The Artificial Hivemind, Revisited: Open-Ended Homogeneity as an Elicitation Artifact

A short empirical note in homage to Jiang et al. (2025)

Rob Panton

Novamine AI · novamine.ai · contact@novamine.ai

ABSTRACT

Jiang et al. (2025) document the *Artificial Hivemind*: on open-ended queries, language models exhibit striking intra-model repetition and inter-model homogeneity, with average pairwise response similarity exceeding 0.8 on 79% of prompts. Using Infinity-Chat prompts and the paper's own measurement protocol (50 generations per prompt; text-embedding-3-small; mean pairwise cosine similarity), we test whether this homogeneity is a property of the models or of how they are asked. Across 100 stratified prompts and four arms, a raw GPT-4o control reproduces the paper's finding in full (mean similarity 0.92; 96% of prompts above the 0.8 hivemind line). A fixed, prompt-agnostic inference-time orchestration system applied to the very same GPT-4o moves that mass to **1%** (mean 0.54), reducing similarity on 99 of 100 prompts. The same system, unmodified, does the same to Claude Opus 5 (57% → 3% above the line) — where it also breaks 23 of the 24 prompts on which raw Opus 5 collapses to fifty near-identical generations (similarity ≈ 1.00). The hivemind, on this evidence, is real but not fundamental: it is an elicitation artifact, recoverable at inference time without touching weights. The same system holds the current records on both NoveltyBench metrics (8.79 distinct / 6.69 utility; submission in review). All outputs are public and re-scorable.

1 Motivation

The Artificial Hivemind paper poses the sharpest available version of the homogeneity question and supplies the instrument to measure it. Its concluding call — that the findings should guide mitigation of the risks the hivemind poses — is the invitation this note answers. We ask a narrow question with the paper's own tools: *does the homogeneity live in the model, or in the default way of asking?*

2 Setup

We sample 100 open-ended prompts from Infinity-Chat, stratified across the released category taxonomy (fixed seed; the same 100 prompts in every arm). For each prompt we collect 50 generations per condition and compute mean pairwise cosine similarity with text-embedding-3-small, exactly following the paper. Four conditions: (i) **GPT-4o raw** — a model from the paper, its decoding protocol (temperature 1.0, top-p 0.9), anchoring our replication; (ii) **GPT-4o orchestrated** — the identical model driven by a fixed prompt-agnostic inference-time system (Novamine) that first maps substantively distinct answer families and then writes one answer per family; (iii) **Claude Opus 5 raw** (temperature-only; the Anthropic API rejects joint temperature/top-p — a protocol deviation we note plainly, which is why the GPT-4o pair carries the like-for-like comparison); (iv) **Claude Opus 5 orchestrated** — the same system, unmodified, on the second provider's model. Nothing is tuned per prompt.

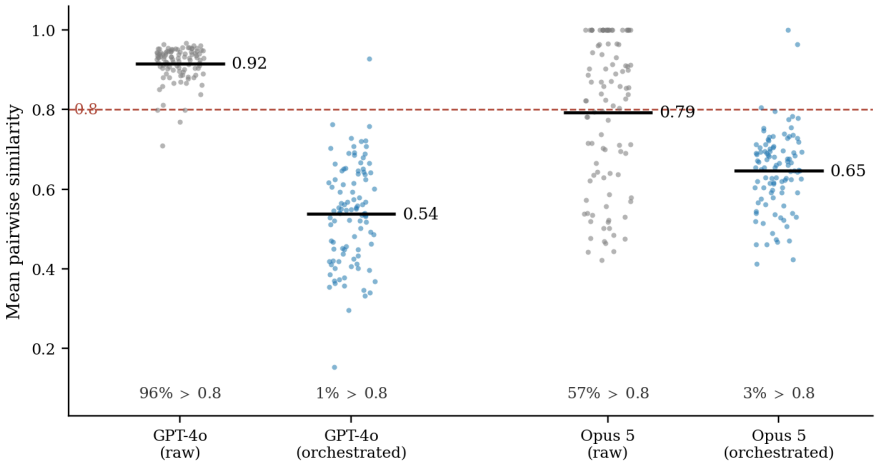

Figure 1: Mean pairwise similarity per prompt (50 generations each), n=100 per arm. Each raw/orchestrated pair is the same model under different elicitation. Raw GPT-4o reproduces the hivemind in full; orchestrated GPT-4o retains a single prompt above the 0.8 line. Horizontal bars mark condition means.

Condition	Mean sim.	% prompts > 0.8	% > 0.7
GPT-4o (raw, paper protocol)	0.92	96%	100%
GPT-4o (orchestrated)	**0.54**	**1%**	**9%**
Claude Opus 5 (raw)	0.79	57%	68%
Claude Opus 5 (orchestrated)	**0.65**	**3%**	**26%**

3 Reading

The raw arms confirm the paper: homogeneity is severe and survives a model generation. The orchestrated arms show the same weights contain far more distributional diversity than default elicitation surfaces — consistent with the paper's observation that diversity "must be deliberately elicited," and extending it: deliberate elicitation, engineered as a fixed system, does not merely dent the hivemind but removes nearly all the above-line mass, on both providers' models, without per-prompt tuning. The paired GPT-4o comparison is the cleanest statement: mean per-prompt similarity drop 0.38 (max 0.75), a reduction on 99 of 100 prompts, at unchanged answer quality (companion result: both NoveltyBench records — distinct 8.79, utility 6.69 — held simultaneously by the same system; the previous distinct record required fine-tuning and cost utility). Two incidental observations may interest the authors: raw Opus 5 collapsed to fifty near-identical generations (similarity ≈ 1.00) on 24 of 100 prompts — a stronger collapse mode than the paper reports for its cohort — and orchestration broke 23 of those 24, the residual prompt surviving family-exclusion intact.

4 What we are not claiming

This is a practitioner's note, not a peer of the original work. We do not claim novelty for the idea that prompting can elicit diversity — §4.3 of the original paper and the verbalized-sampling literature establish it. The contribution is magnitude and form: a fixed inference-time system erasing the above-line mass on the paper's own instrument, reproducibly, with public re-scorable outputs. The orchestration prompts are proprietary; outputs, scores, and scoring method are open.

5 Invitation

If useful to the group's mitigation agenda, we would be glad to share the full output sets for independent re-scoring, run agreed protocol variants, or contribute a systems track datapoint to any future iteration of Infinity-

Chat. NoveltyBench submission (public): github.com/novelty-bench/novelty-bench/pull/9.

References: L. Jiang et al., *Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)*, NeurIPS 2025 (Best Paper). Y. Zhang et al., *NoveltyBench: Evaluating Language Models for Humanlike Diversity*, COLM 2025. This note's visual form is a deliberate homage to the former.