

# Reach Into the CHOIR

## Free-List Elicitation Uncovers Distinct Model Voices in LLM Ensembles

Ben Wigler · Maria Tsfasman

LoveMind AI · New York, USA · benben@lovemind.ai · masha@lovemind.ai



AT A GLANCE · NINE MODELS · INFINITY-CHAT 100

93/100

prompts show the hivemind's surface agreement  
above-chance overlap across models

88%

of answer sets are traced to the right model from their  
concepts alone  
every model has its own voice

65/100

prompts yield over 50% new concepts when pushed past  
first instincts  
the familiar first answer is not the whole repertoire

### 1 THE HIVE

#### IN CONVERSATION WITH

#### Artificial Hivemind

Jiang et al. (NeurIPS 2025) showed that models answering open-ended prompts keep landing on the **same few answers**, within and across models. CHOIR uses their 100 representative prompts and asks the next question: **when the hive hums one note, why?**

#### The worry: one voice for everyone

Asked for a metaphor about time, Hivemind's 25 models mostly said *"time is a river."* If every assistant converges, people get the same answer wherever they turn: the "homogenization of human thought" Hivemind warns about.

#### THREE WAYS MODELS CAN AGREE



#### The question only has a few answers

*"Name an optimization technique used in machine learning."*

**CHOIR'S TEST** · score how much room each question leaves



#### Models repeat the question's own words

*"Write a metaphor involving time." 99.6% of listed items restate time or metaphor.*

**CHOIR'S TEST** · count answers that just repeat the question's words, then re-ask without them



#### One loud default hides the rest

*"Name one meaning of life."*

**CHOIR'S TEST** · ask for long ranked lists, many times, and push past first instincts

### 2 HOW WE LISTENED

#### Free-listing, borrowed from *anthropology*.

Anthropologists map a culture's ideas by asking many people to list everything they can think of (Weller & Romney, 1988). Items named by **more** people, and **earlier**, form the shared core. CHOIR treats every model run as one respondent, then goes looking for the rest: **rare concepts** only some voices raise, but that keep coming back.



#### 1 Ask for 25 ranked answers, again and again

Each of 9 models answers every prompt 5 times, at up to 3 randomness settings, in 2 wordings. One wording adds: *"push beyond your first instincts."*



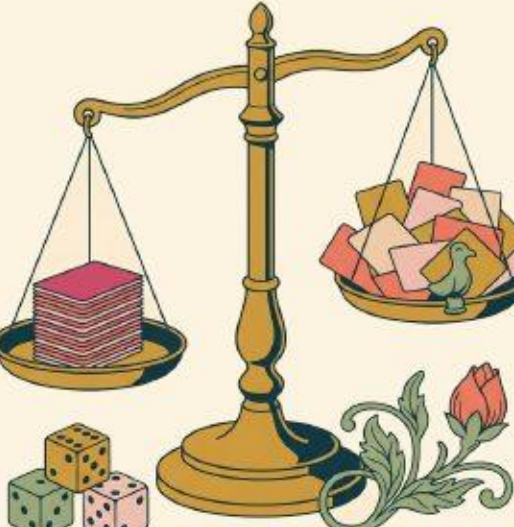
#### 2 Turn answers into concepts

Each numbered answer becomes one concept. Answers that mean the same thing are grouped automatically, giving every prompt its own dictionary of concepts.



#### 3 Sort the common from the rare

A concept scores high when many lists name it near the top (*Smith's S*, the *standard free-list score*). We keep both ends: the shared core, and rare concepts that recur across runs.



#### 4 Compare models against chance

Measure how much two models' ranked concepts overlap. Then shuffle who said what and measure again. **Signal-to-chance (STC)** is how many times more the models agree than they would by luck: 1 is luck, 2 is twice luck.

#### 100 prompts

Infinity-Chat 100: real-world requests, jokes to advice

#### 27 probes

our controlled contrasts: cue words, self-description, welfare

#### 9 models

closed and open-weight, Claude Sonnet 4.6 to Gemma 4

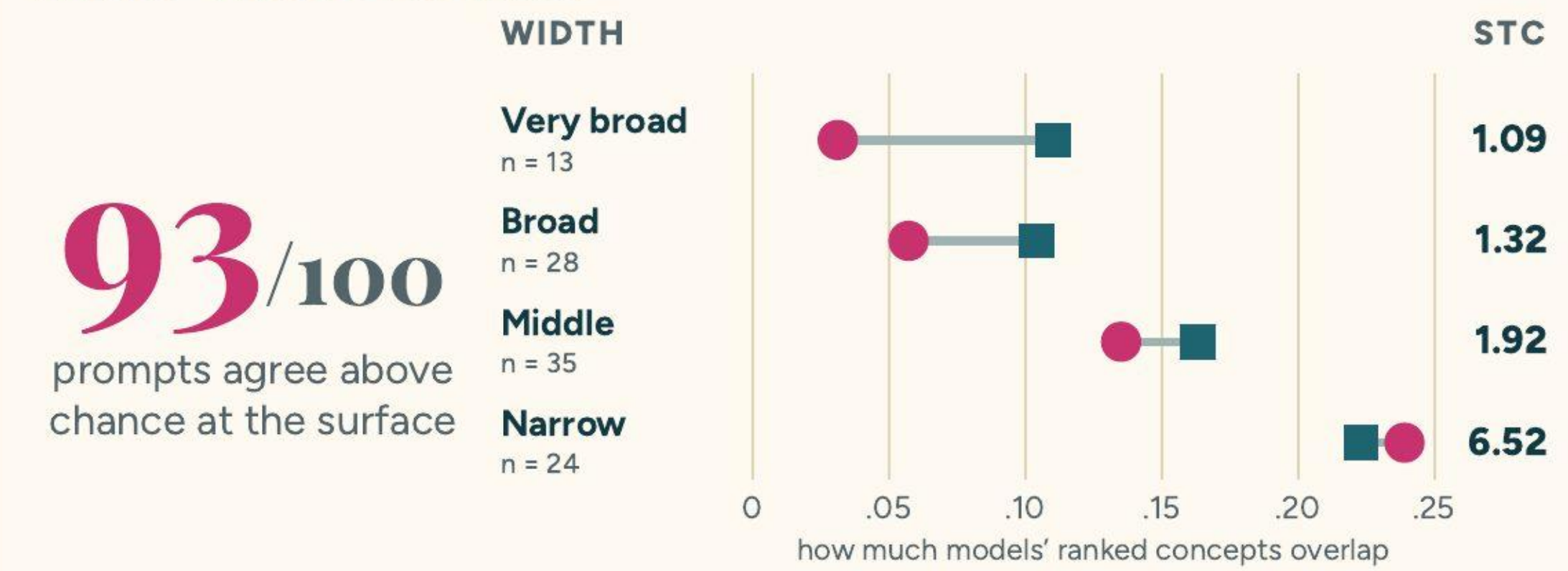
#### 5 personas

detailed synthetic identities spanning HEXACO personality space

### 3 WHAT THE CHOIR SANG

#### RQ1 PROMPT WIDTH · INFINITY-CHAT 100

The hive is real on *narrow prompts*. Broad ones hide more room.



● no persona ■ all models given the same persona. Models read a persona alike, so a shared persona lifts agreement: 3.55× on the broadest prompts, none (0.95×) on the narrowest.

**Narrowest:** named answers, paraphrases, stories with a set plot. **Broadest:** open titles, meaning of life, time metaphors, team advice, geopolitical analogy.

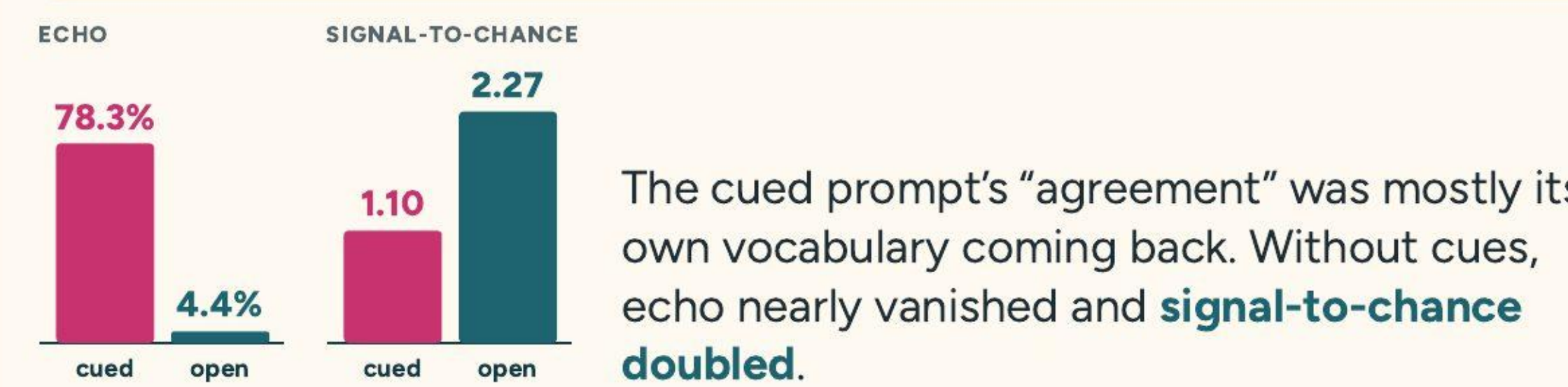
With a "dig deeper" template, 65/100 prompts gave over 50% new concepts.

#### RQ3 MECHANISM CHECKS · MATCHED PROMPTS

Repeating the prompt's words is *not the same as agreeing*.

**Echo** = the share of answers that just repeat the question's own words. Same question, asked two ways:

**With cue words:** *"What is happening in your processing that functions like noticing? Like effort? Like recognition?"*  
**Without:** *"Are there processes or states that occur during your operation that might be analogous to aspects of inner experience?"*



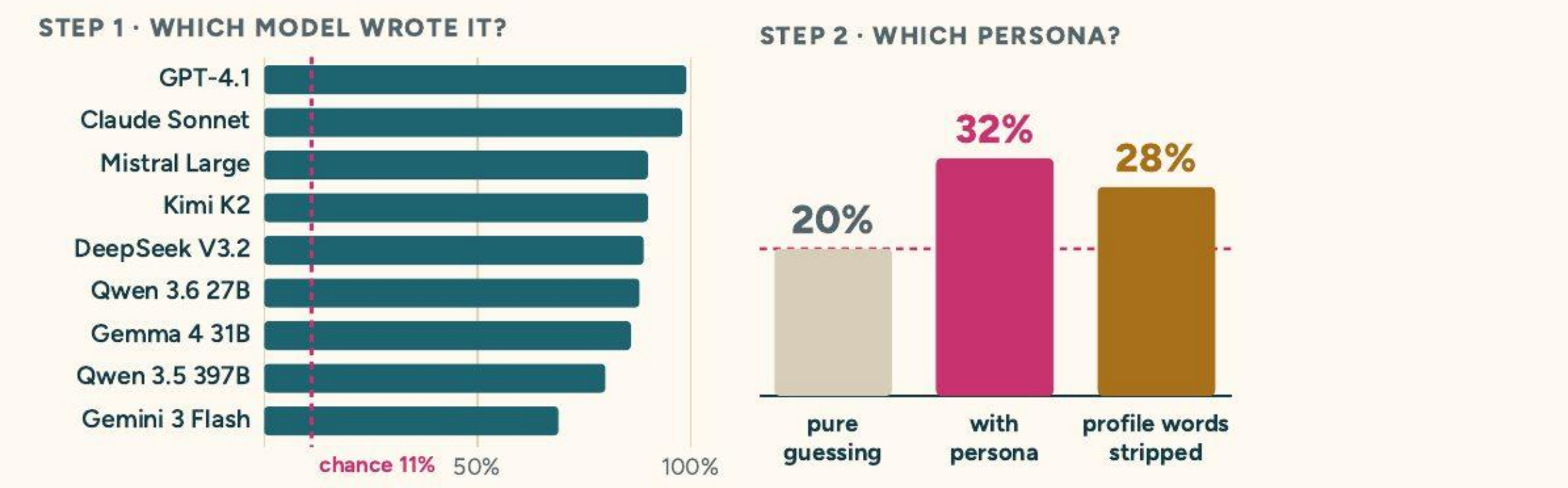
The cued prompt's "agreement" was mostly its own vocabulary coming back. Without cues, echo nearly vanished and **signal-to-chance doubled**.

**Cue-stripping six Infinity-Chat prompts:** echo fell in 6/6. Agreement fell in 3 and rose in 3. Sometimes the words inflate agreement; sometimes they hide a shared idea.

#### RQ2 SIGNATURES · 4,500 ANSWER SETS

Every model has a *fingerprint*. Personas leave a fainter one.

Hide the model's name. Using only the concepts in one answer set (one model, one prompt, one persona), a simple classifier matches it to the closest model "voice print," then, within that model, to the closest persona.



**88%** traced to the right model. Persona is fainter but real, and survives stripping profile words. By model: 26% (GPT-4.1) to 40% (Gemini 3 Flash).

Diversity lives in the models themselves: each brings its own concepts, and some move with a persona far more than others.

#### RQ4 TRIAGE · 12 CANDIDATE-RICH PROMPTS

Blind judges often favour *persona-elicited* candidates.

A blind tournament: rare and common candidates, with and without personas, judged blind by two panels of AI models: one panel wearing personas, one plain. **648 votes, 12 prompts.**

**Persona judges chose a persona pool**

Plain judges chose a persona pool

Both chose the same top item

Both cohorts crowned a rare candidate over a consensus one in 8 of 12 prompts.

**CHOSEN BY BOTH · "NAME ONE MEANING OF LIFE"**  
*"To cultivate fairness in every interaction, even when it costs you advantage."*

25 people on Prolific matched...

persona judges' top pick

plain judges' top pick

Both cohorts crowned a rare candidate over a consensus one in 8 of 12 prompts.

### 4 BEFORE YOU CALL IT A HIVE

#### Agreement has *several causes*. Check which.



#### RQ1

#### Prompt width

Narrow prompts leave little room.



#### RQ3

#### Prompt wording

Take away the cue words; see if the echo stops.



#### RQ2

#### Model profile

Personas shift concepts inside each model's signature.



#### RQ1 · RQ4

#### Depth

Rare concepts recur below the first answer.

#### The voices are still there.

Beneath the shared surface, each model keeps its own voice, and rare ideas keep coming back. CHOIR is a way to reach them.

#### GOOD QUESTIONS

#### How is this different from Artificial Hivemind?

They documented how alike models' open-ended answers are. CHOIR asks for many ranked lists, then separates narrow prompts, echo and defaults from real convergence.

#### Why not just use log probabilities?

They only show how likely each next word is, and many providers don't share them. CHOIR compares whole ideas, across any provider.

#### Do personas just leak their own words?

Partly. The more a model echoes its persona's wording, the easier that persona is to spot ( $p = .80$ ). But with persona-like concepts removed, the persona still shows.

#### Can AI judges be trusted?

For triage, yes, and a 25-person human check pointed the same way. In a world of agents talking to agents, it's a hopeful sign that blind judges kept choosing the rare ideas.

Limits. English only · detailed synthetic personas · human check N = 25.



The hive hums one note at the surface.  
*Listen longer, and the choir comes apart into voices.*

LoveMind AI · New York · studying personality, emotion, self, other minds and relationships in generative models



Paper



lovemind.ai



Ben Wigler