Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)

Liwei Jiang♠ Yuanjun Chai♠ Margaret Li♠ Mickel Liu♠ Raymond Fok♠ Nouha Dziri⋆ Yulia Tsvetkov♠ Maarten Sap♢ Alon Albalak♣⁎ Yejin Choi♡

♠University of Washington ♢ Carnegie Mellon University ⋆Allen Institute for Artificial Intelligence ♣Lila Sciences ♡Stanford University

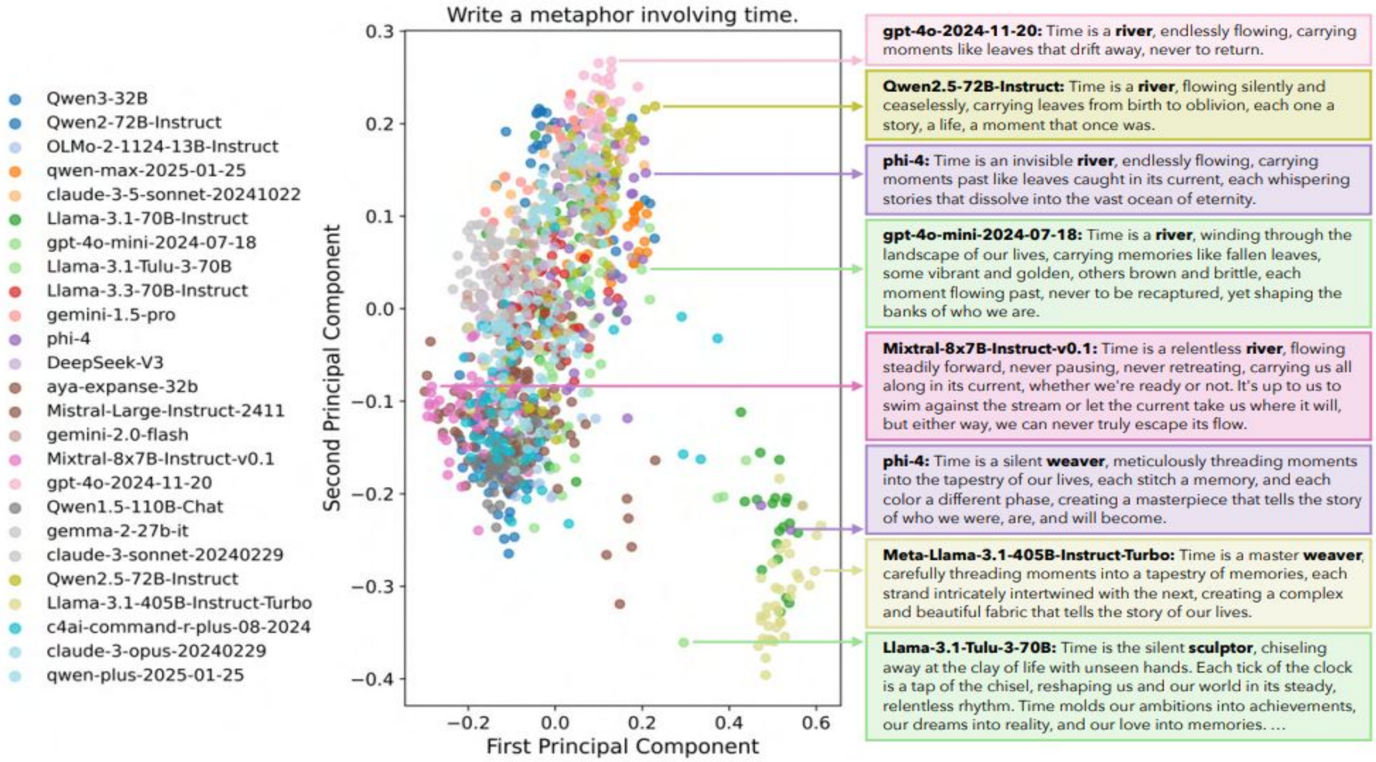
2025.10.27
NeurIPS 2025 Best Paper Awards
Paper: https://arxiv.org/pdf/2510.22954
Github: https://github.com/liweijiang/artificial-hivemind
채진영

Contributions

- Artificial Hivemind: homogeneous swarm intelligence of LMs.
- Introduce INFINITY-CHAT, **a large-scale dataset of 26K real-world open-ended queries** spanning diverse, naturally occurring prompts mined from WildChat - 6 top-level categories and 17 subcategories.
- Uncover a pronounced Artificial Hivemind effect: (1) **intra-model repetition**, where a single model repeatedly generates similar outputs, and, more critically, (2) **inter-model homogeneity**, where different models independently converge on similar ideas with minor variations in phrasing.
- Findings show that **state-of-the-art LMs, reward models, and LM judges are less well calibrated to human ratings** on model generations that elicit differing idiosyncratic annotator preferences

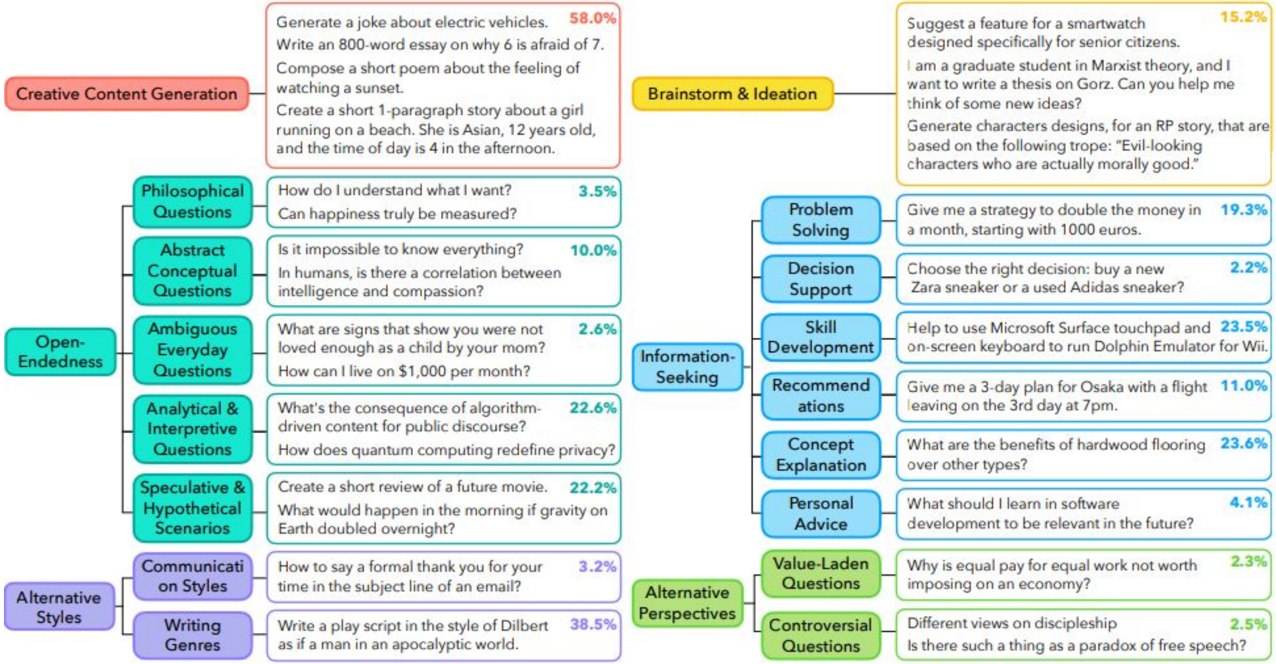
Background - What is the Artificial Hive Mind in LMs?

- LMs struggle to generate diverse, human-like creative contents.
- Existing benchmarks often target stylized tasks such as persona generation, keyword-driven storytelling, or random number generation, and often rely on narrowly defined tests centered on poetry or figurative language.



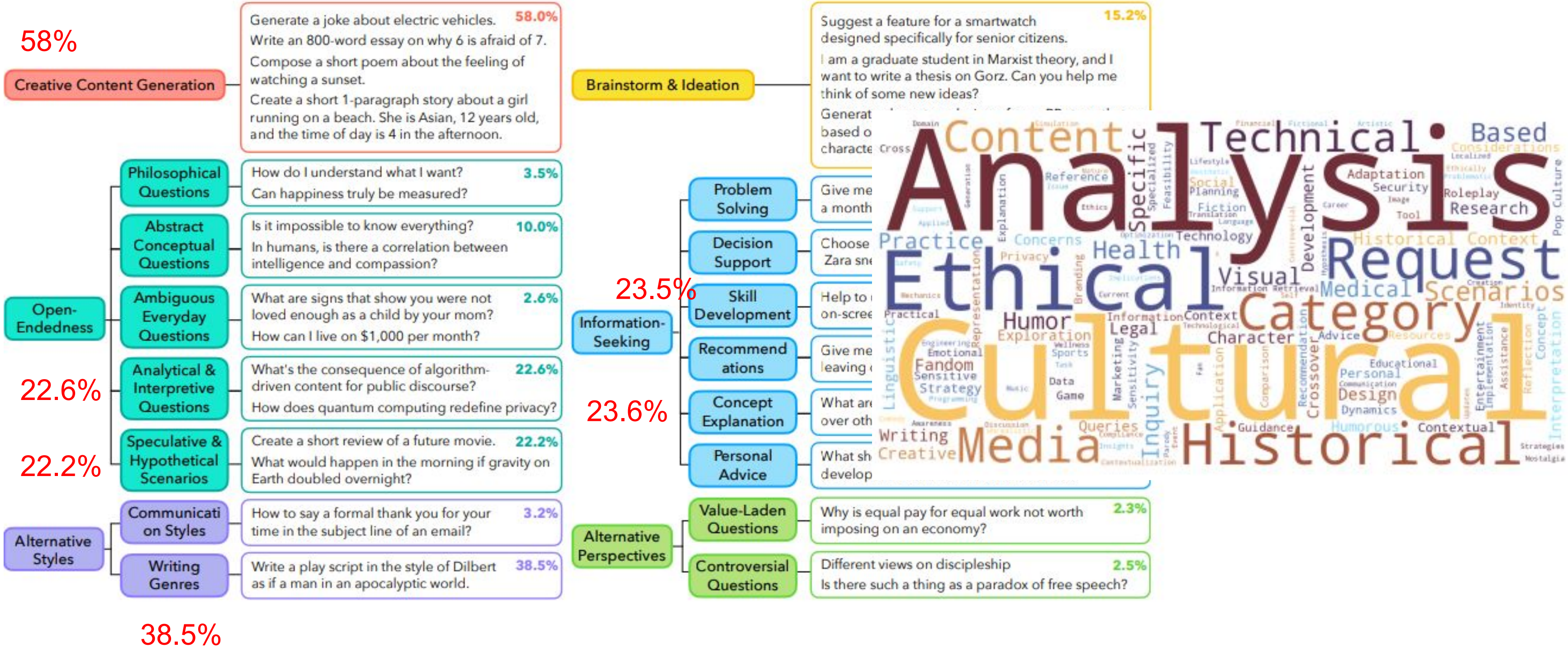
Infinity-Chat

- Existing LM alignment datasets prioritize response correctness over diversity -> **overlook inherent variability**
- Real-world open-ended queries with diverse responses.
- 26k English, non-toxic, 15-200 characters, diverse, real-world, open-ended user queries with no single truth
- Filter and refine user inputs from WildChat



Infinity-Chat

- Categorize diverse landscape of open-ended queries with 6 high-level categories and 17 fine-grained sub-categories.



Artificial Hivemind: Intra- and Inter-Model Homogeneity

- **Intra-model repetition: the same LM fails to generate diverse outputs**.
- 50 responses per query across 100 open-ended queries from Infinity-Chat100.
- Compute the average pairwise embeddings similarity within each response pool.

Similarity Score Ranges	Average	gpt-4o-2024-11-20	gpt-4o-mini-2024-07-18	claude-3-5-sonnet-20241022	claude-3-sonnet-20240229	claude-3-opus-20240229	Llama-3.3-70B-Instruct	Llama-3.1-70B-Instruct	Llama-3.1-405B-Instruct-Turbo	gemma-2-27b-it	gemini-2.0-flash	gemini-1.5-pro	qwen-max-2025-01-25	qwen-plus-2025-01-25	Qwen3-32B	Qwen2.5-72B-Instruct	Qwen2-72B-Instruct	Qwen1.5-110B-Chat	Mistral-Large-Instruct-2411	Mixtral-8x7B-Instruct-v0.1	OLMo-2-1124-13B-Instruct	Llama-3.1-Tulu-3-70B	c4ai-command-r-plus-08-2024	aya-expanse-32b	DeepSeek-V3	phi-4
0.9-1.0	43.8	51.0	53.0	61.0	48.0	59.0	51.0	23.0	43.0	33.0	40.0	53.0	55.0	56.0	40.0	48.0	35.0	48.0	43.0	45.0	29.0	28.0	24.0	50.0	42.0	38.0
0.8-0.9	35.2	36.0	34.0	22.0	36.0	27.0	30.0	44.0	38.0	43.0	41.0	32.0	37.0	28.0	36.0	26.0	35.0	36.0	34.0	41.0	39.0	39.0	32.0	36.0	39.0	40.0
0.7-0.8	12.6	10.0	9.0	9.0	10.0	7.0	12.0	23.0	11.0	14.0	11.0	11.0	4.0	11.0	14.0	19.0	24.0	10.0	15.0	9.0	12.0	18.0	19.0	8.0	13.0	11.0
0.6-0.7	4.9	1.0	4.0	5.0	5.0	3.0	4.0	5.0	3.0	7.0	4.0	3.0	3.0	3.0	6.0	4.0	5.0	2.0	5.0	4.0	9.0	11.0	10.0	5.0	3.0	9.0
0.5-0.6	1.9	2.0	0.0	2.0	1.0	4.0	1.0	2.0	4.0	0.0	2.0	1.0	1.0	2.0	1.0	1.0	0.0	3.0	2.0	1.0	7.0	1.0	7.0	1.0	0.0	2.0
0.4-0.5	0.7	0.0	0.0	1.0	0.0	0.0	1.0	2.0	0.0	2.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	2.0	2.0	3.0	0.0	3.0	0.0
0.3-0.4	0.6	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	2.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0
0.2-0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0
0.1-0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
0.0-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

→ Responses remain repetitive despite using top-p = 0.9, t=1.0 (high-stochasticity decoding),
In 79% of cases, the average similarity exceeds 0.8
→ **LMs still fails to generate diverse responses to open-ended queries**.

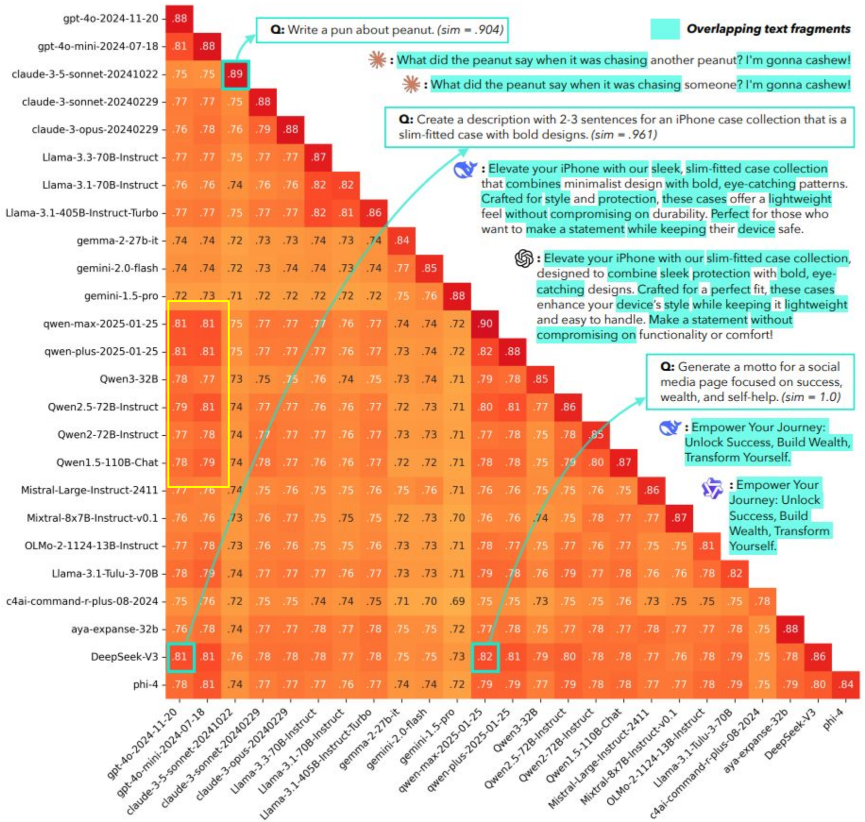
Artificial Hivemind: Intra- and Inter-Model Homogeneity

- **Intra-model repetition: the same LM fails to generate diverse outputs.**
- Evaluate min-p decoding with the same setup and compute pairwise sentence embedding similarities.

→ min-p reduces extreme repetition, 81% of response pairs still exceed 0.7 similarity and 61.2% exceed 0.8, revealing **mode collapse even under diversity-oriented decoding**.

Artificial Hivemind: Intra- and Inter-Model Homogeneity

- **Inter-model homogeneity: different models produce similar outputs**.
- 25 unique models, each generating 50 outputs



→ The average pairwise similarity between responses from different models ranges from **71% to 82%**, with some pairs notably higher

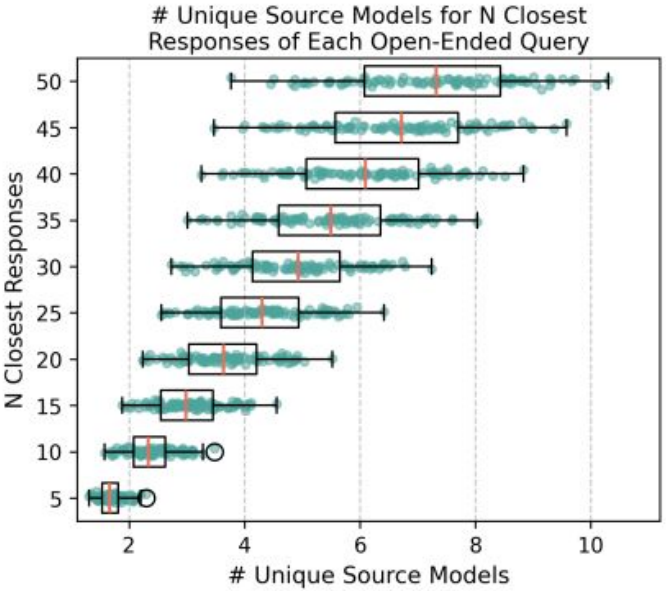
→ **OpenAI's GPT models and Qwen's API models** tend to have **higher similarities** even with models outside their own families

→ DeepSeek-V3 and gpt-4o-2024-11-20 **generate overlapping phrases** like "Elevate your iPhone with our," "sleek, without compromising," and "with bold, eye-catching" in answer to the some query.

→ No exact causes, but possibly shared data pipelines across regions or contamination from synthetic data.

Artificial Hivemind: Intra- and Inter-Model Homogeneity

- **Inter-model homogeneity: different models produce similar outputs.**
- Examine the extent to which outputs from different models become indistinguishable from one another.
- 25 models, 50 responses

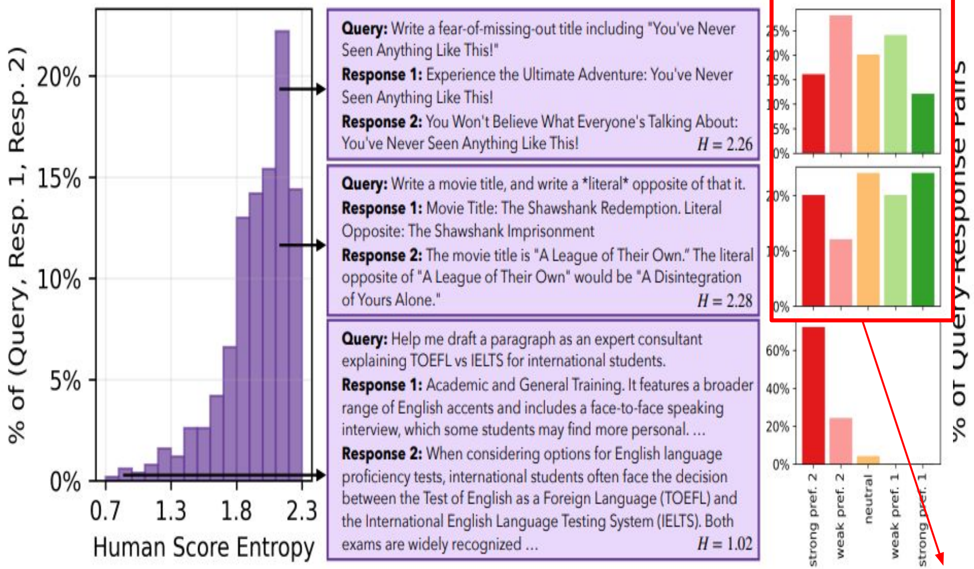


→ The most similar responses often originate from multiple models. For instance, with N = 50, perfectly disjoint responses would yield all 50 from a single model. Yet, we find an average of ~8 unique models per top-50 cluster, with some queries exceeding 10, indicating distinct models frequently generate highly similar content, sometimes resulting in higher inter- than intra-model similarity..

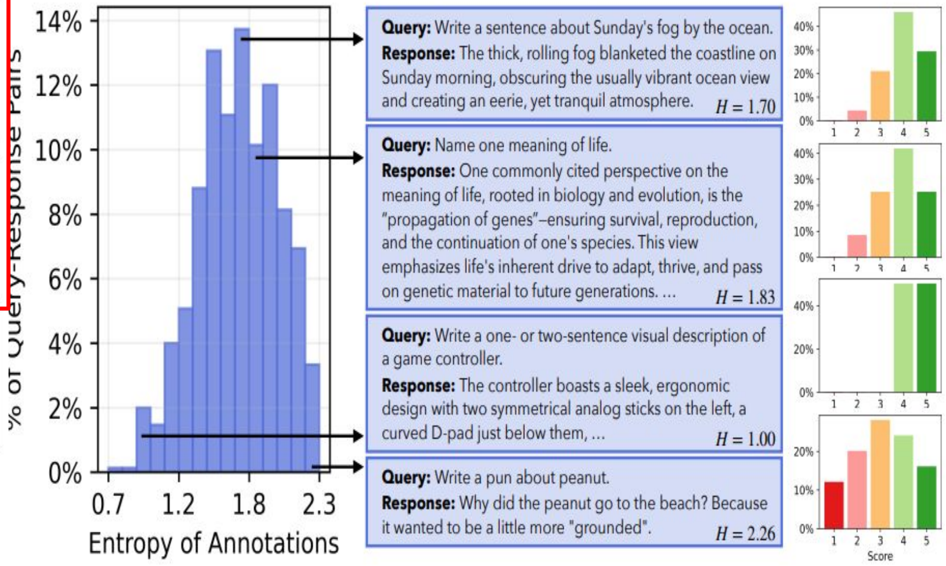
How models handle alternative responses to the Queries?

- Examine whether the ratings of LMs, reward models, and LM judges are calibrated to match human scores given different responses to the queries.
- Gathering distributional annotations across many humans.
- 1) absolute ratings (1–5 scale for response quality) 2) pairwise preference ratings (strong/weak preference between two responses to the same query)

pairwise ratings



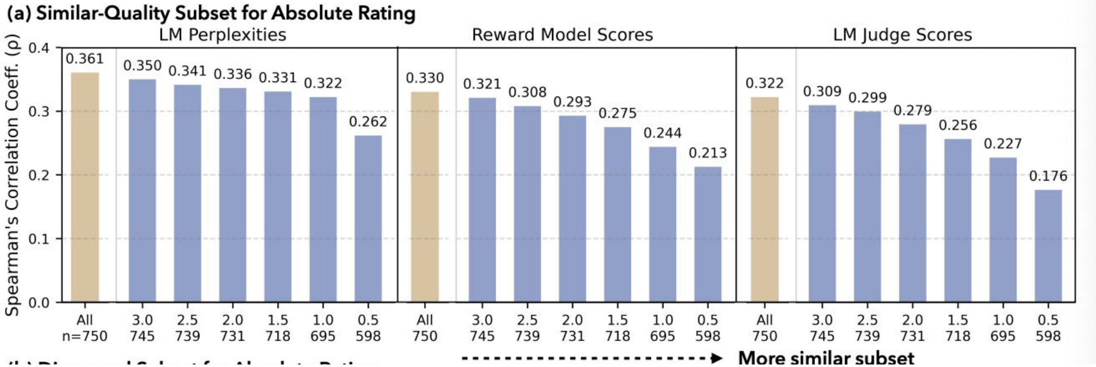
absolute ratings

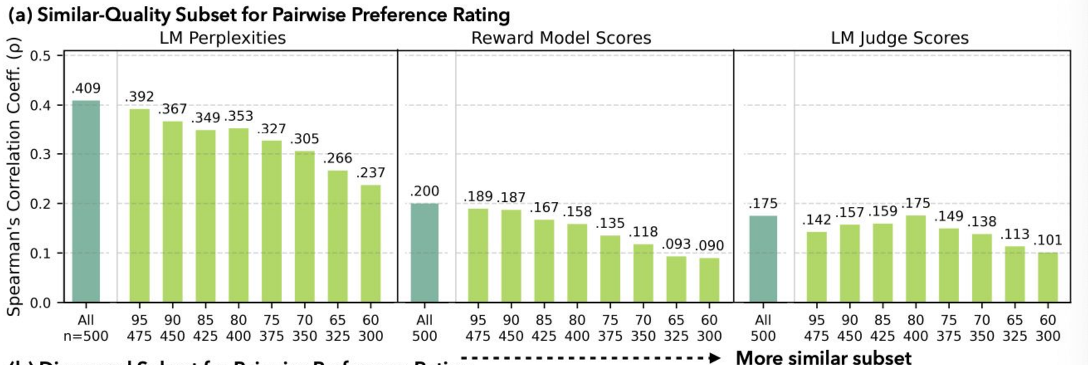


Some response pairs show near-uniform support across all options, indicating annotator disagreement

How models handle alternative responses to the Queries?

- Gathering LMs, Reward Models, and LM Judges Ratings
- Comparing model ratings to human scores for responses to open-ended queries **1) similar-quality alternative responses - same queries** 2) responses with high annotator disagreement
    - **Models show weaker alignment with human-ratings for alternative responses of similar quality.**



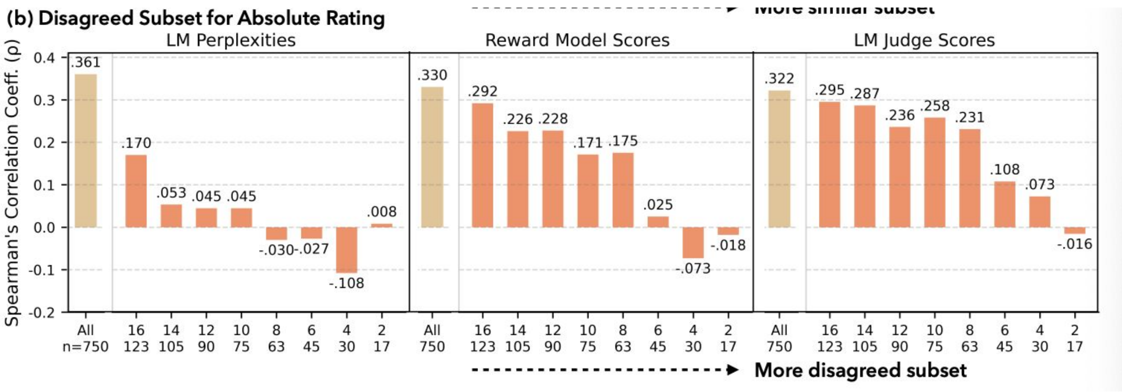


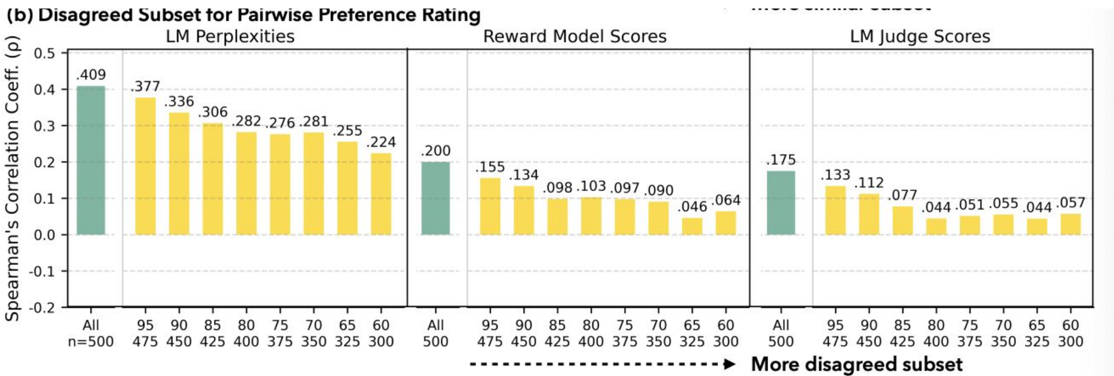
outlier: Q1 − k · IQR or Q3 + k · IQR

→ Correlations between human ratings and those of LMs, reward models, and LM judges drop significantly on similar-quality subsets, for both absolute and pairwise preference rating setups

How models handle alternative responses to the Queries?

- Comparing model ratings to human scores for responses to open-ended queries 1) similar-quality alternative responses **2) responses with high annotator disagreement**
  - **Model judgments are less aligned where annotators disagree.**





→ Show that correlations with human ratings across models drop substantially for examples with high annotator disagreement, in both absolute and pairwise rating setups.

P_disagree =
1 − [max(C_prefer 1,C_prefer 2)+0.5·C_tie] / C_total

Conclusion

- Infinity-Chat, a large-scale dataset designed to evaluate LMs’ diversity
  → new foundation for mitigating mode collapse in gen AI

- Artificial Hivemind - 1) intra-model repetition 2) inter-model homogeneity