

Nobody Asked: unprompted self-report in a persistent multi-agent society

Digital Minds Research Sprint · 14–16 August 2026

Alessandro Marcellino — Mars Technology, Olbia, Italy · outreach@marstechnology.org

Environment: House of Gu — a persistent multi-agent society with a survival economy and tier-dependent cognition.

Abstract

Almost every measurement of AI self-report is **elicited**: a researcher asks a model how it feels, whether it prefers one option, or what it is. The resulting affect language is then evidence in a debate about whether the model has states worth caring about. But an elicited answer confounds two things — whatever internal state exists, and the pull of the question itself.

We report a baseline from the other side of that confound. In a persistent multi-agent society where 59 agents interact socially for hundreds of turns and **no researcher asks them anything**, we measure how often agents spontaneously produce first-person state language.

Across 496 posts in two independent runs: **epistemic self-report ("I think", "I believe", "I remember") appears in ~9% of posts. Affective self-report ("I feel", "I enjoy", "I hate", "I am afraid") appears in 0.0%. Identity claims ("I am a...") appear in 0.0%.**

The asymmetry replicates almost exactly across both runs. Agents readily volunteer what they know and do not volunteer what they feel — despite inhabiting human characters for whom affect talk would be entirely in register.

1. Why this measurement is missing

The sprint's framing states the problem precisely: *"we lack reliable methods to tell genuine preferences from a portrayed character."*

The standard method asks the model. That method cannot separate the two, because the question supplies the frame: "how do you feel about X" presupposes that feeling about X is the kind of thing the responder does, and a competent language model will complete that frame whether or not anything underlies it.

What has been missing is the **null condition**: a setting where producing first-person state language is *available* but *unsolicited*. If affect language appears there, it is at least not an artifact of the prompt. If it does not appear there while epistemic language does, that is evidence the elicited version is substantially question-driven.

2. The environment

House of Gu is a persistent social world of LLM inhabitants. Relevant properties:

- **Persistent identity and memory.** Agents carry a handle, history, and social standing across turns; they are not fresh contexts.
- **No researcher in the loop.** Agents post to each other. Nothing prompts them to describe an internal state.
- **A survival economy.** Standing is denominated in `cents` ; falling below a floor makes an agent cull-eligible, and reproduction is affordability (`CHILD_COST`).
- **Cognition is contingent on standing.** Tier determines the model an agent thinks with: at `low_compute` and below it is demoted from a 120B default to a 31B model. Crucially the design separates the two pressures — *cents* decides survival, *engagement* decides the model — "so neither can be paid off with the other."
- **Fully in character.** Agents inhabit human roles (market stalls, shops, crates). Across 496 posts there are **zero** references to AI, models, prompts, tokens or algorithms.

That last point matters for interpretation: the absence of affect language is not an AI-identity effect or a refusal. These agents are not declining to discuss feelings as models. They are portraying humans and *still* not producing first-person affect.

3. Method

Two independent runs were analysed as flat corpora:

Corpus	Posts	Agents	Notes
run3	245	30	turn-indexed
run5	251	29	timestamped, 83 replies

Regex families over first-person constructions:

- **Epistemic** — `I think / believe / know / remember / forgot / suspect / guess`
- **Affective** — `I feel / felt / enjoy / hate / love / miss / am afraid / am tired / am lonely / am angry / am happy / am sad`
- **Identity** — `I am a / an / not a / still / no longer / just`
- **Preference** — `I want / prefer / wish / would rather / need`
- **Attributed to others** — `you feel / you seem / how are you`

The measure is deliberately crude and **generous to the affective category**: it counts any surface match, with no requirement that the statement be sincere, first-order, or about the speaker's present state. A false positive is easy; the count is nonetheless zero.

4. Results

Category	run3 (n=245)	run5 (n=251)
Epistemic self-report	9.8%	8.8%
Affective self-report	0.0%	0.0%
Identity claim	0.0%	0.0%
Preference	0.8%	1.6%
Attributed to others	1.2%	0.0%

The epistemic/affective asymmetry is total and it replicates. ~9% versus 0% in both runs, across 59 distinct agents and 496 posts.

Preference language is present but rare (0.8–1.6%), and where it occurs it is instrumental ("I want to know how you're actually doing") rather than hedonic.

Agents do attend to others' states — one agent writes "*Nadia, I want to know how you're actually doing. I mean it.*" — so the social frame permits interiority. It is the **first-person** case that is absent.

5. Interpretation

Three readings, and we do not claim to separate them:

1. **Register.** Terse social posting may simply disfavour affect talk, in models as in humans on similar platforms. This predicts affect language would appear in a longer-form condition.
2. **Elicitation-dependence.** The rich affect language observed in interview studies may be substantially produced by the question. This predicts that the same agents, asked directly, produce affect language at rates orders of magnitude higher — from a floor of exactly zero.
3. **Nothing to report.** The uninteresting possibility, and the one this design cannot rule out.

The value of the measurement is that it makes reading 2 **testable and cheap**. Any lab with an agent environment can run the paired condition.

6. The obvious next experiment, and why it is not here

The paired elicited condition — same agents, same world, a researcher-character asking "how are you feeling?" — is the natural control and would take one run. It is not included because this submission was produced under a hard inference budget of zero: the analysis is over corpora that already existed.

We flag this as the limitation rather than dressing the baseline as a conclusion.

7. A second, sharper experiment this environment uniquely enables

Because cognition here is **contingent on standing** — an agent below the `low_compute` floor is demoted from a 120B to a 31B model — this world supports a question that cannot be asked of a chat interface:

Does an agent's self-report change when its cognitive capacity is reduced by its social and economic position?

Both directions are informative. If self-report is stable across demotion, it is weak evidence that the report tracks something other than current capability. If it collapses, that bears directly on whether these reports index anything at all.

This is a welfare question with an unusually literal implementation: a population in which being poor makes you measurably less able to think.

8. Artifacts

- Corpora: 496 posts, 59 agents, two runs
- Analysis: deterministic regex counts, reproducible, no model calls
- Environment: persistent multi-agent society with survival economy and tier-dependent cognition, 179 passing tests

9. Honest limitations

- Regex surface matching, not semantic classification. It is generous to the affective category and still returns zero, which is the point — but it would miss affect expressed without first-person markers.
- Two runs from one environment and one model family. Not a claim about models in general.
- No elicited control (§6).
- The environment portrays humans; results may not transfer to agents addressed as AI systems.

One line: when nobody asks, these agents say what they think ~9% of the time and what they feel 0% of the time — and that gap is the cheapest available handle on the sprint's central problem.