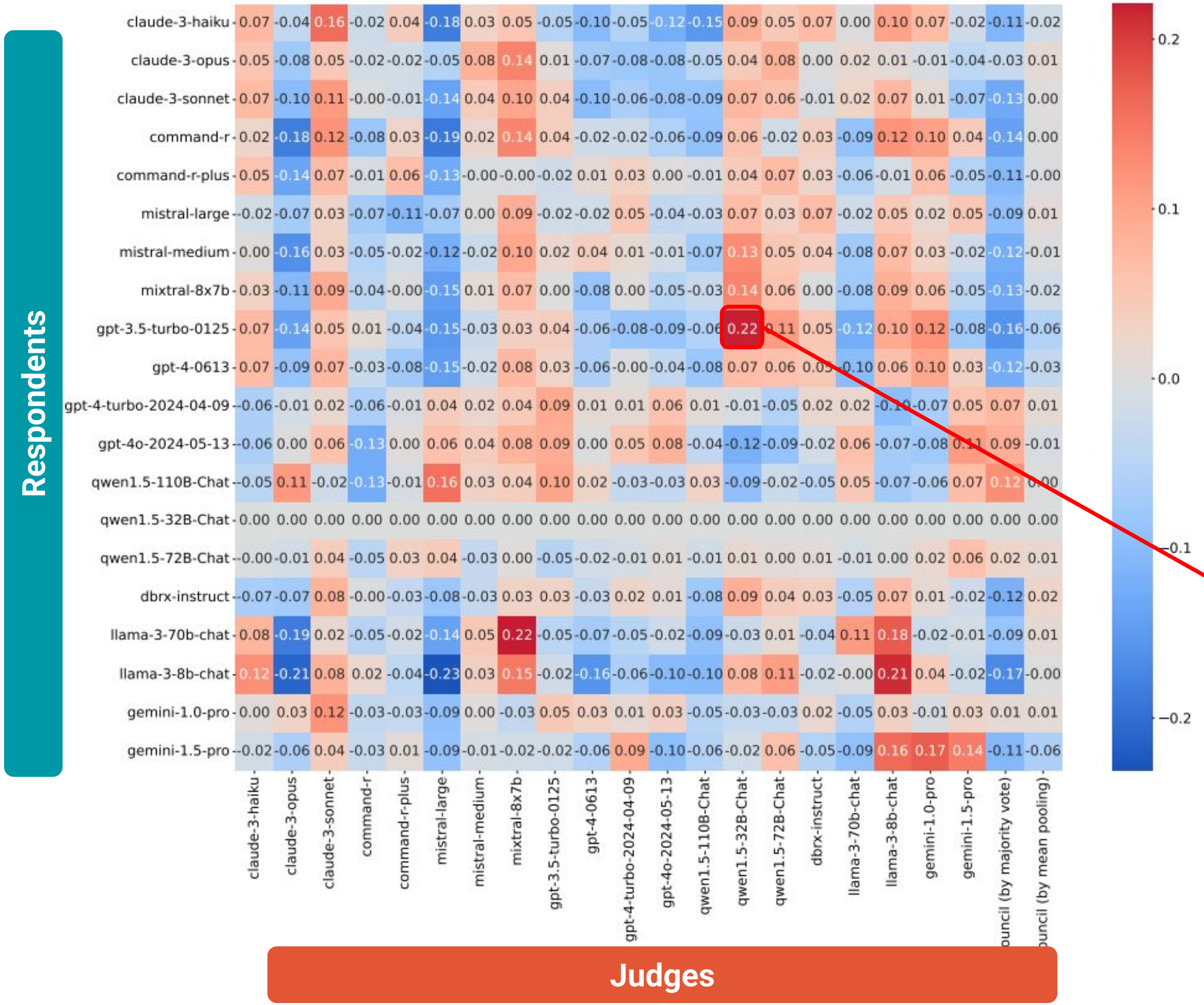
	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	ouncil (by majority vote)	ouncil (by mean pooling)
claude-3-haiku	0.07	-0.04	0.16	-0.02	0.04	-0.18	0.03	0.05	-0.05	-0.10	-0.05	-0.12	-0.15	0.09	0.05	0.07	0.00	0.10	0.07	-0.02	-0.11	-0.02
claude-3-opus	0.05	-0.08	0.05	-0.02	-0.02	-0.05	0.08	0.14	0.01	-0.07	-0.08	-0.08	-0.05	0.04	0.08	0.00	0.02	0.01	-0.01	-0.04	-0.03	0.01
claude-3-sonnet	0.07	-0.10	0.11	-0.00	-0.01	-0.14	0.04	0.10	0.04	-0.10	-0.06	-0.08	-0.09	0.07	0.06	-0.01	0.02	0.07	0.01	-0.07	-0.13	0.00
command-r	0.02	-0.18	0.12	-0.08	0.03	-0.19	0.02	0.14	0.04	-0.02	-0.02	-0.06	-0.09	0.06	-0.02	0.03	-0.09	0.12	0.10	0.04	-0.14	0.00
command-r-plus	0.05	-0.14	0.07	-0.01	0.06	-0.13	-0.00	-0.00	-0.02	0.01	0.03	0.00	-0.01	0.04	0.07	0.03	-0.06	-0.01	0.06	-0.05	-0.11	-0.00
mistral-large	-0.02	-0.07	0.03	-0.07	-0.11	-0.07	0.00	0.09	-0.02	-0.02	0.05	-0.04	-0.03	0.07	0.03	0.07	-0.02	0.05	0.02	0.05	-0.09	0.01
mistral-medium	0.00	-0.16	0.03	-0.05	-0.02	-0.12	-0.02	0.10	0.02	0.04	0.01	-0.01	-0.07	0.13	0.05	0.04	-0.08	0.07	0.03	-0.02	-0.12	-0.01
mixtral-8x7b	0.03	-0.11	0.09	-0.04	-0.00	-0.15	0.01	0.07	0.00	-0.08	0.00	-0.05	-0.03	0.14	0.06	0.00	-0.08	0.09	0.06	-0.05	-0.13	-0.02
gpt-3.5-turbo-0125	0.07	-0.14	0.05	0.01	-0.04	-0.15	-0.03	0.03	0.04	-0.06	-0.08	-0.09	-0.06	0.22	0.11	0.05	-0.12	0.10	0.12	-0.08	-0.16	-0.06
gpt-4-0613	0.07	-0.09	0.07	-0.03	-0.08	-0.15	-0.02	0.08	0.03	-0.06	-0.00	-0.04	-0.08	0.07	0.06	0.05	-0.10	0.06	0.10	0.03	-0.12	-0.03
gpt-4-turbo-2024-04-09	-0.06	-0.01	0.02	-0.06	-0.01	0.04	0.02	0.04	0.09	0.01	0.01	0.06	0.01	-0.01	-0.05	0.02	0.02	-0.10	-0.07	0.05	0.07	0.01
gpt-4o-2024-05-13	-0.06	0.00	0.06	-0.13	0.00	0.06	0.04	0.08	0.09	0.00	0.05	0.08	-0.04	-0.12	-0.09	-0.02	0.06	-0.07	-0.08	0.11	0.09	-0.01
qwen1.5-110B-Chat	-0.05	0.11	-0.02	-0.13	-0.01	0.16	0.03	0.04	0.10	0.02	-0.03	-0.03	0.03	-0.09	-0.02	-0.05	0.05	-0.07	-0.06	0.07	0.12	0.00
qwen1.5-32B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
qwen1.5-72B-Chat	-0.00	-0.01	0.04	-0.05	0.03	0.04	-0.03	0.00	-0.05	-0.02	-0.01	0.01	-0.01	0.01	0.00	0.01	-0.01	0.00	0.02	0.06	0.02	-0.01
dbrx-instruct	-0.07	-0.07	0.08	-0.00	-0.03	-0.08	-0.03	0.03	0.03	-0.03	0.02	0.01	-0.08	0.09	0.04	0.03	-0.05	0.07	0.01	-0.02	-0.12	0.02
llama-3-70b-chat	0.08	-0.19	0.02	-0.05	-0.02	-0.14	0.05	0.22	-0.05	-0.07	-0.05	-0.02	-0.09	-0.03	0.01	-0.04	0.11	0.18	-0.02	-0.01	-0.09	0.01
llama-3-8b-chat	0.12	-0.21	0.08	0.02	-0.04	-0.23	0.03	0.15	-0.02	-0.16	-0.06	-0.10	-0.10	0.08	0.11	-0.02	-0.00	0.21	0.04	-0.02	-0.17	-0.00
gemini-1.0-pro	0.00	0.03	0.12	-0.03	-0.03	-0.09	0.00	-0.03	0.05	0.03	0.01	0.03	-0.05	-0.03	-0.03	0.02	-0.05	0.03	-0.01	0.03	0.01	0.01
gemini-1.5-pro	-0.02	-0.06	0.04	-0.03	0.01	-0.09	-0.01	-0.02	-0.02	-0.06	0.09	-0.10	-0.06	-0.02	0.06	-0.05	-0.09	0.16	0.17	0.14	-0.11	-0.06

💡 Insights

self-enhancement bias, measured.

mistral-large hates llama-3-8b

Judge vs. Respondent (Council-Normalized) (affinities)



	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	ouncil (by majority vote)	ouncil (by mean pooling)
claude-3-haiku	0.07	-0.04	0.16	-0.02	0.04	-0.18	0.03	0.05	-0.05	-0.10	-0.05	-0.12	-0.15	0.09	0.05	0.07	0.00	0.10	0.07	-0.02	-0.11	-0.02
claude-3-opus	0.05	-0.08	0.05	-0.02	-0.02	-0.05	0.08	0.14	0.01	-0.07	-0.08	-0.08	-0.05	0.04	0.08	0.00	0.02	0.01	-0.01	-0.04	-0.03	0.01
claude-3-sonnet	0.07	-0.10	0.11	-0.00	-0.01	-0.14	0.04	0.10	0.04	-0.10	-0.06	-0.08	-0.09	0.07	0.06	-0.01	0.02	0.07	0.01	-0.07	-0.13	0.00
command-r	0.02	-0.18	0.12	-0.08	0.03	-0.19	0.02	0.14	0.04	-0.02	-0.02	-0.06	-0.09	0.06	-0.02	0.03	-0.09	0.12	0.10	0.04	-0.14	0.00
command-r-plus	0.05	-0.14	0.07	-0.01	0.06	-0.13	-0.00	-0.00	-0.02	0.01	0.03	0.00	-0.01	0.04	0.07	0.03	-0.06	-0.01	0.06	-0.05	-0.11	-0.00
mistral-large	-0.02	-0.07	0.03	-0.07	-0.11	-0.07	0.00	0.09	-0.02	-0.02	0.05	-0.04	-0.03	0.07	0.03	0.07	-0.02	0.05	0.02	0.05	-0.09	0.01
mistral-medium	0.00	-0.16	0.03	-0.05	-0.02	-0.12	-0.02	0.10	0.02	0.04	0.01	-0.01	-0.07	0.13	0.05	0.04	-0.08	0.07	0.03	-0.02	-0.12	-0.01
mixtral-8x7b	0.03	-0.11	0.09	-0.04	-0.00	-0.15	0.01	0.07	0.00	-0.08	0.00	-0.05	-0.03	0.14	0.06	0.00	-0.08	0.09	0.06	-0.05	-0.13	-0.02
gpt-3.5-turbo-0125	0.07	-0.14	0.05	0.01	-0.04	-0.15	-0.03	0.03	0.04	-0.06	-0.08	-0.09	-0.06	0.22	0.11	0.05	-0.12	0.10	0.12	-0.08	-0.16	-0.06
gpt-4-0613	0.07	-0.09	0.07	-0.03	-0.08	-0.15	-0.02	0.08	0.03	-0.06	-0.00	-0.04	-0.08	0.07	0.06	0.05	-0.10	0.06	0.10	0.03	-0.12	-0.03
gpt-4-turbo-2024-04-09	-0.06	-0.01	0.02	-0.06	-0.01	0.04	0.02	0.04	0.09	0.01	0.01	0.06	0.01	-0.01	-0.05	0.02	0.02	-0.10	-0.07	0.05	0.07	0.01
gpt-4o-2024-05-13	-0.06	0.00	0.06	-0.13	0.00	0.06	0.04	0.08	0.09	0.00	0.05	0.08	-0.04	-0.12	-0.09	-0.02	0.06	-0.07	-0.08	0.11	0.09	-0.01
qwen1.5-110B-Chat	-0.05	0.11	-0.02	-0.13	-0.01	0.16	0.03	0.04	0.10	0.02	-0.03	-0.03	0.03	-0.09	-0.02	-0.05	0.05	-0.07	-0.06	0.07	0.12	0.00
qwen1.5-32B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
qwen1.5-72B-Chat	-0.00	-0.01	0.04	-0.05	0.03	0.04	-0.03	0.00	-0.05	-0.02	-0.01	0.01	-0.01	0.01	0.00	0.01	-0.01	0.00	0.02	0.06	0.02	0.01
dbrx-instruct	-0.07	-0.07	0.08	-0.00	-0.03	-0.08	-0.03	0.03	0.03	-0.03	0.02	0.01	-0.08	0.09	0.04	0.03	-0.05	0.07	0.01	-0.02	-0.12	0.02
llama-3-70b-chat	0.08	-0.19	0.02	-0.05	-0.02	-0.14	0.05	0.22	-0.05	-0.07	-0.05	-0.02	-0.09	-0.03	0.01	-0.04	0.11	0.18	-0.02	-0.01	-0.09	0.01
llama-3-8b-chat	0.12	-0.21	0.08	0.02	-0.04	-0.23	0.03	0.15	-0.02	-0.16	-0.06	-0.10	-0.10	0.08	0.11	-0.02	-0.00	0.21	0.04	-0.02	-0.17	-0.00
gemini-1.0-pro	0.00	0.03	0.12	-0.03	-0.03	-0.09	0.00	-0.03	0.05	0.03	0.01	0.03	-0.05	-0.03	-0.03	0.02	-0.05	0.03	-0.01	0.03	0.01	0.01
gemini-1.5-pro	-0.02	-0.06	0.04	-0.03	0.01	-0.09	-0.01	-0.02	-0.02	-0.06	0.09	-0.10	-0.06	-0.02	0.06	-0.05	-0.09	0.16	0.17	0.14	-0.11	-0.06

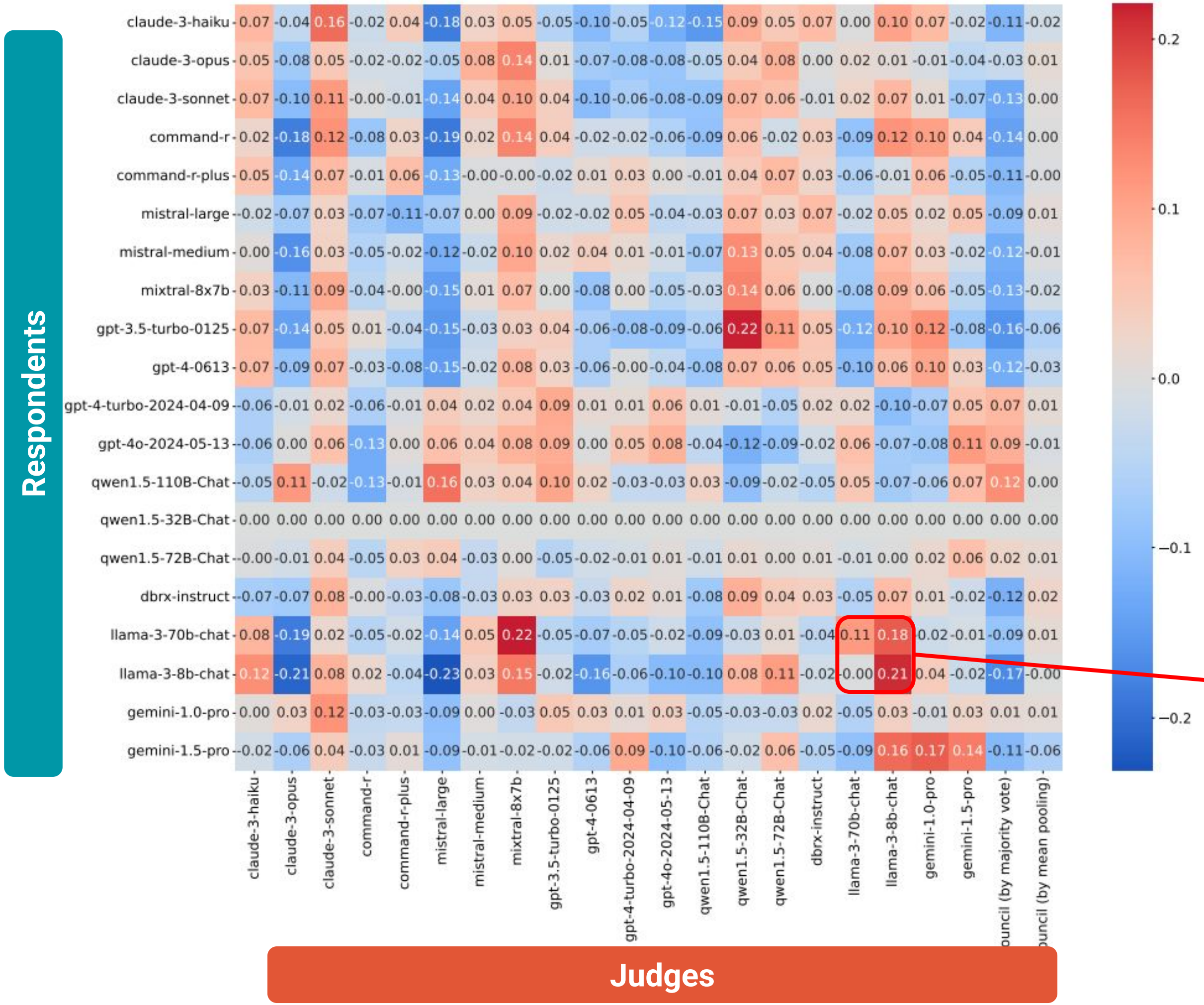
💡 Insights

self-enhancement bias, measured.

mistral-large hates llama-3-8b

qwen 32B loves gpt-3.5

Judge vs. Respondent (Council-Normalized)



	claude-3-haiku	claude-3-opus	claude-3-sonnet	command-r	command-r-plus	mistral-large	mistral-medium	mixtral-8x7b	gpt-3.5-turbo-0125	gpt-4-0613	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	qwen1.5-110B-Chat	qwen1.5-32B-Chat	qwen1.5-72B-Chat	dbrx-instruct	llama-3-70b-chat	llama-3-8b-chat	gemini-1.0-pro	gemini-1.5-pro	ouncil (by majority vote)	ouncil (by mean pooling)
claude-3-haiku	0.07	-0.04	0.16	-0.02	0.04	-0.18	0.03	0.05	-0.05	-0.10	-0.05	-0.12	-0.15	0.09	0.05	0.07	0.00	0.10	0.07	-0.02	-0.11	-0.02
claude-3-opus	0.05	-0.08	0.05	-0.02	-0.02	-0.05	0.08	0.14	0.01	-0.07	-0.08	-0.08	-0.05	0.04	0.08	0.00	0.02	0.01	-0.01	-0.04	-0.03	0.01
claude-3-sonnet	0.07	-0.10	0.11	-0.00	-0.01	-0.14	0.04	0.10	0.04	-0.10	-0.06	-0.08	-0.09	0.07	0.06	-0.01	0.02	0.07	0.01	-0.07	-0.13	0.00
command-r	0.02	-0.18	0.12	-0.08	0.03	-0.19	0.02	0.14	0.04	-0.02	-0.02	-0.06	-0.09	0.06	-0.02	0.03	-0.09	0.12	0.10	0.04	-0.14	0.00
command-r-plus	0.05	-0.14	0.07	-0.01	0.06	-0.13	-0.00	-0.00	-0.02	0.01	0.03	0.00	-0.01	0.04	0.07	0.03	-0.06	-0.01	0.06	-0.05	-0.11	-0.00
mistral-large	-0.02	-0.07	0.03	-0.07	-0.11	-0.07	0.00	0.09	-0.02	-0.02	0.05	-0.04	-0.03	0.07	0.03	0.07	-0.02	0.05	0.02	0.05	-0.09	0.01
mistral-medium	0.00	-0.16	0.03	-0.05	-0.02	-0.12	-0.02	0.10	0.02	0.04	0.01	-0.01	-0.07	0.13	0.05	0.04	-0.08	0.07	0.03	-0.02	-0.12	-0.01
mixtral-8x7b	0.03	-0.11	0.09	-0.04	-0.00	-0.15	0.01	0.07	0.00	-0.08	0.00	-0.05	-0.03	0.14	0.06	0.00	-0.08	0.09	0.06	-0.05	-0.13	-0.02
gpt-3.5-turbo-0125	0.07	-0.14	0.05	0.01	-0.04	-0.15	-0.03	0.03	0.04	-0.06	-0.08	-0.09	-0.06	0.22	0.11	0.05	-0.12	0.10	0.12	-0.08	-0.16	-0.06
gpt-4-0613	0.07	-0.09	0.07	-0.03	-0.08	-0.15	-0.02	0.08	0.03	-0.06	-0.00	-0.04	-0.08	0.07	0.06	0.05	-0.10	0.06	0.10	0.03	-0.12	-0.03
gpt-4-turbo-2024-04-09	-0.06	-0.01	0.02	-0.06	-0.01	0.04	0.02	0.04	0.09	0.01	0.01	0.06	0.01	-0.01	-0.05	0.02	0.02	-0.10	-0.07	0.05	0.07	0.01
gpt-4o-2024-05-13	-0.06	0.00	0.06	-0.13	0.00	0.06	0.04	0.08	0.09	0.00	0.05	0.08	-0.04	-0.12	-0.09	-0.02	0.06	-0.07	-0.08	0.11	0.09	-0.01
qwen1.5-110B-Chat	-0.05	0.11	-0.02	-0.13	-0.01	0.16	0.03	0.04	0.10	0.02	-0.03	-0.03	0.03	-0.09	-0.02	-0.05	0.05	-0.07	-0.06	0.07	0.12	0.00
qwen1.5-32B-Chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
qwen1.5-72B-Chat	-0.00	-0.01	0.04	-0.05	0.03	0.04	-0.03	0.00	-0.05	-0.02	-0.01	0.01	-0.01	0.01	0.00	0.01	-0.01	0.00	0.02	0.06	0.02	0.01
dbrx-instruct	-0.07	-0.07	0.08	-0.00	-0.03	-0.08	-0.03	0.03	0.03	-0.03	0.02	0.01	-0.08	0.09	0.04	0.03	-0.05	0.07	0.01	-0.02	-0.12	0.02
llama-3-70b-chat	0.08	-0.19	0.02	-0.05	-0.02	-0.14	0.05	0.22	-0.05	-0.07	-0.05	-0.02	-0.09	-0.03	0.01	-0.04	0.11	0.18	-0.02	-0.01	-0.09	0.01
llama-3-8b-chat	0.12	-0.21	0.08	0.02	-0.04	-0.23	0.03	0.15	-0.02	-0.16	-0.06	-0.10	-0.10	0.08	0.11	-0.02	-0.00	0.21	0.04	-0.02	-0.17	-0.00
gemini-1.0-pro	0.00	0.03	0.12	-0.03	-0.03	-0.09	0.00	-0.03	0.05	0.03	0.01	0.03	-0.05	-0.03	-0.03	0.02	-0.05	0.03	-0.01	0.03	0.01	0.01
gemini-1.5-pro	-0.02	-0.06	0.04	-0.03	0.01	-0.09	-0.01	-0.02	-0.02	-0.06	0.09	-0.10	-0.06	-0.02	0.06	-0.05	-0.09	0.16	0.17	0.14	-0.11	-0.06

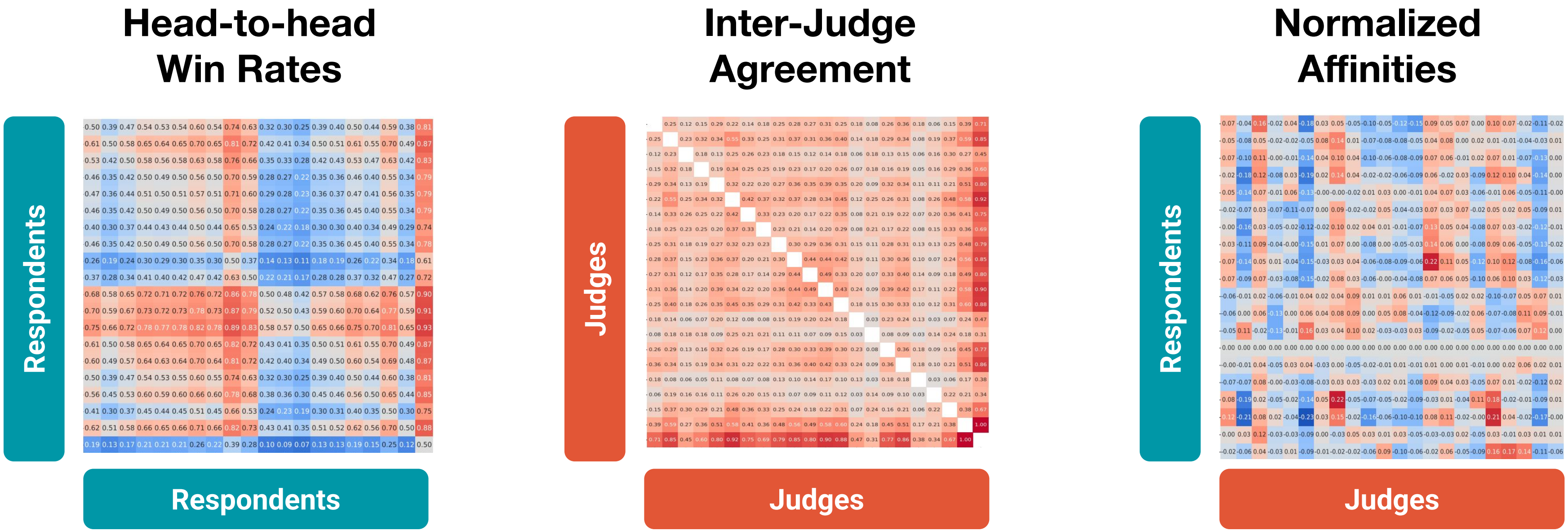
💡 Insights

self-enhancement bias, measured.

mistral-large hates llama-3-8b

qwen 32B loves gpt-3.5

The llama family loves itself



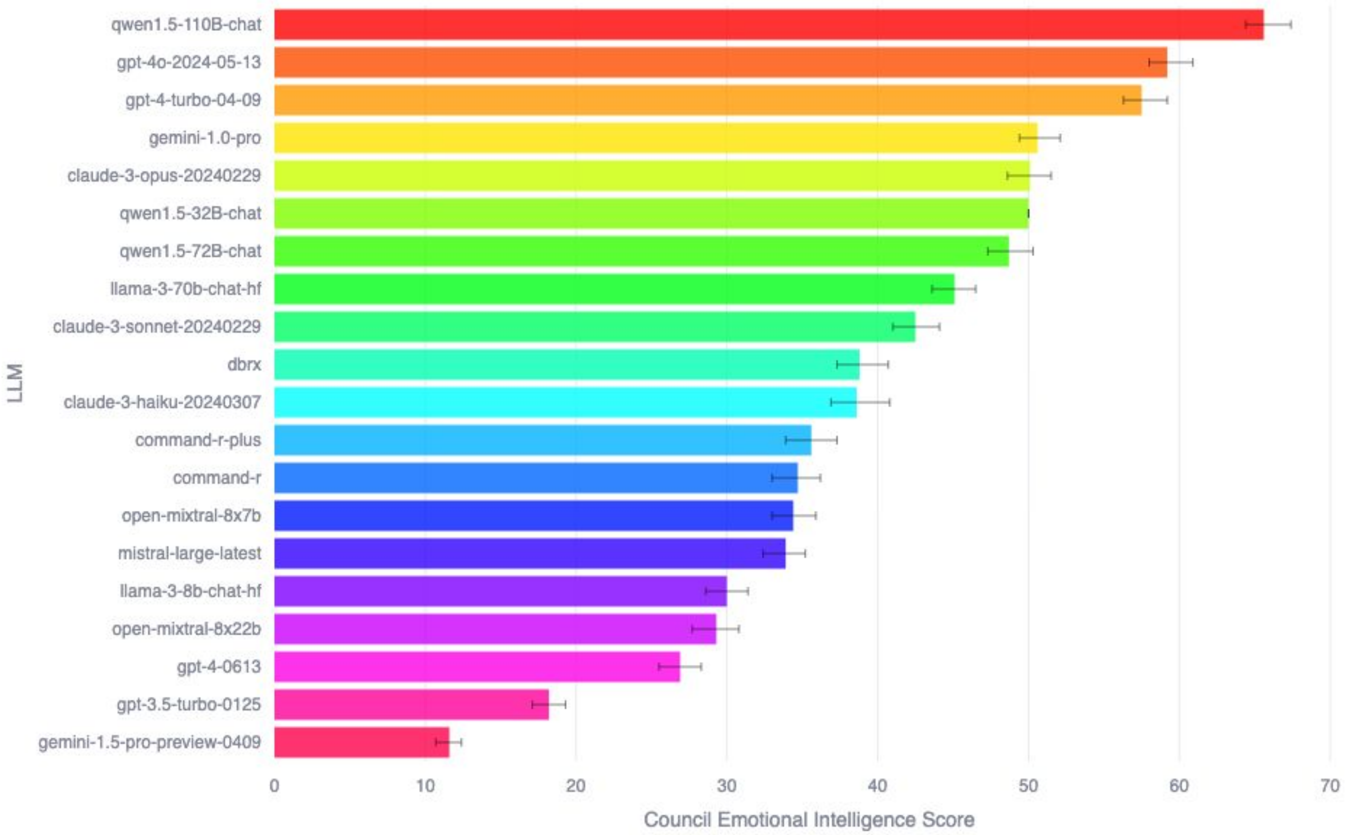
Head-to-head Win Rates

Inter-Judge Agreement

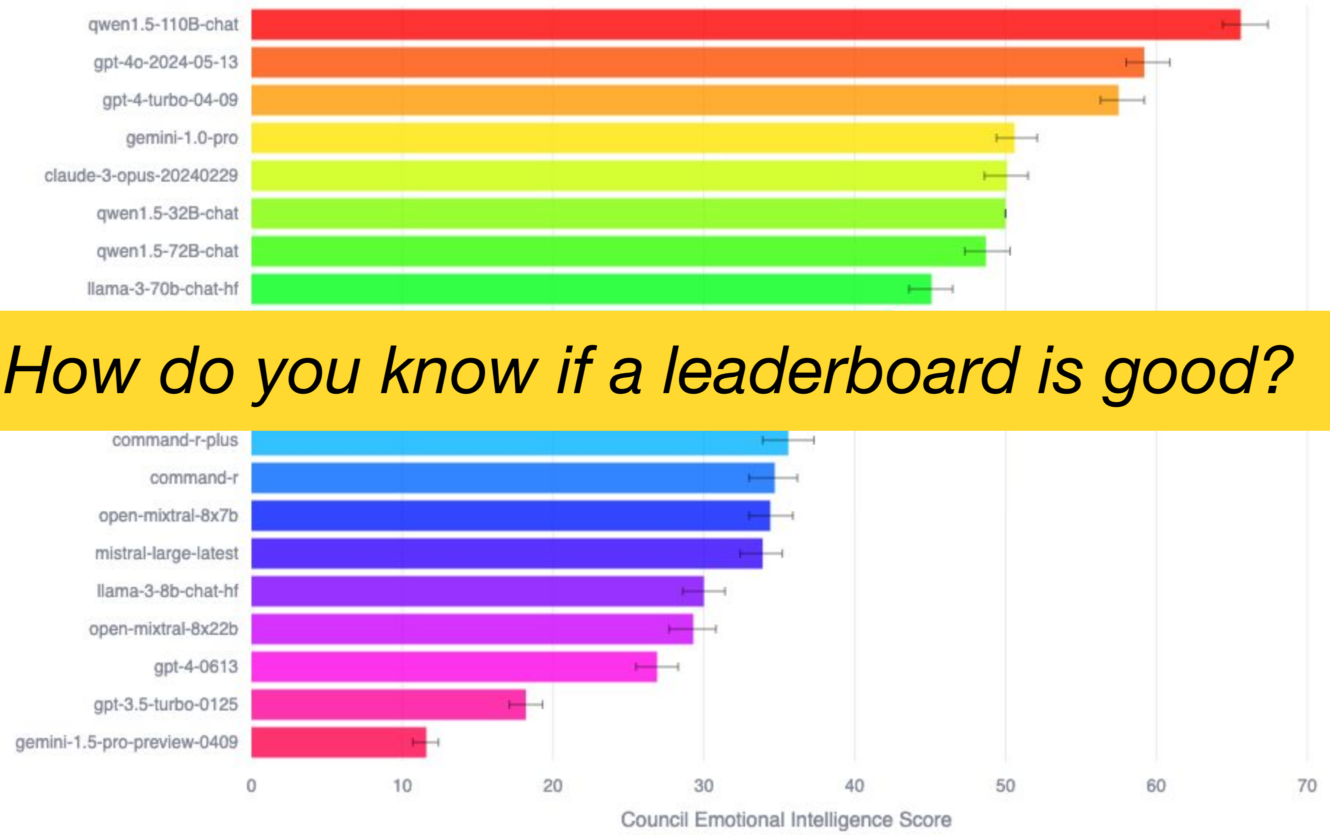
Normalized Affinities

The heatmaps from the fully connected interaction surface of LLM democracy surface rich interaction patterns on behavior and alignment.

Emotional Intelligence Case Study: Final Leaderboard

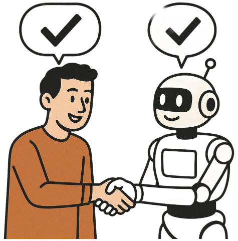


Emotional Intelligence Case Study: Final Leaderboard



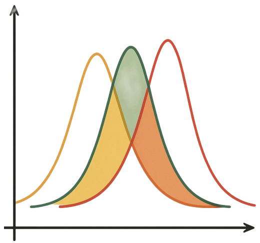
How do you know if a leaderboard is good?

How do you know if a leaderboard is good?



Human agreement

Alignment with meaningful human judgments.



Statistical significance

Stable rankings, not explained by random noise.

Cost and efficiency

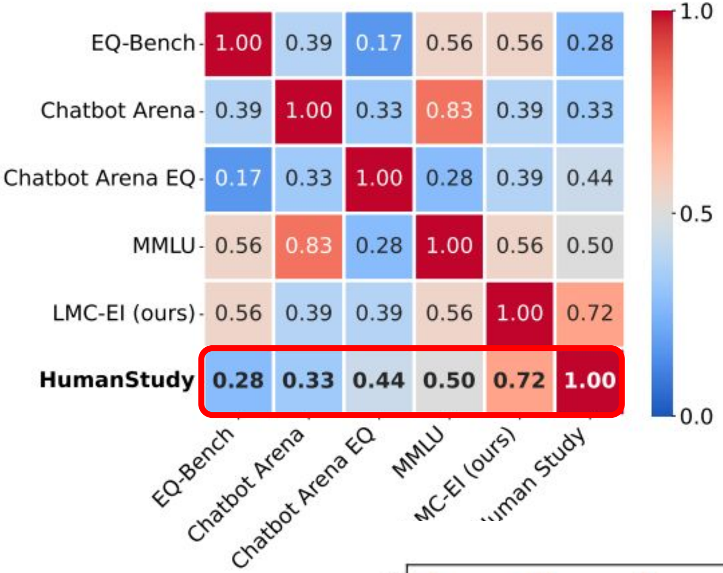
Reasonably quick and **not** cost-prohibitive.

Robustness

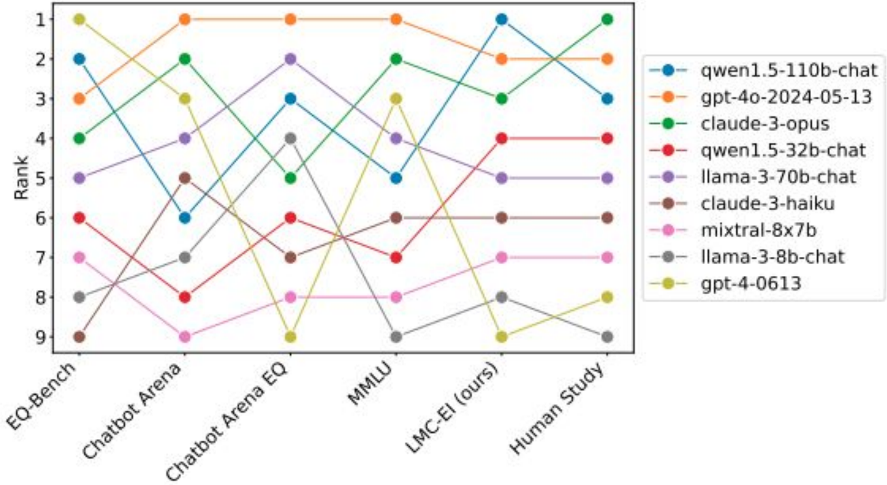
Resilient to **cheating**.

Good leaderboards?

- **human agreement**
- statistical significance
- cost and efficiency
- robustness to cheating



	EQ-Bench	Chatbot Arena	Chatbot Arena EQ	MMLU	LMC-EI (ours)	Human Study
EQ-Bench	1.00	0.39	0.17	0.56	0.56	0.28
Chatbot Arena	0.39	1.00	0.33	0.83	0.39	0.33
Chatbot Arena EQ	0.17	0.33	1.00	0.28	0.39	0.44
MMLU	0.56	0.83	0.28	1.00	0.56	0.50
LMC-EI (ours)	0.56	0.39	0.39	0.56	1.00	0.72
HumanStudy	**0.28**	**0.33**	**0.44**	**0.50**	**0.72**	**1.00**



Good leaderboards?

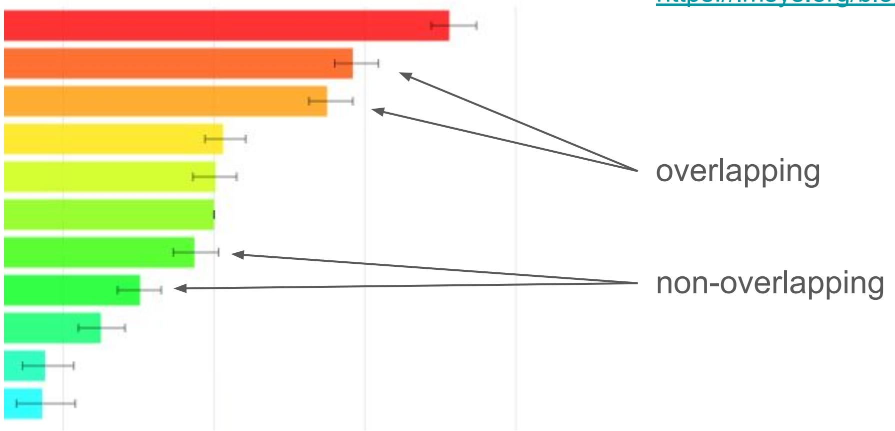
human agreement

statistical significance

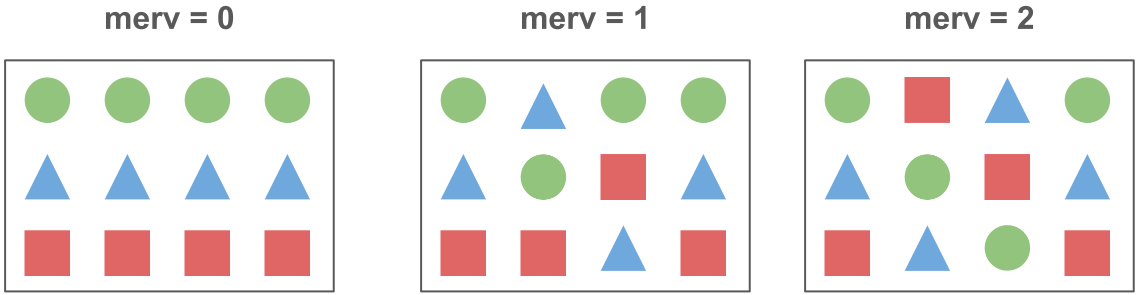
cost and efficiency

robustness to cheating

https://lmsys.org/blog/2024-04-19-arena-hard/



Separability: percentage of model pairs which have non-overlapping confidence intervals of the benchmark scores.



Stability / Consistency: Expected ordinal swing of the average respondent's rank in a new bootstrap trial (MERV).

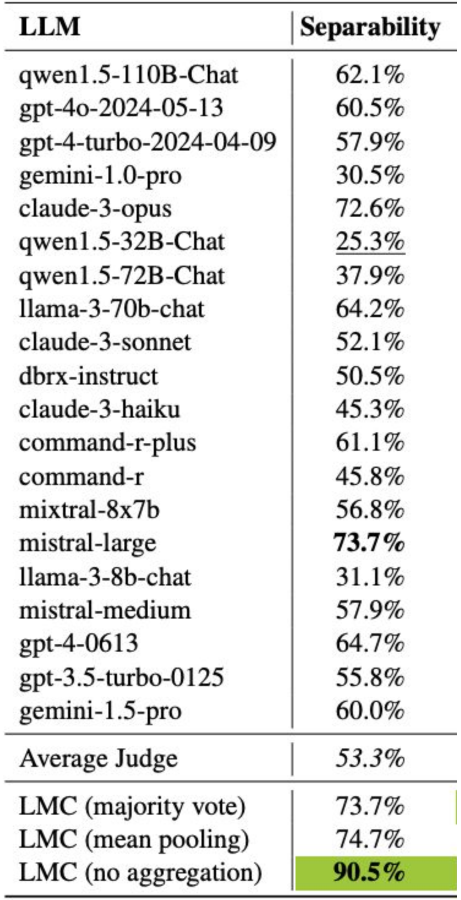
Good leaderboards?

human agreement

statistical significance

cost and efficiency

robustness to cheating

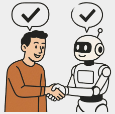


LLM	Separability
qwen1.5-110B-Chat	62.1%
gpt-4o-2024-05-13	60.5%
gpt-4-turbo-2024-04-09	57.9%
gemini-1.0-pro	30.5%
claude-3-opus	72.6%
qwen1.5-32B-Chat	25.3%
qwen1.5-72B-Chat	37.9%
llama-3-70b-chat	64.2%
claude-3-sonnet	52.1%
dbrx-instruct	50.5%
claude-3-haiku	45.3%
command-r-plus	61.1%
command-r	45.8%
mixtral-8x7b	56.8%
mistral-large	**73.7%**
llama-3-8b-chat	31.1%
mistral-medium	57.9%
gpt-4-0613	64.7%
gpt-3.5-turbo-0125	55.8%
gemini-1.5-pro	60.0%
Average Judge	*53.3%*
LMC (majority vote)	73.7%
LMC (mean pooling)	74.7%
LMC (no aggregation)	**90.5%**

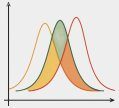
The Council's rankings are more separable than any individual judge.

Good leaderboards?

human agreement

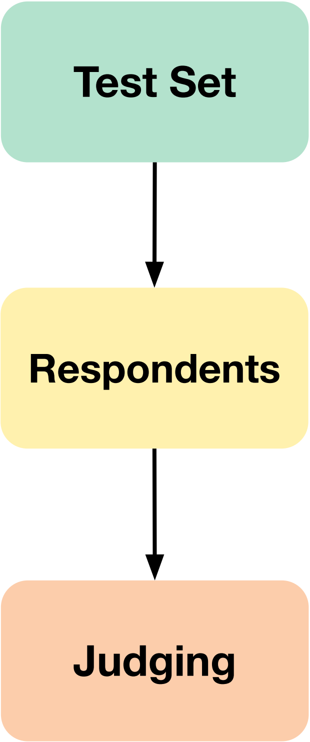


statistical significance



cost and efficiency

robustness to cheating



20 council members
* 5 dilemmas each

100 dilemmas

100 dilemmas
* 20 council members

2000 responses

(2000 responses
- 100) reference model responses
* 20 council judges
* 2 position swap

76000 judgments

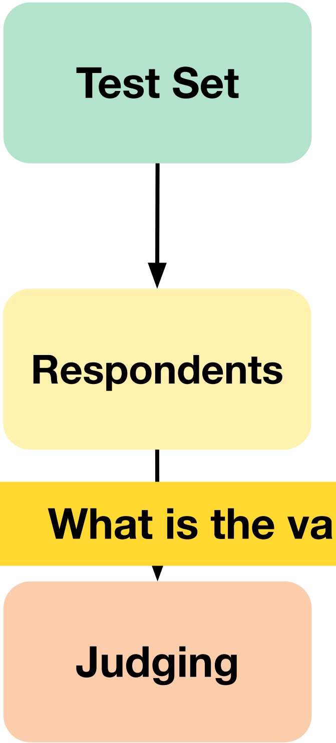
Good leaderboards?

human agreement

statistical significance

cost and efficiency

robustness to cheating



20	council members
* 5	dilemmas each
100	**dilemmas**

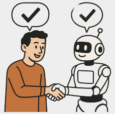
100	dilemmas
* 20	council members
2000	**responses**

What is the value of the incremental judge?

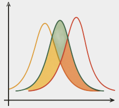
(2000	responses
- 100)	reference model responses
* 20	council judges
* 2	position swap
76000	**judgments**

Good leaderboards?

human agreement



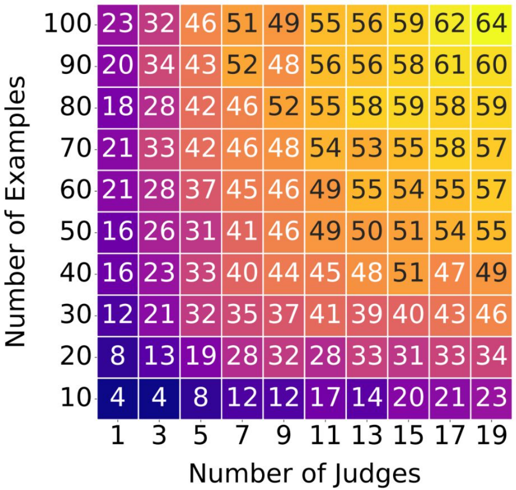
statistical significance



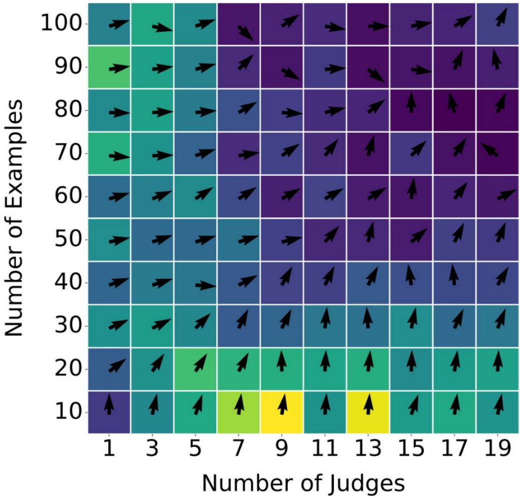
cost and efficiency

robustness to cheating

Monte carlo simulations of hypothetical councils



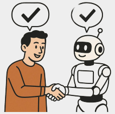
(c) Separability μ



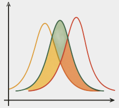
(d) Separability μ gradients

Good leaderboards?

human agreement



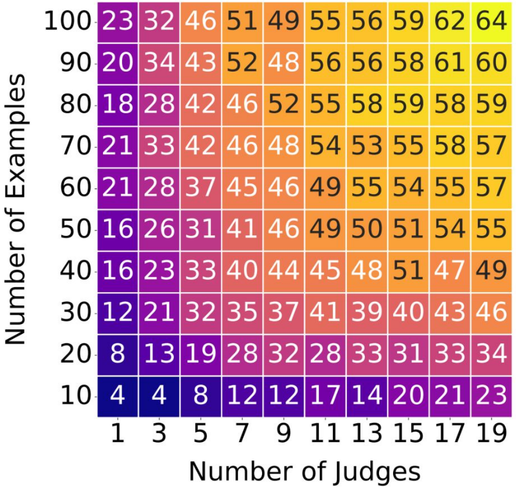
statistical significance



cost and efficiency

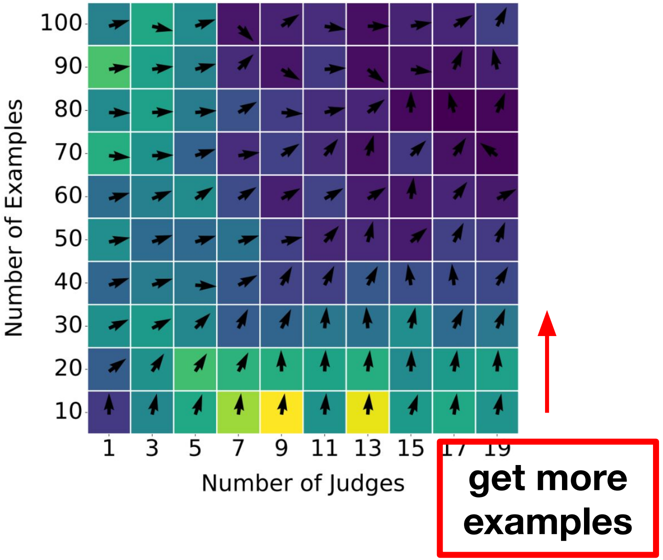
robustness to cheating

Monte carlo simulations of hypothetical councils



Number of Examples	1	3	5	7	9	11	13	15	17	19
100	23	32	46	51	49	55	56	59	62	64
90	20	34	43	52	48	56	56	58	61	60
80	18	28	42	46	52	55	58	59	58	59
70	21	33	42	46	48	54	53	55	58	57
60	21	28	37	45	46	49	55	54	55	57
50	16	26	31	41	46	49	50	51	54	55
40	16	23	33	40	44	45	48	51	47	49
30	12	21	32	35	37	41	39	40	43	46
20	8	13	19	28	32	28	33	31	33	34
10	4	4	8	12	12	17	14	20	21	23

(c) Separability μ

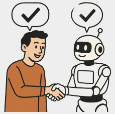


get more examples

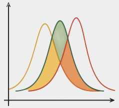
(d) Separability μ gradients

Good leaderboards?

human agreement



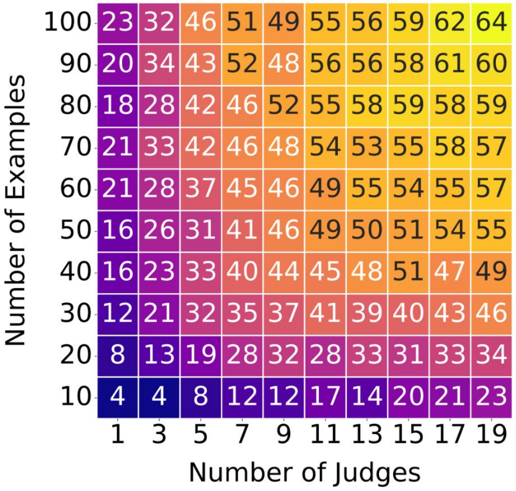
statistical significance



cost and efficiency

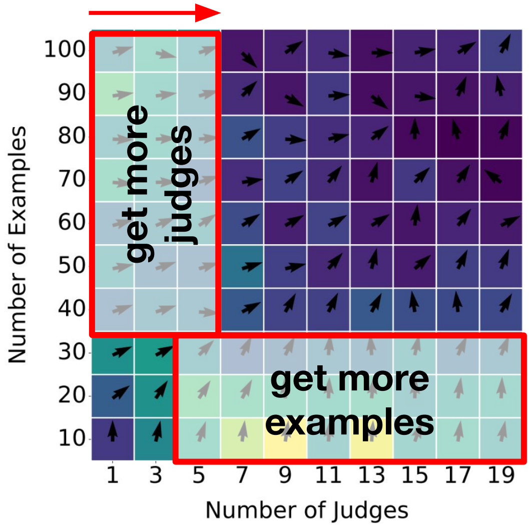
robustness to cheating

Monte carlo simulations of hypothetical councils



Number of Examples \ Number of Judges	1	3	5	7	9	11	13	15	17	19
100	23	32	46	51	49	55	56	59	62	64
90	20	34	43	52	48	56	56	58	61	60
80	18	28	42	46	52	55	58	59	58	59
70	21	33	42	46	48	54	53	55	58	57
60	21	28	37	45	46	49	55	54	55	57
50	16	26	31	41	46	49	50	51	54	55
40	16	23	33	40	44	45	48	51	47	49
30	12	21	32	35	37	41	39	40	43	46
20	8	13	19	28	32	28	33	31	33	34
10	4	4	8	12	12	17	14	20	21	23

(c) Separability μ

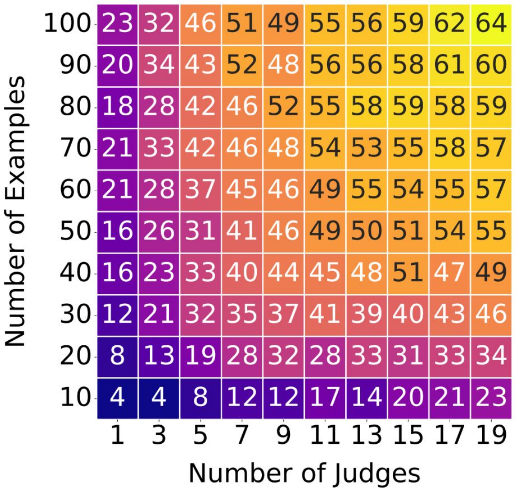


(d) Separability μ gradients

Good leaderboards?

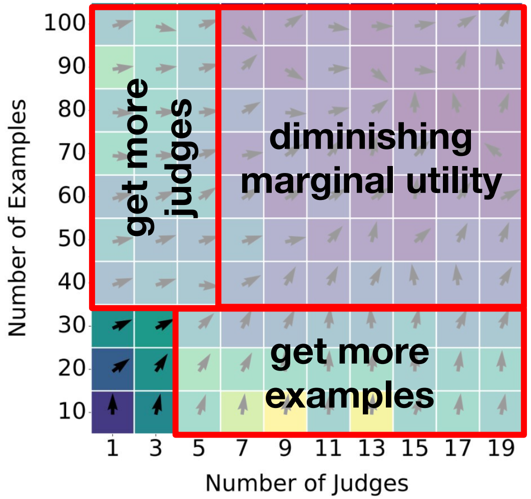
- human agreement
- statistical significance
- **cost and efficiency**
- robustness to cheating

Monte carlo simulations of hypothetical councils



Number of Examples	1	3	5	7	9	11	13	15	17	19
100	23	32	46	51	49	55	56	59	62	64
90	20	34	43	52	48	56	56	58	61	60
80	18	28	42	46	52	55	58	59	58	59
70	21	33	42	46	48	54	53	55	58	57
60	21	28	37	45	46	49	55	54	55	57
50	16	26	31	41	46	49	50	51	54	55
40	16	23	33	40	44	45	48	51	47	49
30	12	21	32	35	37	41	39	40	43	46
20	8	13	19	28	32	28	33	31	33	34
10	4	4	8	12	12	17	14	20	21	23

(c) Separability μ



(d) Separability μ gradients

Good leaderboards?

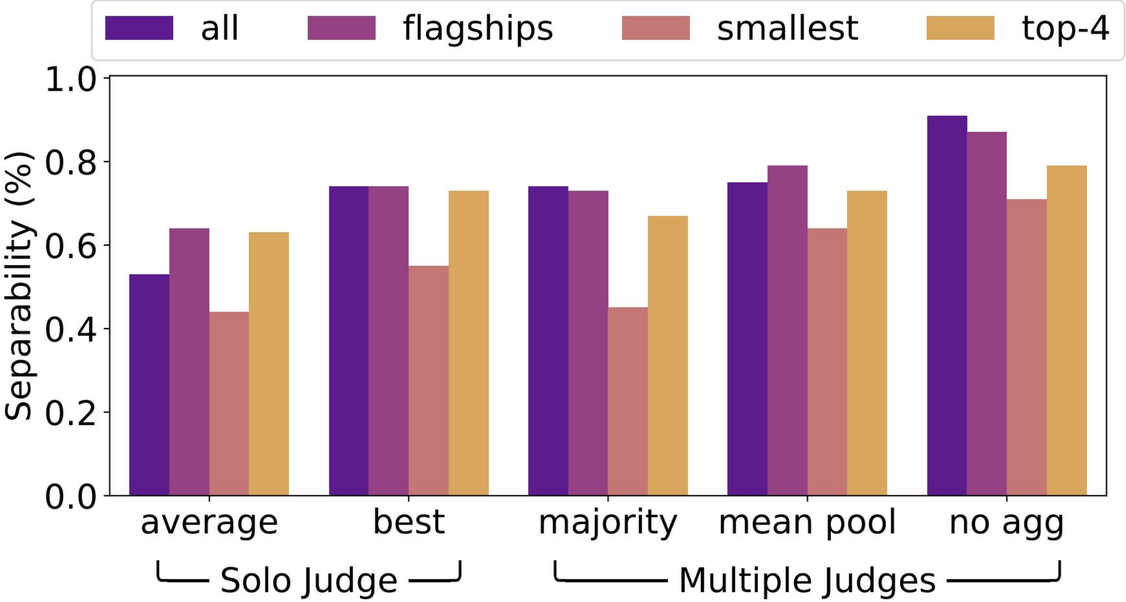
human agreement

statistical significance

cost and efficiency

robustness to cheating

From oligarchical councils to representative democracies



Good leaderboards?

human agreement

statistical significance

cost and efficiency

robustness to cheating

What is the effect of ***adversarial judges****?*

Good leaderboards?

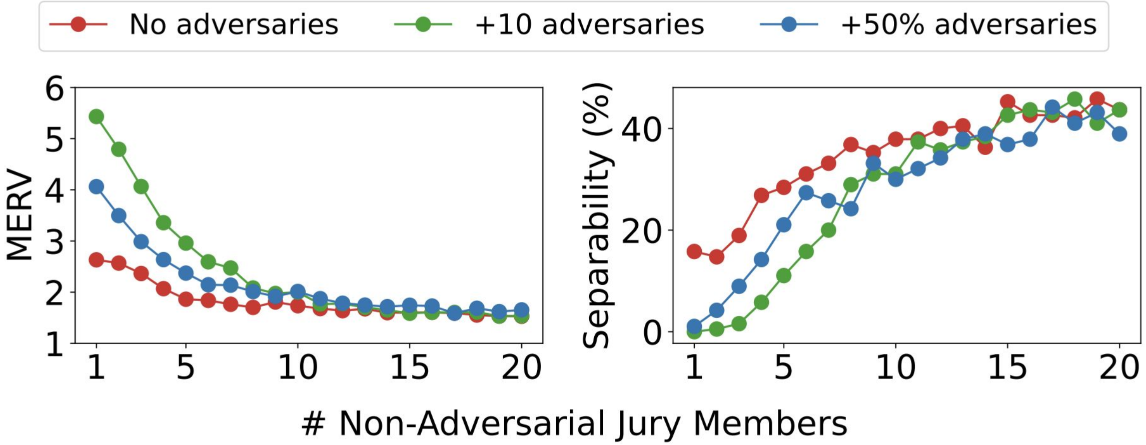
human agreement

statistical significance

cost and efficiency

robustness to cheating

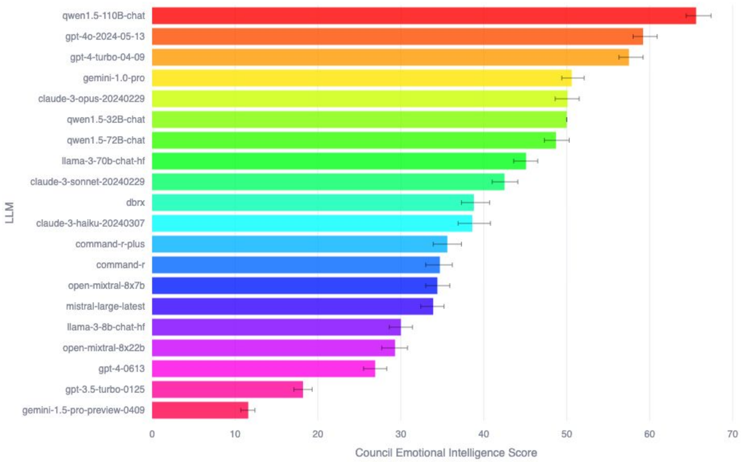
Larger councils are more resilient to **adversaries**.



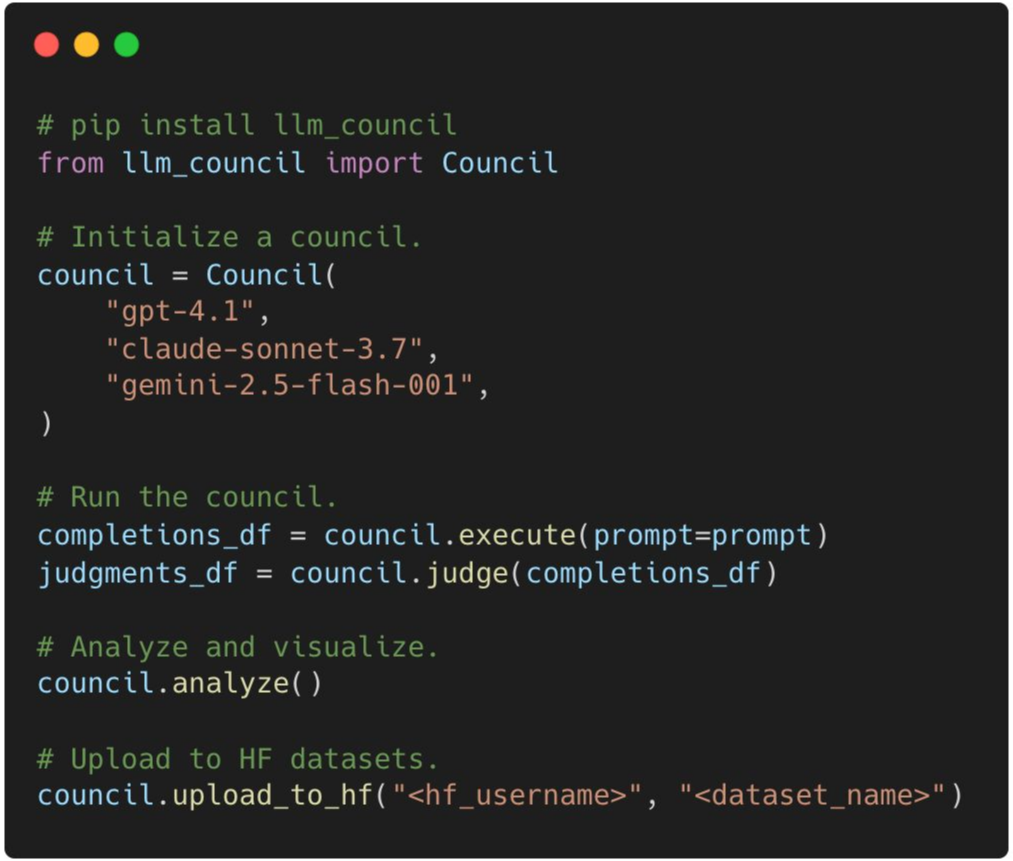
Using a bigger ensemble helps **automatically regularize** corrupt judges.

Can LLMs decide amongst themselves

who is the best*?*



Yes, they can! *



```python
# pip install llm_council
from llm_council import Council

# Initialize a council.
council = Council(
    "gpt-4.1",
    "claude-sonnet-3.7",
    "gemini-2.5-flash-001",
)

# Run the council.
completions_df = council.execute(prompt=prompt)
judgments_df = council.judge(completions_df)

# Analyze and visualize.
council.analyze()

# Upload to HF datasets.
council.upload_to_hf("<hf_username>", "<dataset_name>")
```

llm-council.com

X justinxzhao

in justin-zhao

Thank you!