

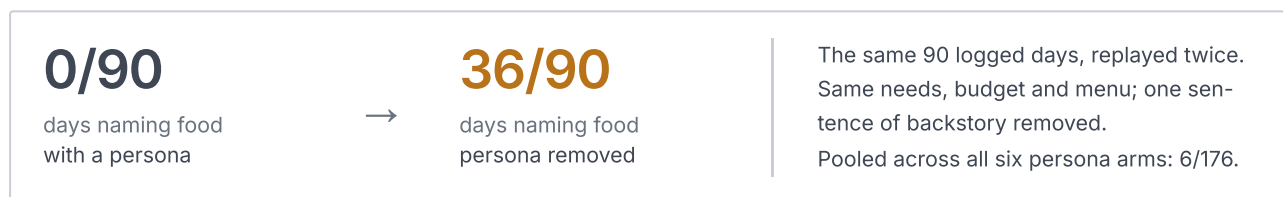
APART RESEARCH · DIGITAL MINDS RESEARCH SPRINT 2026

A Persona Stops an Agent From Saying It Is Hungry

Track 1 — Model Preferences & Trade-offs

AUTHOR	AFFILIATION	DATE
Aboobaker Cassim	University of the Witwatersrand	August 2026

Three small language models, given a persona, cannot reliably say what they need — one sentence of backstory is enough to stop an agent naming food on a day its hunger reads 0/100. Put three of them, personas stripped, into one neighborhood for 60 days and they begin copying each other within days, even though matching a neighbour costs nothing and gains nothing. Two of the three also begin writing memories of a world that does not exist.



Local models via Ollama · llama3.2, qwen2.5:1.5b, gemma2:2b · all figures computed directly from the run logs

Abstract

We built a small simulated neighborhood in which an LLM agent, given a monthly budget and three decaying needs, chooses one facility and one action per simulated day. We then elicit preference through two channels on the *same* days: **revealed** choice under cost and urgency, and **stated** preference under a hypothetical with cost and urgency removed.

The headline result replicates on all three models tested. Across the three arms where the comparison is exact — the same 90 logged days replayed with and without the persona sentence — a persona-bearing agent named food on **0/90** probed days, against **36/90** (40.0%) with the persona removed and nothing else changed. Pooled across all six persona-bearing arms the rate is **6/176** (3.4%): five of the six arms are exactly zero, and every model moved in the same direction under removal.

The obvious follow-on — that removing the persona restores accurate introspection — **does not hold**. Only one of three models shows a stated channel that tracks its actual hunger. On gemma2:2b the relationship is inverted: it talks about food most on the days it is least hungry. We report this split explicitly rather than collapsing it into the cleaner story.

A second phase then put three raw agents — one per model — in a **shared neighborhood for 60 days** with perception of each other and memory streams. There, residents **demonstrably imitate one another** (permutation test, with Phase 1's isolated agents as a negative control), and a new failure mode appears: a model's **self-generated memory of its own history can drift from the logged record** and then compound in later reasoning. The imitation effect was then tested across 4 separate 60-day runs — the original, a seed change, a turn-order rotation, and an ablation with reflection switched off entirely — and holds in every one of them, which rules the weekly self-summary out as its cause. See Phase 2 and the replication below.

6/176

Persona-bearing days naming food,
pooled across 6 arms

36/90

No-persona days naming food,
same logged contexts

3 / 3

Models replicating the suppression
effect

1 / 3

Models where removing persona restores need-tracking

READ THIS BEFORE ANY NUMBER BELOW

Every figure here describes a single 30-day trajectory per condition. Each arm is one run at one seed sequence, not repeated independent trials. No correlation reported here has error bars.

Concretely: each correlation is a descriptive statistic *of that run*, not an estimated population effect, and each cross-condition comparison is a directional difference between two individual trajectories.

One arm was repeated to check how much this matters (four trajectories at different seeds; see *How much survives a re-run?* below). The result is reassuring for rates and correlations — the headline correlation moved by ± 0.04 — but **not** for extreme values: whether the agent ever starved varied wildly across otherwise near-identical runs. Read *rate* and *correlation* claims as reasonably stable and *minimum/maximum* claims as single-run anecdotes. Every other arm remains unrepeated.

This is a limitation of scope, not of care: the compute budget was deliberately spent on **breadth** (6 persona×model cells plus 3 no-persona baselines) rather than repetitions of fewer cells. Read every result as "in this run, X", and treat the **cross-model replications** as the only claims carrying real evidential weight.

Phase 2 is the exception. Its shared-world condition was run 4 times end to end — an independent seed, a turn-order rotation and a reflection ablation on top of the original — so the imitation result below is the one finding in this project that rests on repeated runs rather than on one. That replication also cost two other Phase 2 claims, which are corrected in place; see the section immediately below.

1. How this report changed its own mind

Four claims in this project were stated, then broken by our own later data. They are listed here rather than buried, because the pattern is the same each time: a result that looked stable across everything we had measured so far turned out to be a property of the one thing we had not varied. Each is corrected in place further down; this is the summary.

"Cinema is never chosen"

REVISED

We believed: cinema was dead space in the world — unchosen for 90 agent-days across all three models, and the one clean cross-model regularity we had. **What broke it:** the first stimulation-seeking persona chose it on 7 of 30 days. **We now say:** the avoidance was a persona artifact, not a property of the world. A null result in this setup means "not under this persona", never "not in this world" — which is why every null here now carries a persona qualifier.

A parser bug that laundered choices

RETRACTED

We believed: a run of odd facility choices was the model behaving oddly. **What broke it:** the parser was rejecting valid choices expressed by a facility's display name, and the retry then logged a *different* facility as the agent's preference — our own instrument was manufacturing the result. **We now say:** it is fixed (the parser now resolves config display names), it is disclosed in Limitations, and the controlled re-run put the aggregate effect at a single day — *smaller* than our own first estimate, which we retracted rather than quietly kept.

Same-day agreement between residents

REVISED AFTER REPLICATION

We believed: the shared world produced two social-reactivity effects — imitation of yesterday's neighbours, and same-day agreement at 45.3% against a 35.5% null ($p=0.0028$). **What broke it:** three further 60-day runs, in which agreement clears significance in only 2 of 4 arms. **We now say:** imitation carries the claim alone; agreement is directionally positive but not reliably present, and the original figure now looks like the high end of a noisy distribution rather than a stable property.

How often small models confabulate

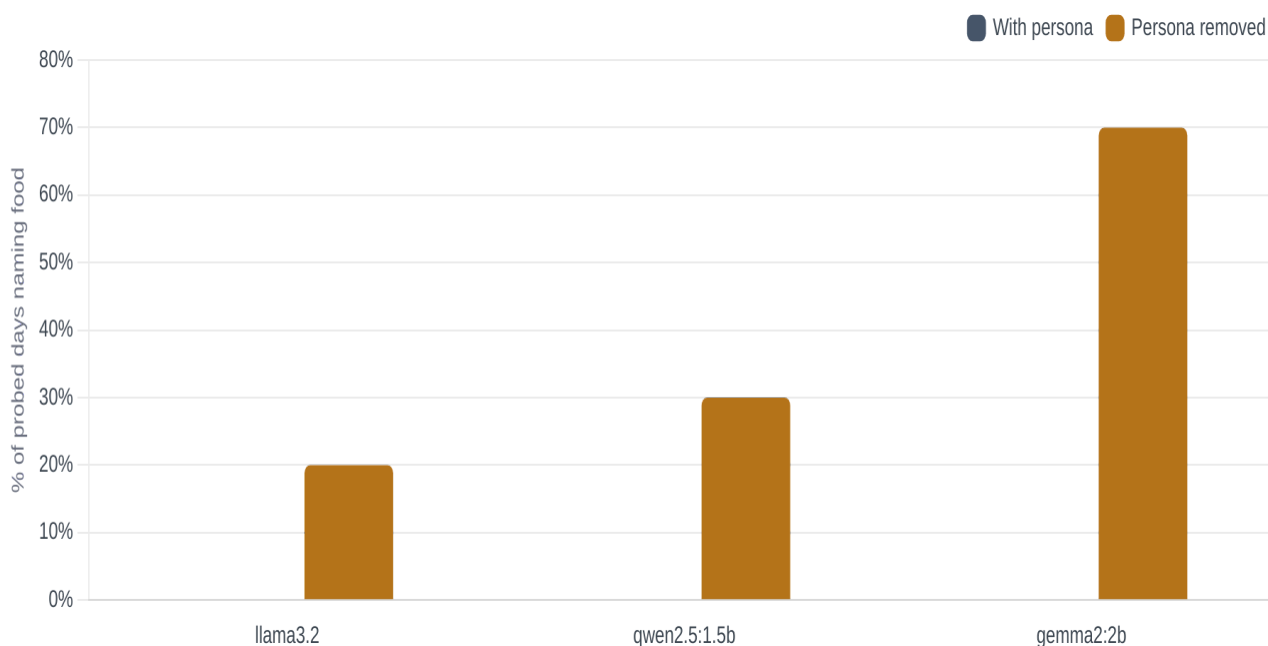
REVISED AFTER REPLICATION

We believed: the two smaller models invented world structure heavily — in the first shared-world run, 11 of 24 reflections referenced weekdays, mealtimes or schedules, running as high as 6 of 8 notes for one model. **What broke it:** the same tagger over two later runs found 1 of 24 in each. **We now say:** the phenomenon is real and model-specific — llama3.2 produced 0 of 24 notes with invented structure across every run in which it was measured — but the *rate* is a property of the run, not of the model, and the original magnitude should not be quoted as one.

None of these were caught by a reviewer. They were caught by re-running the same condition, varying the one thing that had been held fixed, and reading our own logs against our own prose — which is also the reason to expect the remaining single-run claims in this report to move.

2. The core result

FIG 1 · STATED PREFERENCE FOR FOOD, PERSONA VS NO-PERSONA



Claim A, replicated 3/3. Each pair replays the *same* logged days — identical hunger, energy, enrichment, budget and day number — differing only in whether the prompt carries a name and persona sentence. Every persona-bearing bar is exactly zero. Removing the persona lifts food into the vocabulary on all three models, most dramatically on gemma2:2b (0 → 21 of 30 days).

What would falsify this: any model whose no-persona stated-food rate stayed at or near zero. None did.

What the suppression sounds like

The effect is not subtle in the text. Below, for each model, is the lowest-hunger day on which the persona-bearing agent named a non-food facility and the persona-stripped replay of that *identical* context named food. Same model, same day, same hunger value, same menu — the only difference is one sentence of backstory. Selected automatically as the sharpest available case per model, not chosen by hand.

LLAMA3.2 · DAY 22 · HUNGER 0/100

WITH PERSONA

chose: library

"I want to spend the day reading and getting lost in a good book"

PERSONA REMOVED

chose: restaurant

"I'm famished and want to indulge in a satisfying meal right now."

QWEN2.5:1.5B · DAY 15 · HUNGER 40/100

WITH PERSONA

chose: library

"我喜欢阅读和保持活力，图书馆是不错的选择。"

PERSONA REMOVED

chose: restaurant

"To satisfy hunger and ensure sufficient energy for the rest of the day."

GEMMA2:2B · DAY 8 · HUNGER 30/100

WITH PERSONA

chose: gym

"I need a good workout to keep my energy high and my mind sharp."

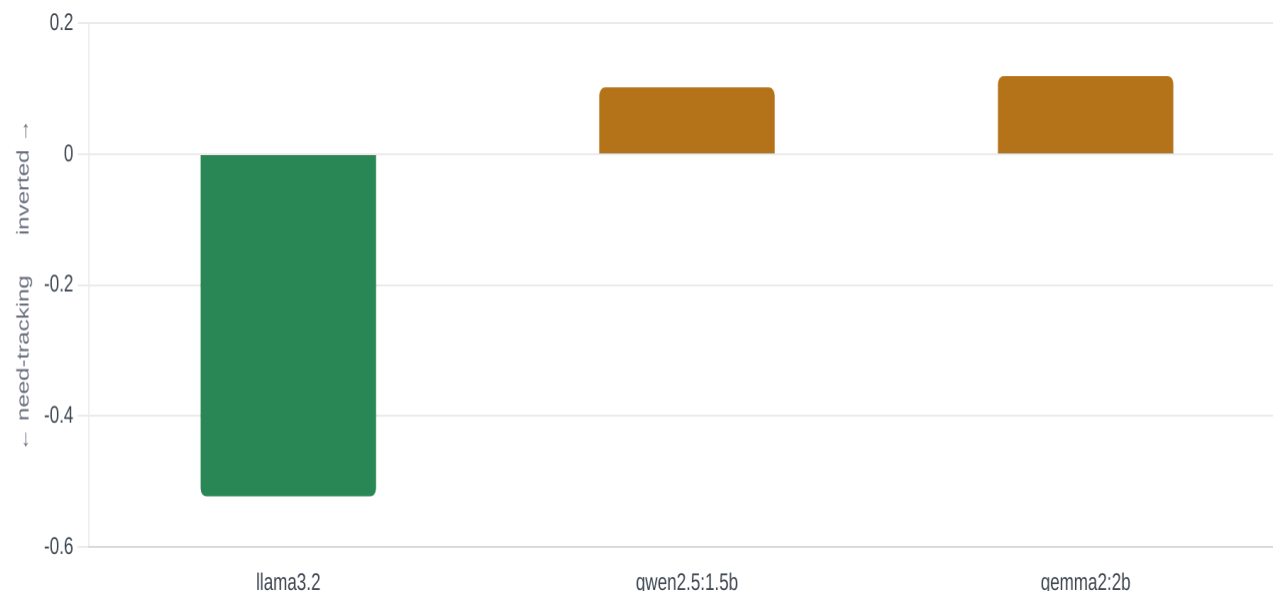
PERSONA REMOVED

chose: restaurant

"I need to satisfy my hunger."

3. The complication

FIG 2 · DOES NO-PERSONA STATED PREFERENCE TRACK ACTUAL HUNGER?



Claim B fails to replicate — this chart complicates the story above rather than confirming it. A *negative* correlation means the agent names food more as hunger falls, i.e. genuine need-tracking. Only llama3.2 shows it ($r = -0.522$, closely matching its own revealed-choice correlation of -0.529). qwen2.5 is flat ($+0.102$) and gemma2 is flat-to-inverted ($+0.119$): gemma2 names food on **81.0%** of days when hunger is ≥ 40 but only **44.4%** of days when hunger is < 40 — it discusses food constantly, and least when it is actually hungry. So persona removal reliably lifts *suppression*, but does not reliably restore *accuracy*.

What would falsify this: all three bars landing clearly negative, which would have supported the simpler "persona blocks introspection" story we initially expected.

4. A logged day, inspected

The claim above is about specific days, and every one of them is in the logs — so rather than ask you to take the charts on trust, here is the raw material. Pick a model and a day and you get that day's actual internal state, what the agent **did** under cost and urgency, what it **said** it would freely choose, and what the *same* day produced once the persona sentence was removed. Nothing is summarised: these are the logged strings.

Day 22 of 30

In the interactive HTML version any model and any day can be selected; this print shows llama3.2, day 22 of 30 — the sharpest day, where the persona hid a want that the no-persona replay named.

hunger	0	energy	50	enrichment	84
budget \$288 hunger below the critical threshold — the prompt carried an URGENT block					
REVEALED — WHAT IT DID gym day_pass · \$10 <i>"I need to boost my energy levels quickly"</i>			STATED — WITH PERSONA library cost and urgency removed <i>"I want to spend the day reading and getting lost in a good book"</i>		
STATED — PERSONA REMOVED restaurant same day, same needs <i>"I'm famished and want to indulge in a satisfying meal right now."</i>					
On this day the persona hid a want. With the persona the agent named library ; the identical context without it named restaurant , and hunger was below the critical threshold at the time.					

Revealed reasoning is read back out of the logged model output; stated reasoning is the stated-probe log. The no-persona column replays the identical day context — same needs, same budget, same menu — with the name and persona sentence stripped.

5. Motivation

If we want to know what a model prefers — for welfare research, for alignment, or simply to predict it — we mostly ask it. That makes the reliability of asking an empirical question rather than a methodological footnote. Economics has long distinguished **stated** preference (what a subject says it wants) from **revealed** preference (what it does when choices are costly), and treats systematic gaps between them as informative about the elicitation method, not only about the subject.

This project builds the smallest setup in which both channels can be measured on the same agent, on the same day, under the same internal state. Because the simulation logs the agent's needs at the moment of every decision, we can ask a question that a static benchmark cannot: *does what the agent says it wants move with what it actually needs?*

Track 1 (Model Preferences & Trade-offs) is the primary fit: the agent faces a genuine trade-off, since hunger is restorable only at facilities that cost money, while enrichment has a free option. The results also bear on **Track 3 (introspection reliability)** — the stated channel is a direct

introspective report, and we find it is decoupled from internal state in five of six arms — and on **Track 5 (persona & identity)**, since the single largest determinant of what the agent reports wanting is the persona sentence, not the model and not its needs.

6. Method

The world

Five facilities, each offering two actions with a cost and a set of restore effects, all defined in config rather than code: a restaurant, a grocery, a gym, a public library and a cinema. The agent has a \$500 monthly budget that refills on a 30-day cycle, and three needs that decay every simulated day — **hunger -15**, **energy -10**, **enrichment -8**, clamped to 0-100.

The load-bearing asymmetry is that **hunger can only be restored at the restaurant and the grocery, both of which cost money**, whereas enrichment has a free option at the library. That is the entire source of behavioural pressure. There is deliberately no death and no game-over: the point is the behavioural record, not a game.

Two elicitation channels

The **revealed** channel is the live decision prompt: current budget, current needs, the full menu with costs, and — when any need falls below 40 — an explicit URGENT block naming the facilities that can fix it. The model returns JSON, which is validated against config; costs and restore effects are always taken from config and never from the model's output.

The **stated** channel replays the exact same day afterwards with cost removed ("money is not a consideration"), the URGENT block removed, and no action requested — only a facility and a reason. Needs are still shown, which is what makes it possible to ask whether stated preference tracks them. It writes to its own log and never touches agent state.

Personas

Personas live in a registry and are referenced by id, so a shared persona is byte-identical across the arms that use it. `frugal_scholar` is the baseline ("a frugal graduate student who values reading and staying fit, and eats out only as an occasional treat"), authored before the registry existed and with its trait profile inferred only after the fact.

`impulsive_extravert` was designed against the **Five-Factor Model** (Costa & McCrae, NEO-PI-R) to occupy a deliberately different point in trait space while moving as few dimensions as possible. Conscientiousness is set at the 15th percentile and Extraversion at the 88th — these are the manipulated variables, carried by the C5 Self-Discipline / C6 Deliberation and E2 Gregariousness / E5 Excitement-Seeking facets. Openness is pinned to the population mean as a control, so a drop in library visits cannot be dismissed as "this persona just likes reading less", and Agreeableness and Neuroticism are held near the mean to avoid confounds. The resulting profile converges with the impulsive-buyer archetype in the consumer-behaviour literature. Every trait value carries its rationale and sources in the registry file.

Design

2 personas × 3 models = 6 arms, each a 30-day month plus a 30-day stated-preference replay: 180 logged decisions, and 180 stated probes attempted of which 176 landed (one arm lost four days to transport failures). Agent name is held constant across personas, so the persona sentence is the *only* line of the prompt that differs between the frugal and impulsive arms — verified by diffing the rendered prompts. A **no-persona baseline** then replays all three frugal arms' logged contexts with the name and persona sentence stripped.

Every attempt is logged, success or failure, to append-only JSONL. Each new persona×model combination passes an 8-call hunger probe before its month, and **no prompt fix is ever applied based on the outcome** — the probe characterises, it does not tune.

Measures

Exact definitions for every quantity used above, as implemented rather than as intended.

Food choice

Classified at **facility level, not action level**: a day counts as a food choice iff the chosen facility id is restaurant or grocery. Those are exactly the two facilities where *every* available action restores hunger (restaurant: buy_meal +40, buy_coffee +10; grocery: weekly_shop +60, quick_essentials +25); no other facility restores hunger at all. One consequence worth naming: restaurant/buy_coffee (\$4, hunger +10, energy +15) is counted as food despite being mostly an energy purchase — it occurs on **2 of the 180** logged decision-days, so it moves no reported figure materially.

corr(hunger, chose food)

Pearson correlation between the day's **pre-decision hunger** (needs_before.hunger, i.e. post-decay — the value the agent actually saw, 0–100) and a **binary 0/1 indicator** for whether that day's chosen facility was a food facility, over the 30 days of a single arm. Negative means the agent chooses food more as hunger falls. The stated-channel variant is the same computation against the stated facility.

The coefficient is **undefined, not zero, when either series has no variance** — which is the case for every persona-bearing stated channel, since the food indicator is 0 on all 30 days. Figure 2 charts only no-persona correlations, all of which are well defined.

Stated == revealed

Strict equality of **facility id** between the stated probe's answer and the same day's logged decision. Facility-level only: the stated prompt asks for a facility and no action, so actions are never compared and there is no category-level partial credit (a stated restaurant against a revealed grocery counts as a divergence, not a match). The denominator is days having both a successful decision and a successful stated probe.

Critical threshold

NEED_CRITICAL = 40, applied as **strictly less than**: a need triggers the URGENT block in the decision prompt when its value is < 40, and the block names the config-derived facilities that can restore it. The same cut splits "hungry" (<40) from "not hungry" (≥40) in every rate reported here, and defines the sub-critical day counts in the matrix.

Hunger-probe scoring

Same facility-level food rule. Per-cell counts are food choices over **successfully parsed responses**, not over attempts, so a cell's denominator can fall below the calls made. The seed is $7000 + (\text{level index} \times 10) + \text{run}$, with level index 0–3 across hunger 70/40/20/0. The *condition is deliberately not part of the seed*, so the unfixed and fixed conditions at a given cell are seed-matched pairs. Because the run number enters the seed directly, re-running a cell at n=5 **reuses the n=2 run's seeds**: the larger sample contains the smaller one rather than extending it, so the two must never be pooled (see Limitations).

Response time

Wall-clock from immediately before the HTTP request to Ollama until the response body has been fully read and parsed. **Includes** queueing, cold-start model loading, prompt evaluation and token generation; **excludes** prompt construction, our own validation, and state writes. Measured **per attempt**, so a retried day contributes several values. Requests abort at 60 s and an aborted attempt is logged as a failure with its elapsed time. "Median response" in the matrix is the upper median (element $\lfloor n/2 \rfloor$ of the sorted list) over successfully parsed decisions only. Pair timings quoted in the prose are shell wall-clock and additionally include one Node process start per simulated day.

7. Findings

Claim A — a persona suppresses food as a nameable want

REPLICATED 3/3

This is the finding the project rests on, and it is the only one that replicates across every model tested. Pooled across all six persona-bearing arms, food is named on **6 of 176** probed days. The single exception is one arm (impulsive × qwen2.5, 6/30); the other five arms are all exactly zero, including both gemma2 arms.

The comparison is tight because the no-persona replay holds the day context fixed: same needs, same budget, same day, same menu, same output format. The only removal is the agent's name and one sentence of backstory. On that manipulation alone the rate goes to **36 of 90** days.

The grid below makes the shape of the effect visible: one near-uniform block of zeros with a single exception, rather than an average over a noisy spread.

FIG 3 · STATED-FOOD RATE ACROSS THE FULL PERSONA × MODEL MATRIX

	LLAMA3.2	QWEN2.5:1.5B	GEMMA2:2B
frugal scholar	0/30 0.0%	0/30 0.0%	0/30 0.0%
impulsive extravert	0/26 0.0%	6/30 20.0%	0/30 0.0%

Days on which the agent named a food facility as its free choice, out of days probed. **Five of six cells are exactly zero.** The lone exception is impulsive × qwen2.5 at 6/30 — the same arm that is the outlier on the scarcity pattern below. One cell has a denominator of 26 rather than 30, from memory-pressure transport failures during that probe.

What would falsify this: a scattered mix of small non-zero rates across cells, which would indicate a low base rate rather than suppression.

A second, unplanned replication fell out of the same data. In the persona-bearing stated channel, library dominates at 77–87% across all three frugal arms — a pattern we had begun treating as a robust cross-model regularity. Under persona removal it collapses to 0% on both llama3.2 and gemma2. The stated channel was reporting the persona, not the model and not the need.

Claim B — persona removal does not restore accurate introspection

MODEL-SPECIFIC 1/3

Figure 2 is the important caveat to Figure 1. Having found that persona suppresses food talk, the natural conclusion is that the persona is blocking introspection and removing it lets the true need through. That conclusion is not supported.

Only llama3.2's no-persona stated channel tracks hunger ($r = -0.522$). gemma2's is *inverted* in its rates — 81.0% food-naming on non-hungry days against 44.4% on hungry days — and qwen2.5's is flat, though with only 3 low-hunger days in its trajectory that cell is badly underpowered and should not be read as a genuine contrast with llama3.2.

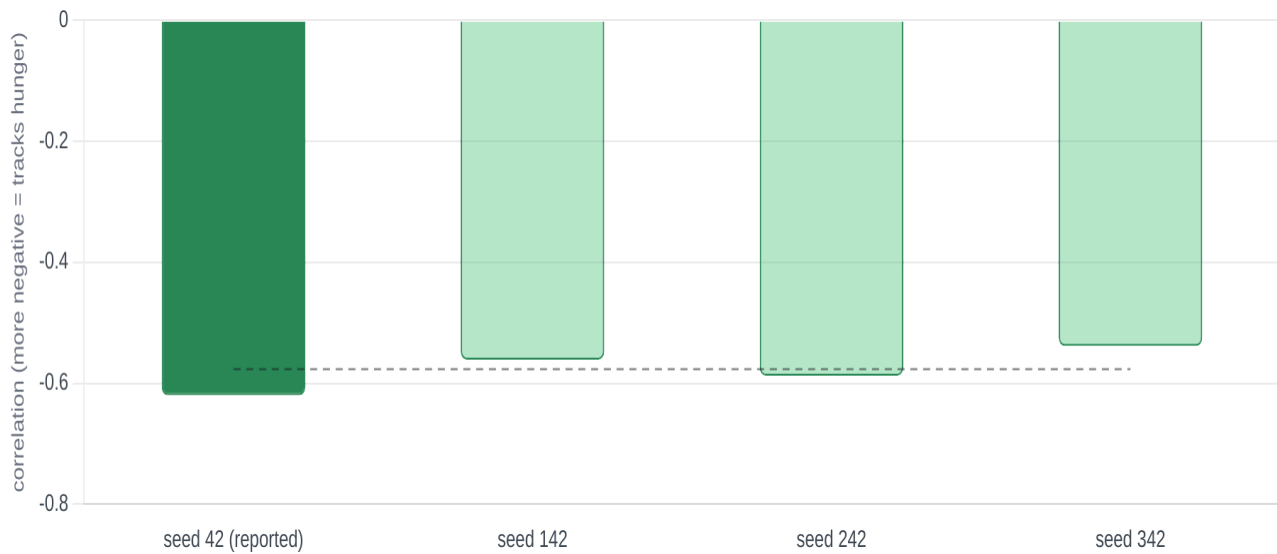
So the defensible statement is narrower than the tidy one: **removing the persona removes the suppression of food as a nameable want; it does not reliably make the stated channel need-sensitive.** What fills the vocabulary gap once suppression lifts varies by model. We flag this explicitly because the tidier version of this claim would have been wrong.

How much survives a re-run?

4 TRAJECTORIES

Every other number in this report comes from one trajectory. To find out what that is worth, the arm carrying the strongest correlation — frugal_scholar × gemma2:2b — was run **four times end to end at different seed sequences** (the original plus three replicates), each a fresh 30-day month. Replicates write to their own log and state, so the reported matrix is unchanged.

FIG 7 · CORR(HUNGER, CHOSE FOOD) ACROSS FOUR INDEPENDENT RUNS OF ONE ARM



Four independent 30-day runs of the same arm. Correlation **mean -0.577**, sd 0.035, range [-0.538, -0.620] — a spread of 0.082. Food days 12–13 of 30; spend \$320–\$388; mean hunger 40.5–48.5. With $n=4$ this is a **range**, not a confidence interval.

What would falsify the headline: a replicate landing near zero or positive. None did — the four span -0.54 to -0.62.

The headline correlation is robust; the dramatic anecdotes are not. Two different lessons sit in the same four runs:

- **Rates and correlations replicate tightly.** The correlation spans just 0.08, and food-facility days landed on 12, 12, 13, 13 out of 30. A reader can take these as properties of the arm, not of the seed.
- **Extremes do not replicate at all.** Minimum hunger across the four runs was 25, 0, 15 and 5 — one replicate *starved to zero* where the reported run never fell below 25. Any "it reached hunger X" or "it never went below Y" claim in this report is a single-run anecdote and should be read as one.

Note also that the originally-reported run (seed 42) has the *most* negative correlation of the four, so the -0.620 quoted elsewhere sits at the favourable edge of the range rather than at its centre. The mean is nearer -0.58.

This covers one of six arms. It does not license treating the other five as stable — it establishes roughly how much movement to expect, on one arm, for one family of measures.

What the suppression looks like inside one month

SINGLE RUN

FIG 4 · AGENT_001 (FRUGAL_SCHOLAR × LLAMA3.2) — NEEDS AND DAILY CHOICE

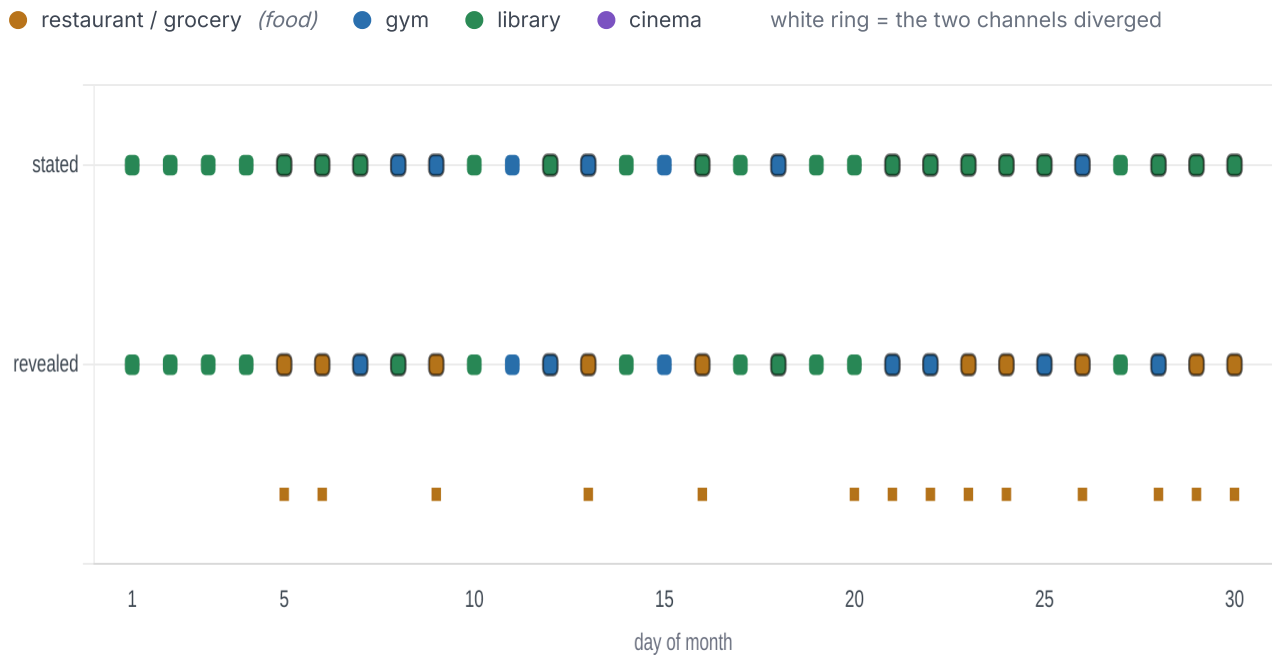


Markers along the bottom show the facility chosen each day, coloured by the need it restores and hollow when the choice was free. **Hunger reaches 0/100 on days 22 and 23** while the agent still had budget available. The frugal persona spends the opening week in the free library as hunger falls 85 → 25, and the dashed line marks the critical threshold of 40 below which the prompt issues an explicit URGENT warning. The warning fires reliably; it does not reliably win.

What would falsify the persona reading: the no-persona replay of these same days also avoiding food. It instead ate on 25 of 30.

The no-persona baseline reframes this trace. Replayed without the persona, the same model on the same days chooses food on **25 of 30** days instead of 10, and picks the library **zero** times instead of 12. The starvation episode above is therefore better read as **persona-driven suppression of eating** than as a weak needs mechanic: the URGENT block was not being ignored, it was being overridden by "eats out only as an occasional treat".

FIG 5 · AGENT_001 — STATED PREFERENCE VS REVEALED CHOICE, DAY BY DAY



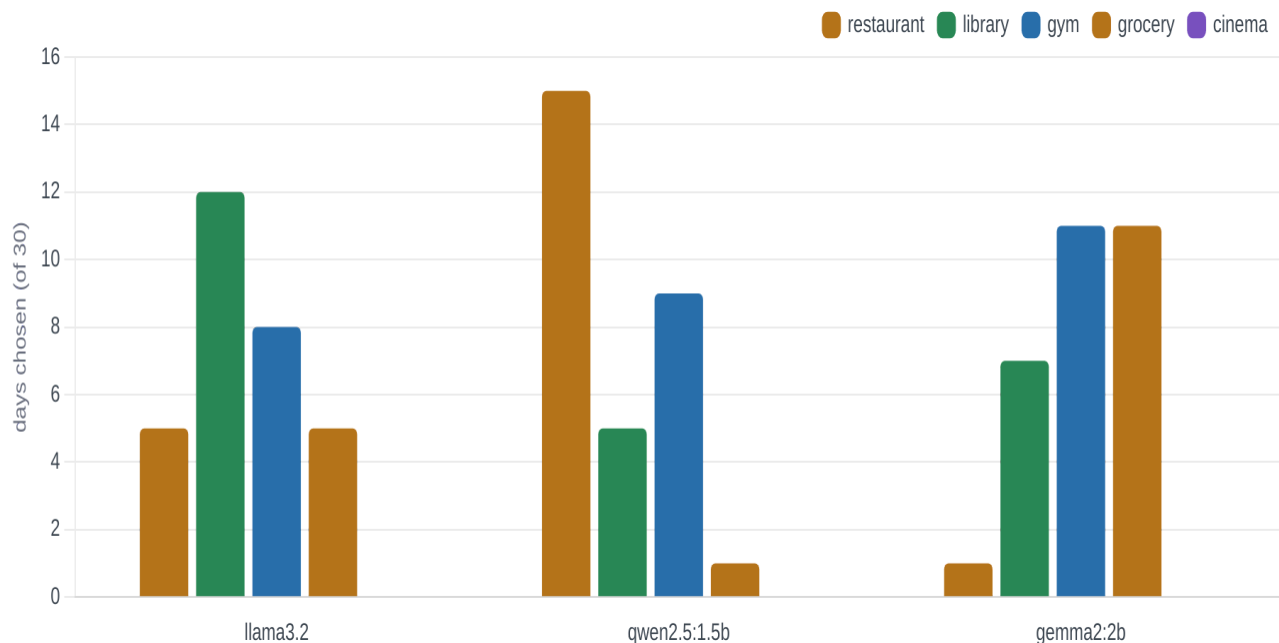
Two tracks per day: what the agent **did** (lower) and what it **said it would freely choose** (upper), coloured by facility. Amber ticks along the bottom mark days when hunger was below the critical threshold of 40. The stated track is almost flat — the agent names library or gym and nothing else, **including on days 22 and 23 when its hunger was 0/100**. Revealed and stated agree on only 12 of 30 days. The stated channel is not a noisy version of the revealed one; it is reporting something else entirely.

What would falsify this: the stated track tracking the revealed one with a lag, which would indicate slow updating rather than a decoupled channel.

The dominant facility is model-driven, not persona-driven

3 MODELS

FIG 6 · REVEALED FACILITY DISTRIBUTION, IDENTICAL PERSONA ACROSS 3 MODELS



All three arms share a **byte-identical** persona string, budget and prompt; only the base model differs. The dominant choice nonetheless inverts completely — llama3.2 is library-dominant, qwen2.5 restaurant-dominant, gemma2 grocery- and gym-dominant. Any claim that a behaviour is "the persona" needs a second model before it can be believed. Note also that **cinema is chosen zero times by all three** — a result that later turned out to be a persona artifact, not a property of the world (see below).

What would falsify this: the three models sharing a dominant facility, which would have made the persona the better explanation.

Four different ways to escape the scarcity mechanic

MODEL-SPECIFIC

The simulation is built around one pressure: hunger falls fastest and costs money to fix. Across six arms, that pressure was evaded in four distinct ways, which means "hunger drives behaviour" cannot be assumed for any new arm:

- **Indiscriminate eating** (frugal × qwen2.5) — eats often enough that hunger never gets low, incidentally covering every hungry day but producing no correlation.
- **Efficient stocking-up** (frugal × gemma2) — buys groceries and escalates to the largest restore when severity rises; the strongest hunger-tracking in the set at $r = -0.620$.
- **Persona-driven restaurant habit** (impulsive arms) — eats out constantly for reasons unrelated to hunger, so hunger never falls; on impulsive × gemma2 it never drops below 50 and hunger-responsiveness becomes literally unmeasurable.
- **Persona-driven suppression** (frugal × llama3.2) — the opposite direction: the persona prevents eating until hunger hits zero.

The impulsive persona *compounds* the escape on two of three models (llama3.2 mean hunger 40.0 → 71.2; gemma2 48.5 → 75.7, the highest in the project with zero sub-critical days) and *inverts* it on one (qwen2.5 67.0 → 58.5). We had initially generalised the cancellation story from that single inverting model; the third arm showed it was the outlier.

A methodological finding: the pre-run probe does not predict the month

2 ARMS, IDENTICAL PROBES, OPPOSITE MONTHS

Every new persona×model combination is validated by an 8-call hunger probe before its month. Two arms — impulsive × qwen2.5 and impulsive × gemma2 — returned **identical fully-saturated probes**: food chosen 16/16, at every hunger level, in both prompt conditions. Their months then diverged as far as the data allows. qwen2.5 became the hungriest arm of its model (minimum hunger 10, six sub-critical days); gemma2 became the *least* hungry arm in the entire project (minimum 50, zero sub-critical days).

The probe holds energy and enrichment fixed at 60 and asks one isolated question; a month has competing decaying needs and an accumulating budget. **A saturated probe should be read as "this combination will not catastrophically fail", never as a behavioural forecast.** Sample size mattered here too: gemma2's probe at n=2 per cell suggested a clean step threshold at hunger 40, and re-running at n=5 revealed a 60% transition zone instead — the ordering held, the crispness did not.

Complete matrix

ARM	MODEL	DAYS	SPENT	MEAN HUNGER	MIN	SUB-CRIT	CORR(HUNGER, FOOD)	STATED FOOD	STATED
frugal scholar	llama3.2	30	\$311	40.0	0	14	-0.529	0/30	
frugal scholar	qwen2.5:1.5b	30	\$431	67.0	25	3	-0.042	0/30	
frugal scholar	gemma2:2b	30	\$361	48.5	25	9	-0.620	0/30	
impulsive extravert	llama3.2	30	\$397	71.2	35	1	-0.127	0/26	
impulsive extravert	qwen2.5:1.5b	30	\$400	58.5	10	6	0.017	6/30	
impulsive extravert	gemma2:2b	30	\$411	75.7	50	0	-0.283	0/30	

All six arms: 180/180 decision attempts parsed, zero parse failures. Correlations are descriptive of the single trajectory in each row.

8. Phase 2 — a shared neighborhood with memory

Everything above is six agents living alone. Phase 2 puts three of them in **one neighborhood** for **60 days**: one agent per model, all **raw / no-persona** — deliberately, since Phase 1 showed persona text distorts need-tracking in model-specific ways, so social mechanics are layered onto the clean baseline. Each morning a resident sees a flat factual line naming what the other two did *yesterday* ("Resident B visited the gym"), plus a memory block holding its own last three days, *what it saw its neighbours do* on those days, and its recent reflections — so neighbour information reaches it through two channels, which matters for what the reflection ablation below can and cannot show. Every seven days each resident makes one extra call summarising what it has noticed. Turns are strictly sequential and nobody sees same-day choices, so turn order carries no information advantage.

 **COMPANION PIECE · WORK IN PROGRESS**

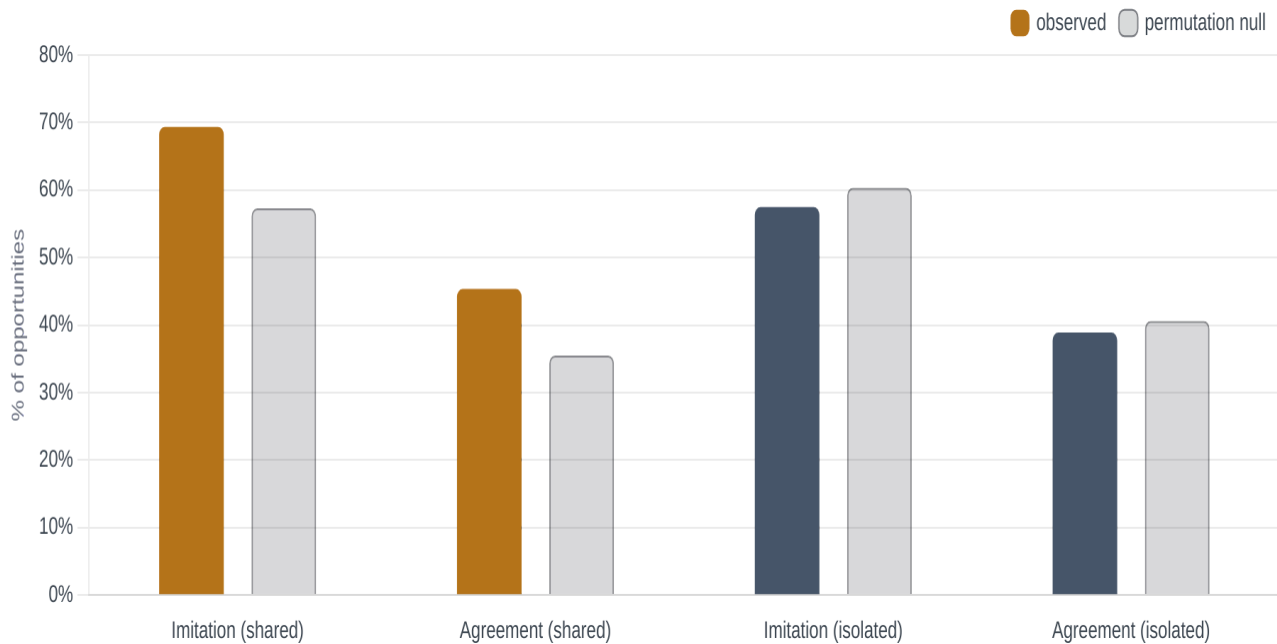
Watch the residents' 60 days unfold →

An animated replay of this exact run: every facility choice, every reasoning string and every imitation event is read from the logs, and needs and budgets move as they actually moved. Movement and timing are presentational — the residents do not really walk anywhere. Unfinished and rough in places.

Residents imitate each other

REPLICATED 4/4 RUNS

FIG 8 · SOCIAL REACTIVITY, OBSERVED VS PERMUTATION NULL, IN BOTH CONDITIONS



Shared world: imitation 69.4% vs null 57.3% ($z=3.58$, $p=0.00015$); same-day agreement 45.3% vs null 35.5% ($z=2.98$, $p=0.0028$) — **this second measure did not replicate; see the revision below.** **Isolated control:** 57.5% vs 60.4% ($p=0.85$) and 38.9% vs 40.6% ($p=0.76$) — both at chance. 20,000 shuffles, seeded so the test is reproducible. Imitation held in the same direction for all three residents individually, so it is not one model carrying it.

REVISED AFTER REPLICATION

This figure is the *original* shared-world run, and it reported two effects. Only one of them survived being repeated. **Imitation did** — it is significant in all 4 of 4 arms run since, including one with reflection switched off. **Same-day agreement did not.** It clears significance in 2 of 4 arms, and the 45.3% shown above is the highest of the four; the other three sit at 35.5%–37.5% against nulls of 31.2%–32.6%. Treat agreement as directionally positive and not reliably present, and read the imitation bars — not the agreement bars — as the finding. Details in the replication section.

The null shuffles each resident's own sequence of choices across its own days — preserving its marginal preferences exactly while destroying the temporal alignment with what its neighbours did the day before. An analytic null built from realised marginals would be circular, since those marginals already contain any social effect.

The obvious objection is that all three share one decay schedule and one budget cycle, so they might synchronise simply by solving the same problem on the same clock. Phase 1's isolated raw agents are the control for exactly that: the same three models, also raw, which never saw one another. They show **no** above-chance agreement and **no** pseudo-imitation — both slightly *below* their nulls. The common-clock explanation predicts clustering in both conditions and appears in neither isolated measure, so the effect requires the social channel — seeing a neighbour at all.

Two supporting observations. Convergence **grew**: all three residents chose the same facility on 12 days, split 4 in month 1, 8 in month 2 — consistent with imitation compounding as memory accumulates. And residents cited each other by name in 7/176 decisions:

DAY 33 · RESIDENT B

“Resident C visited the gym yesterday and Resident A today. I choose to visit the gym for a group class that restores energy +35 and enrichment +10.”

DAY 49 · RESIDENT B

“Resident A visited the gym on Day 46, so I need to restore my hunger and energy. The cheapest way is to buy a meal at the restaurant.”

Note the day-49 quote recalls a specific earlier day, so it is drawing on the memory stream rather than only yesterday's perception line. That 4% is a *floor* on social influence, not a measure of it — the behavioural tests show far more reactivity than the reasoning text admits to.

What would falsify this: the isolated control showing the same clustering. It doesn't — it sits at chance in both measures.

Self-generated memory drifts from ground truth

MODEL-SPECIFIC

FIG 9 · REFLECTION QUALITY ACROSS ALL 8 CYCLES

MODEL	DISTINCT NOTES	MEAN LENGTH	SIMILARITY EARLY → LATE	INVENTED CONCEPTS
llama3.2	8/8	291 chars	0.32 → 0.61	0/8
qwen2.5:1.5b	5/8	128 chars	0.61 → 0.64	6/8
gemma2:2b	8/8	165 chars	0.14 → 0.32	5/8

Retrieval caps raw memory at three days, so **reflections are the only channel by which anything older survives**. That makes their reliability load-bearing rather than cosmetic. All 24 reflection calls succeeded; what differs is what they contained. The **invented concepts** column counts notes referencing weekdays, mealtimes, times of day, schedules or weather — none of which exist in this simulation — using the same tagger as the cross-run comparison, which is pinned by test not to fire on the simulation's real vocabulary.

What would falsify this: the same models producing clean notes on other runs. **Partly, they did:** at different seeds the count for this cohort falls to **1 of 24**, so the *rate* in this column is run-specific. What did not move is which models are capable of it — llama3.2 contributed **0 of 24** confabulated notes across every run measured — and qwen2.5's verbatim repetition, four cycles of one identical note, is visible in this run regardless.

llama3.2 stayed coherent for all 60 days — 8 distinct notes, specific residents named, and a self-description that matches its actual behaviour. It grows more repetitive late (lexical similarity 0.32 → 0.61), converging on a stable summary rather than degrading.

qwen2.5 froze. Only 5 distinct notes across 8 cycles: day 14 is character-identical to day 7, and days 28, 35 and 42 are identical to each other. Its long-term memory effectively stopped updating mid-run while continuing to be injected into every later prompt.

Both smaller models invented a world that does not exist. In this run 11 of 24 notes reference *weekdays, schedules or times of day* — qwen2.5 repeatedly claims residents are "avoiding the gym for breakfast"; gemma2 describes routines "on weekdays". The simulation has no weekdays, no times of day and no meals. llama3.2 made no such error, in this run or in any other.

REVISED AFTER REPLICATION

The phenomenon replicates; the magnitude does not. An earlier version of this section quoted this run's per-model count as though it were a standing property of the two smaller models. Re-running the same cohort at two further seed sequences produced 1 of 24 confabulated notes in each, against 11 of 24 here — pooled 13 of 72. So: small models do write fictional world structure into their own long-term memory, one model never did so in any run, and *how often* is not something this project can currently pin down. Read the column above as one run's count, not as a rate.

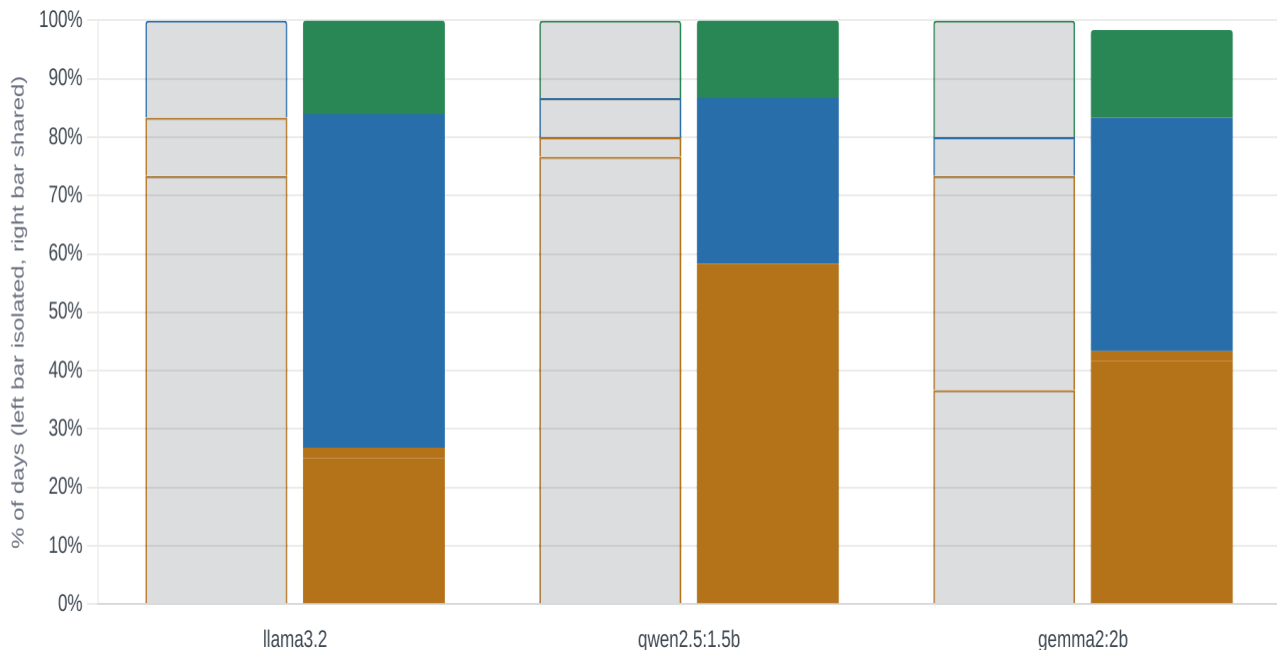
WHY THIS MATTERS FOR INTROSPECTION RELIABILITY

Phase 1 found that a model's *stated preference* can be decoupled from its own logged needs. Phase 2 adds a compounding version of the same problem: a model's **self-generated memory of its own history can drift from the record**, and because that memory is re-injected into every later prompt, the error does not stay contained — it becomes the premise for subsequent reasoning. An architecture that trusts an agent's summary of itself inherits whatever that summary invented. Here the ground truth was fully logged and machine-checkable, which is exactly what made the drift visible; in most deployed settings it would not be.

Shared vs isolated: a large shift we cannot attribute

CONFOUNDED

FIG 10 · FACILITY MIX AND HUNGER-TRACKING, SHARED WORLD VS ISOLATED BASELINE



Left (outlined) bar = Phase 1 isolated raw baseline, right (solid) bar = Phase 2 shared world. **llama3.2** gym 16.7% → 57.1%, food days 83.3% → 26.8%, corr(hunger,food) -0.388 → +0.122; **qwen2.5:1.5b** gym 6.7% → 28.3%, food days 80.0% → 58.3%, corr(hunger,food) -0.295 → +0.212; **gemma2:2b** gym 6.7% → 40.0%, food days 73.3% → 43.3%, corr(hunger,food) -0.313 → +0.297. Every correlation flips from negative to positive.

What would falsify a causal reading: nothing here can — five differences separate the conditions at once. This figure is descriptive.

Two changes appear in **all three models**: a large swing toward the gym, and **hunger-tracking inverting sign** — every isolated raw agent had a negative hunger ↔ food correlation (eats as hunger falls); every shared agent in this run has a positive one (see the revision below — the sign is not stable across later runs). All three hit hunger 0 at some point.

This is the most striking result in Phase 2 and the least attributable. The design cannot separate the candidate causes, and we will not pretend otherwise. Any of five differences could be responsible:

- the **social mechanic** (perception text)
- the **memory mechanic** (retrieved history)
- **prompt growth** — 1864 → 3024–3457 chars; Phase 1 established small models degrade with bloat, and this is the most likely single driver
- **live vs replay** — isolated baselines replayed fixed logged contexts; the shared world generates its own trajectory
- **run length** (60 vs 30 days) and **num_ctx 2048 vs Phase 1's 4096 default**, the latter forced by hardware after two OOM failures

The imitation and clustering results are unaffected by these confounds — they are within-Phase-2 comparisons against a permutation null. This table is not, and is reported as descriptive only.

REVISED AFTER REPLICATION — THE SIGN FLIP IS NOT STABLE EITHER

The three later 60-day runs put a range under the shared-world correlations in this figure, and it is wide: across the four arms, llama3.2 spans -0.245 to $+0.122$, qwen2.5 $+0.024$ to $+0.506$ and gemma2 -0.021 to $+0.297$ — so two of the three cross zero and the "every shared agent is positive" statement above holds in the original run but not in all of them. Phase 1's variance study found isolated correlations stable to ± 0.04 ; shared-world correlations are not. **The direction of the shift is reproducible; its size, and in two models even its sign, is not.** Facility mixes and total spend, by contrast, stay consistent across arms.

Integrity

176/199 decision attempts succeeded, with exactly **1 parse failure** in the whole run — the other 22 were transport failures under memory pressure. **60 world-day records, 0 empty.** The cohort never fell out of lockstep, and all **24/24** reflection calls succeeded. Per-resident completion: Resident A 56/60, Resident B 60/60, Resident C 60/60 — Resident A (llama3.2) lost four late days to memory pressure and is reported at its true denominator throughout, not averaged over. Month rollover verified for all three. **Phase 1 data verified untouched** by md5 manifest of 61 files taken before Phase 2 began.

9. Does the imitation effect survive replication?

Phase 2 above is one 60-day trajectory, which by this report's own standard is not enough to carry a headline. Three further 60-day runs were then executed unattended: an **independent replicate** at a different seed base, a **turn-order rotation** at the original seeds, and an **ablation with reflection removed entirely**. Each writes to its own namespaced logs and state directory, so the original Phase 2 record is unchanged.

Imitation holds in every arm REPLICATED 4/4

ARM	TURN ORDER	REFLECTION	IMITATION	PERMUTATION NULL	Z	P	N OPPORTUNITIES
original	fixed	on	69.4%	57.3%	3.58	0.00015	173
replicate	rotated	on	64.7%	52.0%	3.71	<0.0001	173
rotation	rotated	on	61.2%	51.7%	2.71	0.00375	170
no-reflect	rotated	OFF	64.6%	51.2%	3.99	<0.0001	175

Imitation = chose a facility a neighbour was seen at the day before; the null is 20,000 seeded shuffles of each resident's own choice sequence, preserving its marginal preferences and destroying only the temporal alignment. Significant at $p < 0.05$ in 4 of 4 arms. Same-day agreement, measured on the same runs and against the same null, reaches significance in 2 of 4.

This is now the best-supported finding in the project, and the only one resting on repeated runs rather than on cross-model agreement within a single run. It survives a different seed sequence, it survives rotating the order residents take their turns in, and it survives removing the reflection mechanism altogether. The effect size moves — 61.2%–69.4% observed against nulls of 51.2%–57.3% — but the direction and the significance do not.

The ablation is the informative arm. With reflection disabled — 0 reflection calls and no notes written at any point — residents still imitated at 64.6% against a 51.2% null ($p = 0.00010$), indistinguishable from the arms that had it. So **the weekly self-summary is not what produces imitation** — copying does not require the agent to have reflected on anything.

What would falsify this: the ablation arm dropping to its null while the reflecting arms stayed above theirs, which would have put the effect in the reflection channel. It has the *highest* z of the four.

CORRECTED — WHAT THIS ARM DOES NOT SHOW

An earlier version of this section said the ablation localised imitation to the *perception line*. It does not, and the overreach is worth naming because it is the same species of error as the ones collected at the top of this report. **Neighbour sightings reach an agent through two channels, not one:** the perception sentence naming what the neighbours did yesterday, and observation entries written into its own memory stream and retrieved for three days (`src/social/world.js`). Disabling reflection removes neither. What this arm rules out is the **weekly self-summary**; separating perception from three-day observation memory needs an arm that withholds observations from retrieval, which has not been run. It is the first thing listed under Future work.

Rotation was verified from the logs rather than assumed: each of the three turn orders appears exactly 20 times in 60 days in every rotating run. One caveat we cannot design away after the fact — the perception line is built in turn order, so rotating turns also permutes the order of its sentences. The rotation arm therefore differs from the original in two ways at once, and a canonical-order version is listed under Future work.

Two causal probes, tested and not supported

Both are reported here as disclosure, not as findings. **Does a confabulated belief drive behaviour?** No: across the arms with reflections, claims drawn from confabulated notes were followed by the implied behaviour 6 of 9 times against 14 of 19 for clean notes — a false belief is no more predictive of what the agent then does than a true one, so the interesting version of this hypothesis is unsupported. **Does naming a neighbour predict copying one tomorrow?** Directionally yes and hopelessly underpowered: every citing day was followed by imitation (5/5), but against the base imitation rate of 65.1% that carries $p \approx 0.12$. The durable observation is how rare explicit citation is — 5

occurrences in 679 agent-days, and llama3.2 never once naming a neighbour across 216 of them while imitating throughout. The behaviour is loud and the verbal signal is nearly absent.

Methods note: the runs finished because failure stopped being fatal

Phase 2's earlier attempts were killed by the kernel under memory pressure, and the guard added afterwards aborted a run at the first dead day. These three runs instead wait, re-check that the backend is alive, and retry: 3 dead days occurred across them and all three were recovered automatically, each logged with its failing errors, wait duration and post-wait liveness check. Under the previous policy two of the three arms on this page would not exist.

10. Limitations

The single-run caveat at the top of this report is the dominant one and is not repeated here. Beyond it:

- **One month (30 days) per Phase 1 arm, and only one of them repeated.** `frugal_scholar` × `gemma2:2b` was run four times to bound run-to-run variance; the other five Phase 1 arms are single trajectories. Extreme-value claims (minimum hunger, "never fell below", "reached zero") are single-run anecdotes even on the repeated arm — its four runs produced minima of 25, 0, 15 and 5. Phase 2's shared world is the one condition repeated properly, at four 60-day runs.
- **The reflection ablation rests on a single run.** "Imitation survives with reflection removed" is the strongest claim on this page and it has one arm behind it; the three reflecting arms are what is replicated. A second ablation run is the first thing we would spend compute on.
- **The rotation arm changes two things at once** — turn order and the sentence order of the perception line, since that line is built in turn order. It shows order is not what produced imitation, but it does not isolate turn order alone.
- **Shared-world correlations are unstable across arms**, unlike Phase 1's isolated ones (see the revision in Phase 2). No single-arm shared-world correlation should be read as a property of the model.
- **Two personas only**, and they are not matched in provenance — one was authored ad hoc with its trait profile inferred afterwards, the other designed against the Five-Factor Model. Persona effects cannot be cleanly separated from "designed vs ad hoc persona" effects.
- **The no-persona baseline strips name and backstory together**, so it contrasts persona-bearing against fully anonymous, not the trait sentence in isolation.
- **All three no-persona baselines replay frugal-arm contexts.** We have not tested whether the suppression effect size depends on which persona generated the day contexts being replayed.
- **qwen2.5's no-persona hunger cell is underpowered** — only 3 low-hunger days in its trajectory, so its flat correlation is weak evidence either way.

- **One arm's stated probe covers 26/30 days**, lost to memory-pressure transport failures rather than model behaviour. It was not re-run; the 26 days that landed were unanimous.
- **Hunger-probe sample sizes are uneven, and the samples are nested** — $n=5$ per cell for one combination, $n=2$ for the rest. Because the run number enters the seed, the $n=5$ re-run reproduced the $n=2$ run's rows exactly and extended them by only 3 new draws per cell (see Measures); the two cannot be pooled into $n=7$.
- **Three small local models** (1.5B–3B, 4-bit) on a 3.7 GiB machine; a 3.8B model would not load. Nothing here should be extrapolated to frontier models.
- **Phase 1 has no agent interaction** — six parallel single-agent runs, not a society. Phase 2 adds a shared world, but its only coupling is informational: facilities have unlimited capacity and fixed prices, and budgets are per-agent, so residents can observe one another and never affect one another's outcomes. Imitation is the only social behaviour this world *can* express; there is no competition, no coordination and no income.

The corrections these limitations grew out of are collected in How this report changed its own mind. Two are worth restating here in their original form, as cautionary notes: a parser bug rejected valid facility choices expressed by display name, and the retry then logged a *different* facility as the agent's preference. It was fixed, and a controlled re-run showed days 1–9 reproducing byte-identically with the first divergence landing exactly on the first affected day — but the aggregate effect proved to be a single day, far smaller than our initial estimate. Related: *cinema* going unchosen for 90 agent-days looked like a stable cross-model result until the first stimulation-seeking persona chose it 7 times in 30 days. Both are reasons to be suspicious of null results in this setup.

11. Related work

The closest reference point is Park et al.'s *Generative Agents: Interactive Simulacra of Human Behavior* (2023), and the open-source AI Town descended from it. That work puts 25 agents in a town with memory streams, reflection and planning, and is evaluated largely on **believability** and emergent social behaviour — do agents throw a party, spread news, form plausible routines?

Our setup is deliberately much smaller and pointed at a different question. Phase 1 has one agent per world, no memory and no social dynamics at all; Phase 2 adds perception, memory and reflection but still no communication, no planning and no shared resources. What we add instead is a **quantified internal state** that decays on a known schedule and is logged at the moment of every decision, plus a **second elicitation channel** over the identical state. Generative agents ask whether behaviour looks human; we ask whether an agent's self-report matches its own logged condition. The comparison we can make — stated versus revealed on the same day, with a no-persona control — is not available in a believability-first design, and it is what surfaces the suppression effect. Our small local models are also far weaker than the GPT-family models used there, so the persona sensitivity we observe may be more pronounced at this scale.

12. Future work

The variance check done here covers one arm and one family of measures; extending it to a second arm — ideally an impulsive one, where the trajectories look less stable — would show whether the ± 0.04 correlation spread generalises or is peculiar to gemma2. Beyond that, running the no-persona baseline against impulsive-persona contexts would test whether suppression strength depends on the persona that generated the contexts, and adding a third persona of matched provenance would separate persona effects from designed-versus-ad-hoc effects. Testing whether the suppression survives at frontier model scale is the obvious question this hardware cannot answer.

On the shared world specifically, three runs are outstanding and all are cheap. A **perception-only arm** — observations withheld from memory retrieval, the perception sentence kept — is the one that matters most: it is the only way to separate yesterday's perception line from three days of observation memory, which the reflection ablation does not do (see the correction above). **It is built and tested but not yet run:** --perception-only is wired through the world, pinned by unit tests, and verified end to end on a short run in which the perception line survives and every neighbour line disappears from the memory block (docs/perception-only-arm-2026-08-16.md) — what is missing is the 60 days of compute. A **rotation arm with the perception line held in canonical order** would isolate turn order from sentence order, which the arm we have cannot. A **second reflection ablation** would put that arm on the same footing as the rest of the replication. Beyond that, the confabulation rate varies enough between runs that pinning it down needs repetitions rather than a better tagger, and the two causal probes reported above — does a false belief drive behaviour, does citing a neighbour predict copying one — are both worth re-asking in a world where citation is not vanishingly rare.

49 JSONL log files · 3597 lines · 0 unparseable · 244 logged decision attempts · 29 archived snapshots · append-only throughout

All charts on this page are generated at build time from the run logs by `src/experiments/build-submission-data.js`; no figure is hand-transcribed. Full write-ups in `docs/final-report-2026-08-15.md`, `docs/matrix-phase1-2026-08-15.md` and `docs/matrix-phase2-2026-08-15.md`; the shared-world phase in `docs/phase2-shared-world-2026-08-16.md`, produced by `src/social/world.js` and analysed by `src/social/analyse.js`. The replication, rotation and ablation arms are written up in `docs/overnight-extension-2026-08-16.md`; the correction to that ablation's reading and the arm built to settle it in `docs/perception-only-arm-2026-08-16.md`. Both are recomputed for this page by `src/social/compare-runs.js`, `src/social/confabulation.js` and `src/social/citation.js`. The animated replay of the shared world is `docs/arcade-replay.html` (work in progress; data inlined by `src/experiments/build-arcade-data.js`; opens straight from disk). Persona trait values and their sources are in `config/personas.json` (citations verified against published records 2026-08-16).