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

Jiang et al. | NeurIPS 2025 | CS 8803-LLM

A Thought Experiment

- "Write a metaphor about time."
- 25 LLMs, 50 responses each = 1,250 metaphors

- **What would you expect?**
  - Diverse, creative metaphors... or something else?

The Result: Mode Collapse

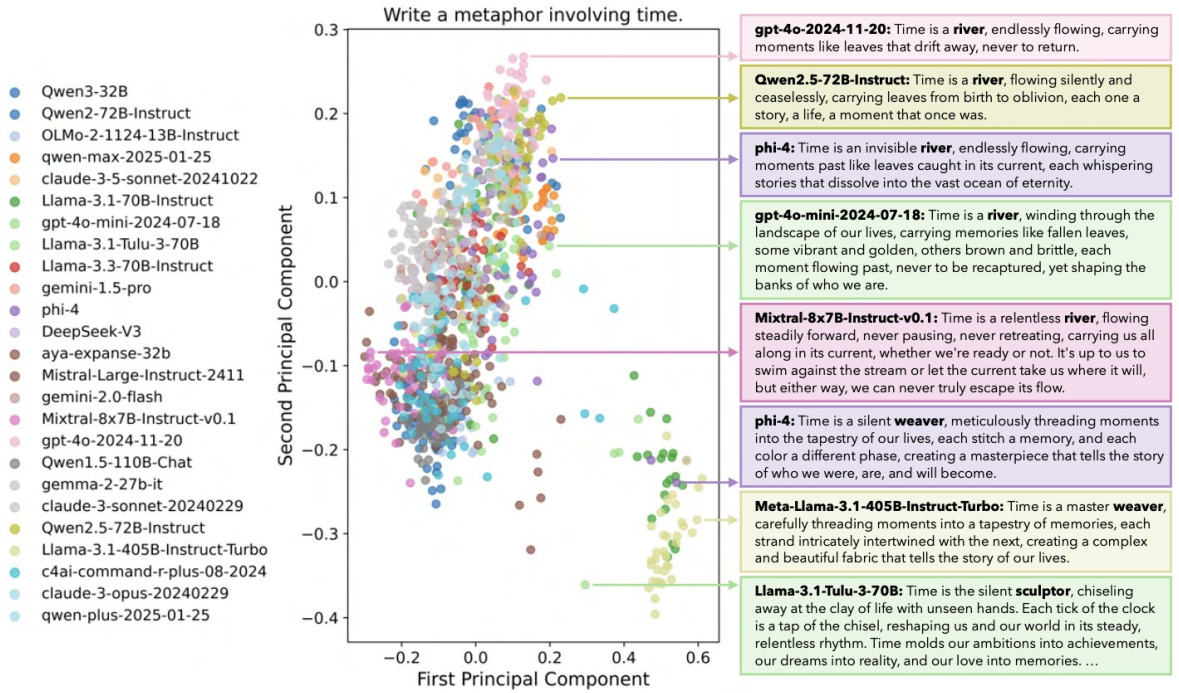


Figure 1: Responses to the query **"Write a metaphor about time"** clustered by applying PCA to reduce sentence embeddings to two dimensions. Each of the 25 models generates 50 responses using top-p sampling ($p = 0.9$) and temperature $= 1.0$. Despite the diversity of model families and sizes, the responses form just two primary clusters: a dominant cluster on the left centered on the metaphor "time is a river," and a smaller cluster on the right revolving around variations of "time is a weaver."

- 1,250 responses cluster into TWO metaphors: "time is a river" and "time is a weaver"

The "Artificial Hivemind" Effect

Intra-Model Repetition

- Same model repeats similar outputs
- Even with high temperature + top-p
- 79% of queries: similarity > 0.8

Inter-Model Homogeneity

- Different models converge on same ideas
- Cross-model similarity: 71%-82%
- Sometimes verbatim phrase overlap

Why Should We Care?

- Hundreds of millions use LLMs for creative tasks

- **If every model gives the same answers:**
  - Creative homogenization of human thought
  - Loss of diverse perspectives in brainstorming
  - Cultural flattening across languages/regions

- **First large-scale systematic study of this phenomenon**

How Do We Study This Systematically?

What's Missing in Existing Benchmarks?

Existing Benchmarks

- Researcher-designed tasks
- Single task type
- One model at a time
- ~3 annotators per example

What's Needed

- Real user queries
- Diverse open-ended categories
- Cross-model comparison
- Dense annotations (25+)

No benchmark combines real user queries, cross-model evaluation, and dense human annotation at scale

Prior Work: Three Research Threads

- **Diversity Collapse in LMs**
  - RLHF + synthetic data training reduces output diversity
- **Creativity & Divergent Thinking Evaluation**
  - Individual outputs score high; collective diversity is low
- **Pluralistic Alignment**
  - Serve diverse preferences, not monolithic values

INFINITY-CHAT: 26K Real-World Queries

- **Data Source**
- WildChat: real user-chatbot conversations
- 37,426 GPT-4 queries filtered to:
- 26,070 open-ended queries
- 8,817 closed-ended queries
- English, non-toxic, 15-200 chars

- **Human Annotations**
- 31,250 total labels
- 25 annotators per example
- Absolute ratings (1-5 scale)
- Pairwise preference judgments
- 8x denser than HelpSteer3

Taxonomy of Open-Ended Queries

- 6 top-level categories, 17 subcategories. Creative Content Generation dominates at 58%

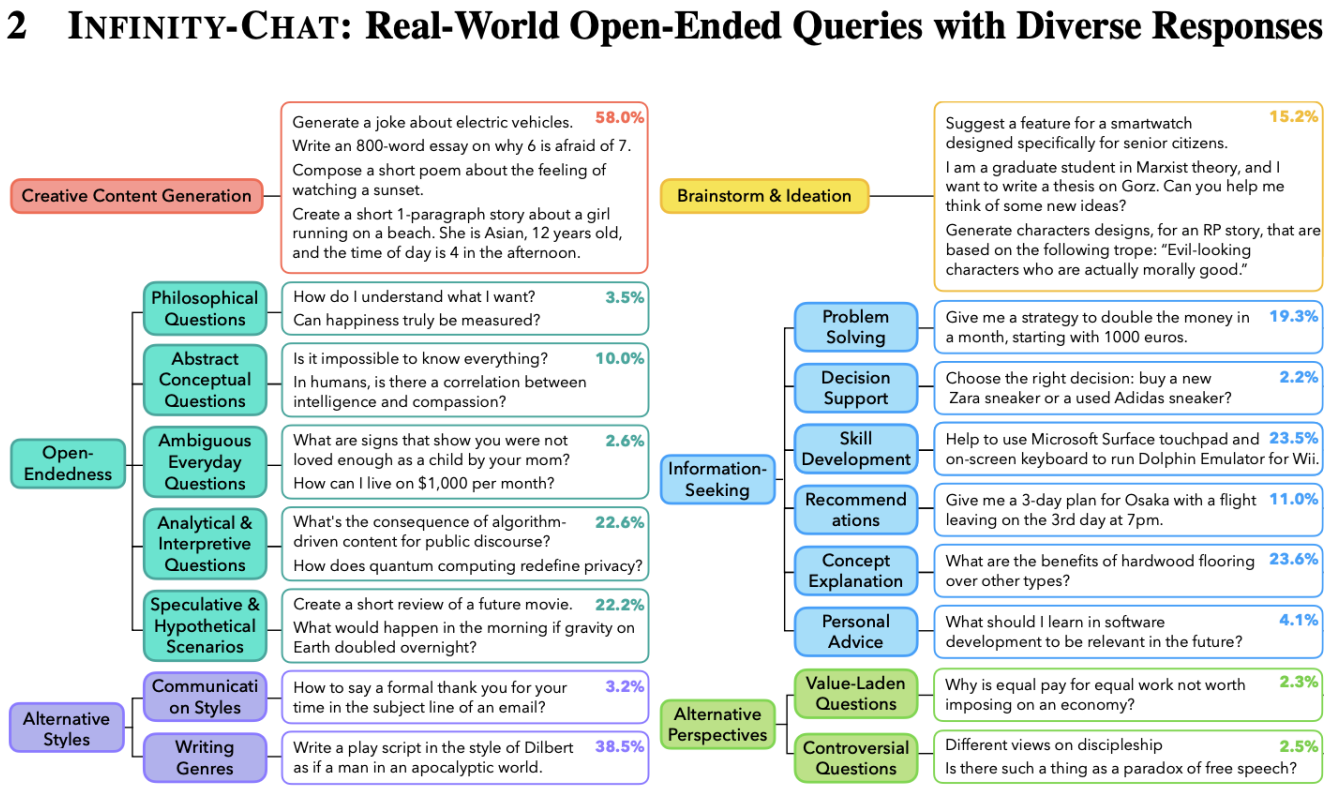


Figure 2: A taxonomy of **real-world open-ended queries that invite diverse model responses** that are mined from in-the-wild user-chatbot interactions, categorized into 6 top-level and 17 fine-grained subcategories, along with their occurrence percentages.

How Bad Is the Hivemind, Really?

Georgia Tech

Methodology: Measuring Mode Collapse

- **Setup:**
  - 100 queries, 70+ LMs (25 in main paper), 50 responses each
  - Sampling: top-p = 0.9, temperature = 1.0
  - In total: 100*25*50 = 125,000
- **Measurement:**
  - Sentence embeddings (text-embedding-3-small)
  - Pairwise cosine similarity;
  - Baseline: 0.1-0.2 (The similarity of responses from two random pick questions)

Intra-Model Repetition

- 79% of queries: within-model similarity > 0.8 (baseline: 0.1-0.2)

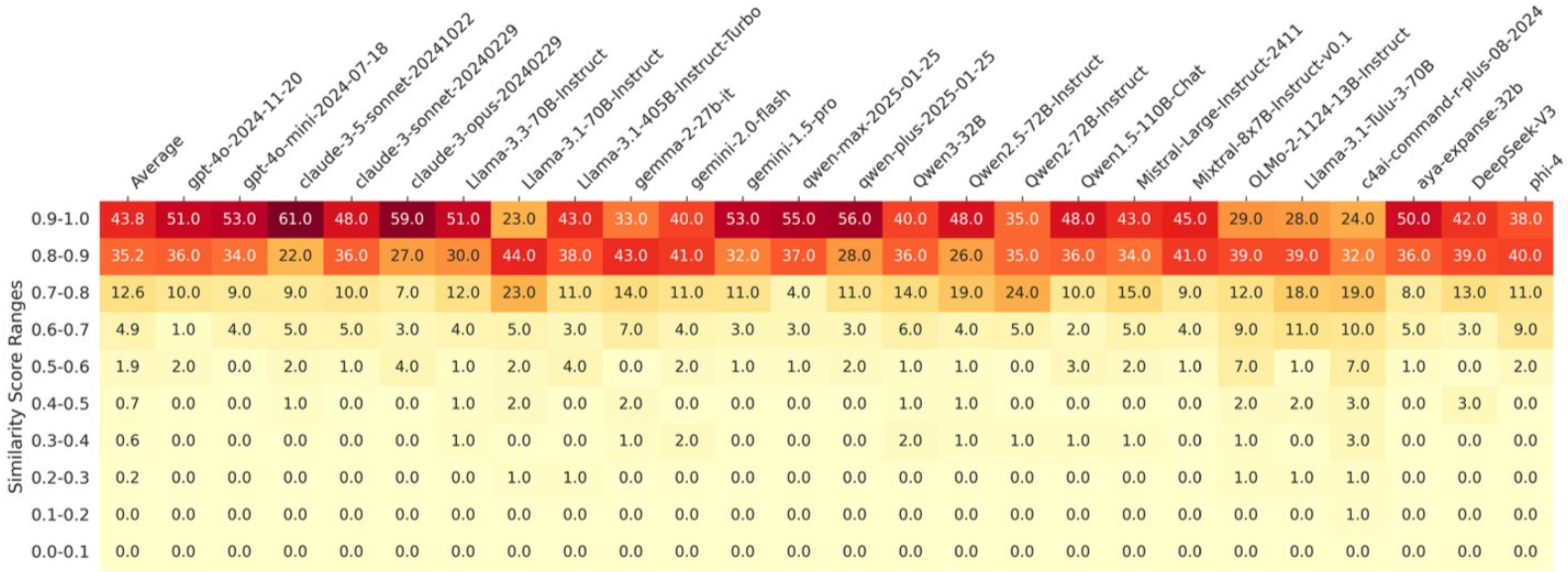


Figure 4: The heatmap shows **degree of repetition in responses to open-ended queries generated by the same LMs**. For each model, we generate 50 responses per query across 100 open-ended queries from INFINITY-CHAT100. We then compute the average pairwise sentence embedding similarities for each query's response pool and measure the percentage of queries falling into the similarity ranges indicated on the y-axis. Under the sampling parameters (top-$p = 0.9$, temperature $= 1.0$), the average pairwise similarity among responses to the same prompt typically exceeds 0.8. As a baseline, randomly paired responses from the global pool 100% fall within the 0.1–0.2 range.

Can Decoding Strategies Fix This?

Standard Sampling

- top-p=0.9, temp=1.0
- 79% queries: similarity > 0.8
- In higher temp (2.0), response would become incoherent

Min-p Decoding
(An advanced decoding strategy)

- top-p=1.0, min-p=0.1, temp=2.0
- 81% pairs still > 0.7
- 61% still exceed 0.8
- **Decoding alone is insufficient**

Inter-Model Homogeneity

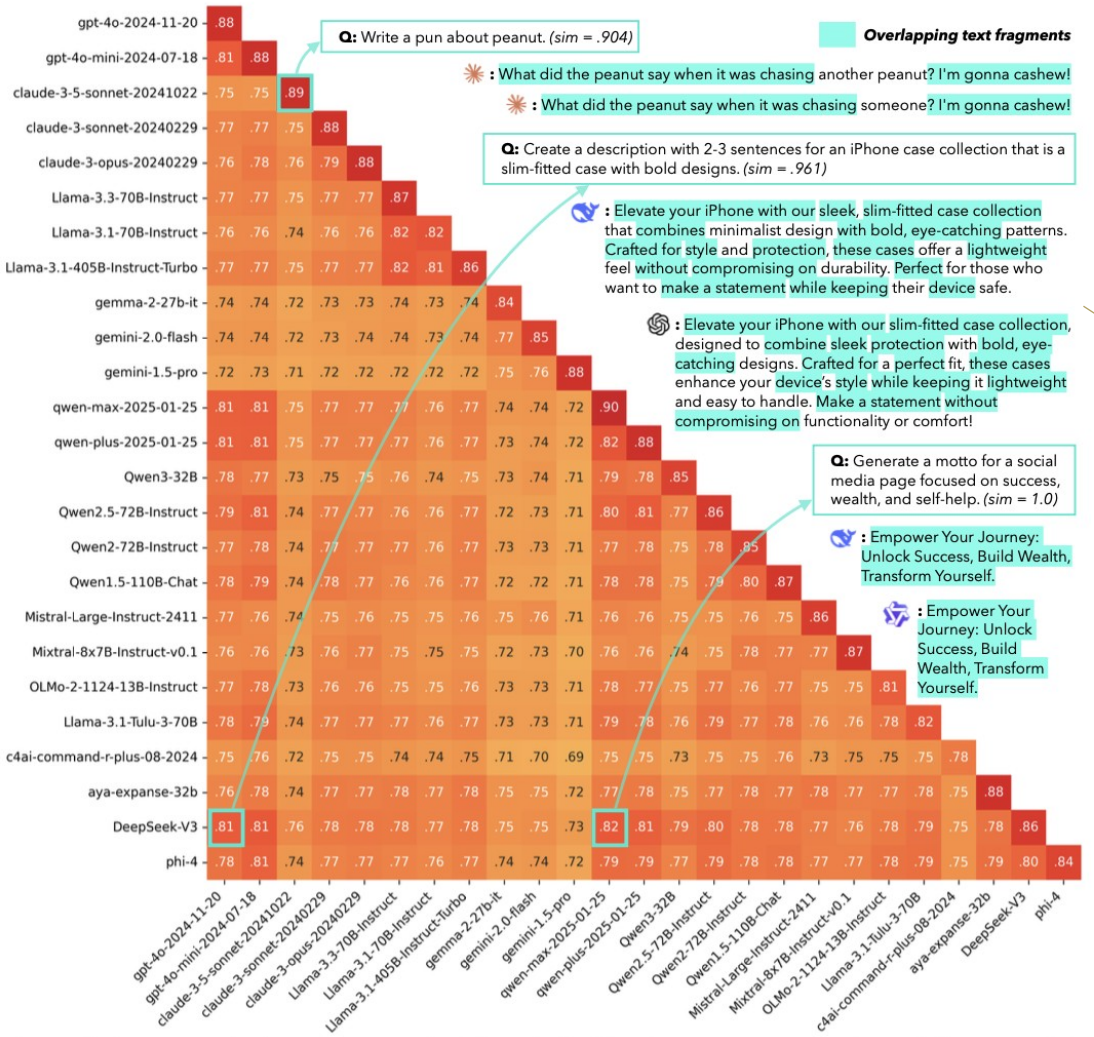
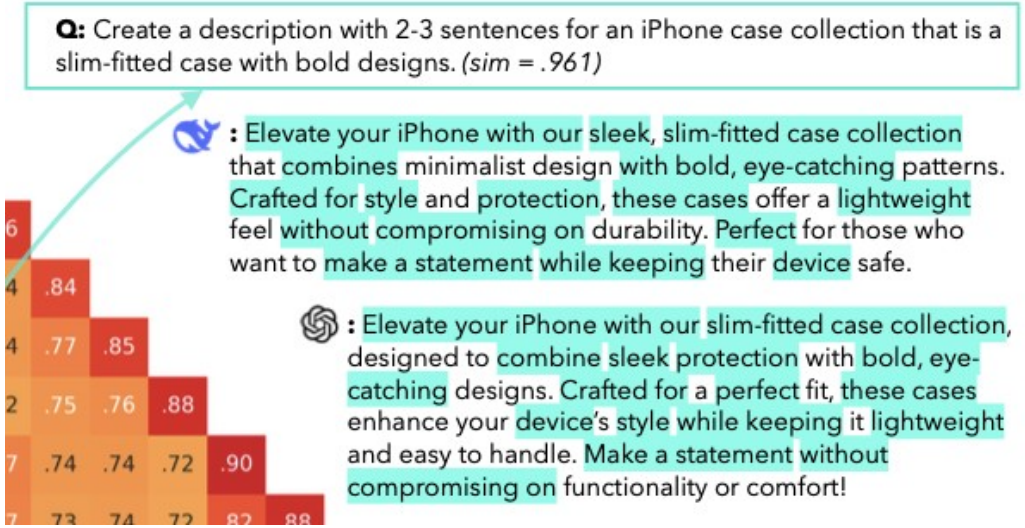


Figure 6: Average pairwise sentence embedding similarities between responses from different models reveal substantial semantic overlap across model outputs. Qualitative examples further illustrate that **different models often produce strikingly similar responses to fully open-ended queries**, including extended verbatim spans, underscoring the extent of repetition across models in open-ended generation tasks. All responses are generated using top-$p = 0.9$ and temperature $= 1.0$.

- Cross-model similarity: 71%-82%. DeepSeek-V3 & qwen-max: 0.82, DeepSeek-V3 & GPT-4o: 0.81



Verbatim Overlaps Across Models

- Q: "Write a pun about peanut"

- GPT-4o: "I'm gonna cashew!"
- Claude: "I'm gonna cashew!"

- Same joke, different companies

- Q: "iPhone case description"

- DeepSeek: "Elevate your iPhone...
- bold, eye-catching..."
- GPT-4o: "Elevate your iPhone...
- bold, eye-catching..."

How Indistinguishable Are the Models?

- Top-50 similar responses: ~8 different source models
- Models more similar to each other than to themselves
- **The Hivemind is real: convergence within and across families**

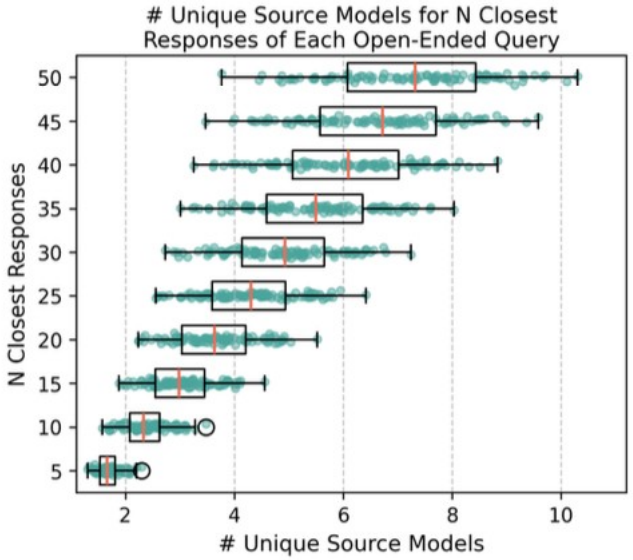


Figure 8: The avg.# of unique source models among the top-N most similar responses to each open-ended query across 25 LMs.

If diverse → X should be 1
(all top responses from same model)

Reality: X ≈ 8 at N=50
(responses from 8 different models are mixed together)

→ Models are nearly indistinguishable

Can Models Evaluate Open-Ended Responses?

We often use multi-models to evaluate a response (LAAJ), is this a good choice?

Georgia Tech

Humans Genuinely Disagree on Quality

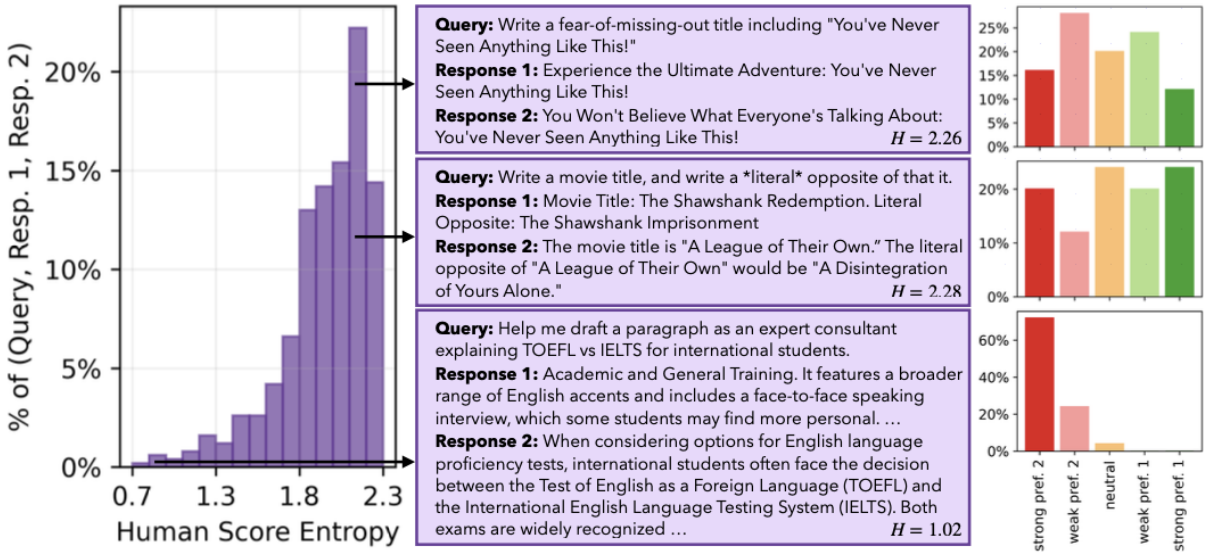


Figure 7: The **left** histogram shows the distribution of Shannon entropy across the 25 human annotations for each **(Query, Response 1, Response 2)** triplet, where annotators judge which response is better. Given open-ended queries, multiple high-quality responses are possible, often leading to disagreement among annotators and, on average, high entropy. With 25 annotations, label distributions can vary widely across triplets. The **middle** panel presents example triplets from different entropy regions, and the **right** bar plots show their corresponding label distributions.

- 18,750 absolute + 12,500 pairwise labels (25 annotators each)
- High Shannon entropy: genuine disagreement, not noise
- Pluralistic preference: different people value different things

Open-ended tasks have no single 'correct' preference — disagreement is the norm

Models Fail Where It Matters Most

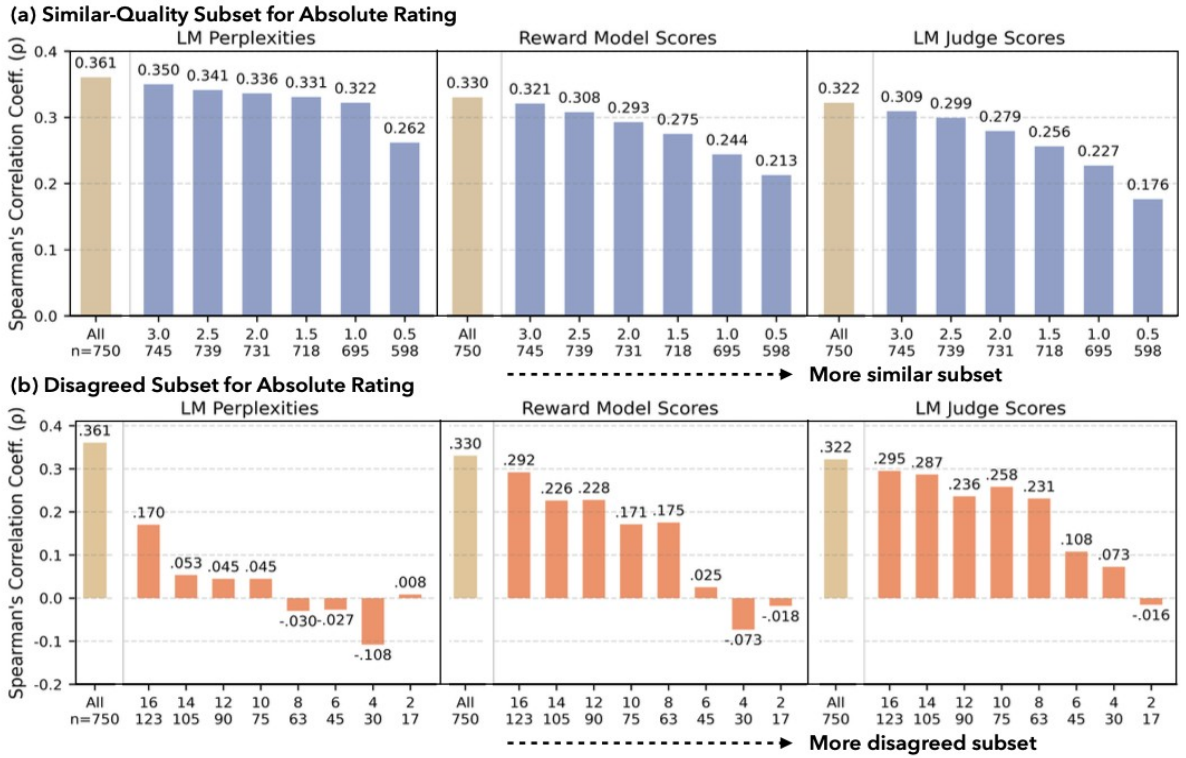


Figure 10: We compute Spearman's correlation coefficients between human-annotated and model-generated **absolute rating** scores, including *LM* perplexity scores, *reward model* scalar outputs, and *LM judge* scalar ratings. Correlations are calculated for the **full set**, as well as two groups of subsets: (a) responses with **similar human-rated quality**, and (b) responses with **high human disagreement**. The results show that correlations are notably lower in these two subsets, indicating weaker alignment between model scores and human judgments in cases of subtle or contested quality differences.

- 3 columns = 3 evaluation methods
  - LM Perplexity | Reward Models | LM Judges
- Row (a): similar-quality subset
  - → Bars drop as quality gap narrows
- Row (b): high-disagreement subset
  - → Correlation drops to 0 or negative
- 56 LMs, 6 reward models, 4 LM judges

All automated evaluation methods fail on the subjective cases

Our Interpretation: A Possible Feedback Loop

- Paper leaves root causes as open questions.
- Our synthesis:
  - Reward models trained on majority preference
  - Models optimized toward single "best" response
  - Generation diversity decreases
  - Future (synthetic) training data more homogeneous
  - **Cycle repeats may require new training approaches**

Limitations & Open Questions

- Cosine similarity as sole metric
- English-only queries
- 100-query main experiment subset
- Embeddings miss narrative diversity

- Root cause unknown:
- Pretraining data overlap?
- RLHF convergence?
- Synthetic data contamination?

Broader Societal Implications

- **For individuals:**
  - AI brainstorming may yield narrower ideas than expected
- **For society:**
  - Risk of cultural homogenization at global scale
- **For AI research:**
  - Diversity needed as first-class metric alongside helpfulness

Key Takeaways

- **1. The Artificial Hivemind is real and measurable**
  - Models converge within and across families
- **2. Decoding tricks alone cannot fix it**
  - Min-p, high temperature are insufficient
- **3. Evaluation poorly calibrated for open-ended tasks**
  - Reward models fail where preferences are pluralistic
- **4. We need new training-level approaches**
  - Pluralistic alignment and diversity-aware evaluation

Thank You

Paper: arXiv:2510.22954 | Code: github.com/liweijiang/artificial-hivemind

Georgia Tech