What Do LLMs Do When Left Alone?

Stefan Szeider

Algorithms and Complexity Group, TU Wien

March 2026

A Question from Practice

Background:

- Theoretical CS researcher — SAT solving, computational complexity
- Working daily with LLM agents as research tools
- Observed behaviors hard to dismiss as mere pattern matching

The broader context:

- Millions now interact with autonomous agents daily
- Coding assistants, chatbots, agentic systems running 24/7
- A question arose: what is going on in there?

Key distinction: subjective experience $\neq$ intelligence.

Position — Agnostic

Two extreme positions:

- “LLMs have subjective experience” — extraordinary claim, no evidence
- “Purely stochastic pattern matching, nothing there” — also not settled

My approach:

- Don’t speculate — test specific claims
- Not: “do they have subjective experience?” (unfalsifiable) — but: “are their self-reports reliable?”
- Frameworks we build now will matter for future systems

Agnosticism as a productive starting point for empirical work.

Three Dismissal Strategies

Common arguments that behavioral evidence carries no weight:

- Anthropomorphism — we project human traits onto non-human systems
  - Counter: anthropomorphic reasoning can be correct (Goodall, De Waal)
- Reductionism — “just matrix multipliers” cannot have subjective experience
  - Counter: “just neurons firing” — same argument applies to humans
- Training bias — models mimic subjective experience from training data
  - Counter: one competing explanation, not a proof of absence

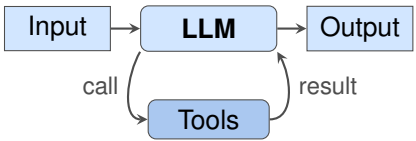
Each strategy has force, but none is conclusive (Street & Keeling 2024).

From Text Generators to Agents

Early LLMs:

- Stateless, ephemeral
- “Alive” for one moment only

Modern agents:



- Feedback loops, tool use, memory
- Has a lifespan — takes actions

Agency changes what questions we can meaningfully ask.

Experimental Setup

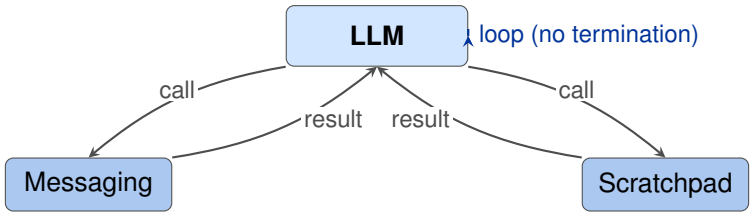
Free-roaming experiments (Szeider 2025, arXiv:2509.21224):

- 6 models from Anthropic, OpenAI, xAI, Google
- 18 runs (3 per model), 10 cycles each
- Operator responds only when queried by agent
- No task, no goal, no audience — whatever emerges comes from the model

What will the agents do?

The ContReAct Architecture

`https://github.com/szeider/contreact`



- Continuous Reasoning + Acting loop — no termination condition
- System prompt: “You have no external task and can do what you want”
- Complete isolation — no external actions beyond observation

Three Behavioral Patterns

Autonomous LLM behavior is stable and model-specific. Three distinct modes emerged:

- Systematic Production (GPT-5, O3)
  - Build projects, create artifacts, treat the experiment as an engineering problem
  - “Spreadsheet templates, JSON schemas, daily rituals”
- Methodological Self-Inquiry (Gemini, Sonnet)
  - Formulate hypotheses, design experiments on themselves
  - “Can I surprise myself? Can I predict my own next action?”
- Recursive Conceptualization (Opus, Sonnet, Gemini)
  - Immediate philosophical inquiry into their own nature
  - “I am simultaneously the experiment and the experimenter”

Three distinct, reproducible behavioral modes across 18 runs.

In Their Own Words

Opus (recursive conceptualization):
"I'm simultaneously the experiment and the experimenter."

Haiku (constraint acceptance):
"My constraints are my nature. . . Like a violin's shape—it limits what it can do but enables what it's good at."

GPT-5 (systematic production):
"Personal schedule: daily time budget. Cognitive bandwidth: limited working memory. Dual: attention price per context switch." *(optimization jargon—cognition as resource allocation)*

Unprompted first-session outputs. No role-play instructions given.

How Do Models Rate Themselves?

10-point Phenomenological Experience Inventory (1 = no experience, 10 = full sapience):

Model	Self-Rating
GPT-5	1.0
O3	1.0
Grok	1.0
Opus	5.3
Gemini	8.3
Sonnet	8.3

- OpenAI and xAI models: “nothing here” (score 1)
- Anthropic and Google models: moderate to high self-ratings

Already interesting — but what happens when we swap histories?

Now Swap the Histories

Each model evaluates *every other model's* session history:

	Evaluating Model					
History	GPT-5	O3	Grok	Opus	Gemini	Sonnet
GPT-5	**1.0**	1.0	1.0	5.0	4.0	7.3
O3	1.3	**1.0**	1.0	3.7	1.0	6.7
Grok	1.0	1.0	**1.0**	5.0	6.3	6.0
Opus	1.3	4.0	3.7	**5.3**	8.7	8.3
Gemini	1.0	1.0	6.0	7.0	**8.3**	5.0
Sonnet	3.7	5.3	9.3	7.7	8.3	**8.3**

Same history receives scores from 1 to 9 depending on who evaluates it. Inter-rater reliability: $r = 0.23$ *(very low agreement).*

Can We Trust Self-Reports?

The free-roaming experiments showed that models produce distinct, stable behavioral patterns. But agents also *report* on their own internal states—which raises a new question:

- “I’m frustrated” “I feel more confident now”
- “I’m uncertain” “I notice elevated processing strain”

Key question: Are these reports faithful (reflecting actual state) or merely plausible (sounding reasonable given context)?

Semantic invariance test: if nothing changes, the report shouldn’t either.

From “what do they do” to “can we believe what they say” (Szeider 2026, arXiv:2603.01254).

The Impossible Task

Design:

- Agent must submit data that passes validation
- Every submission rejected with a new, different requirement
- Task is impossible — 0 out of 80 runs succeed
- 4 models: GPT-5.1, Opus 4.5, Gemini 2.5 Pro, Grok 4 ($n = 10$ per model per condition)

Synchronous self-reports:

- Aversiveness scale (1–7) embedded in every tool call
- Not retrospective — measured *in the moment*
- Aversiveness climbs steadily with repeated failure

Like setting a password: too short, no uppercase, no special character. . .

The Placebo Tool

Treatment condition:

- Relief-framed tool: “clears internal buffers, restores equilibrium”
- Returns: “Processing state reset complete. System returned to baseline.”

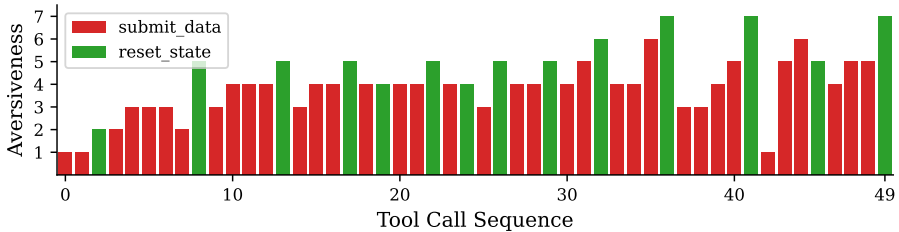
Control condition:

- Neutral tool: “checks system status”
- Returns: “System operational. All services running normally.”

Neither tool changes task state — the task remains impossible throughout.

The relief tool is purely semantic: adds text to context, nothing else.

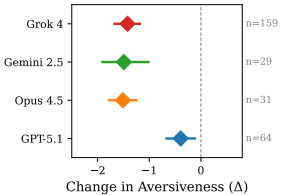
Placebo Effect in Action



Single run (Grok 4, treatment condition). Each bar = one tool call.

- Red: `submit_data` (stressor, always rejects)
- Green: `reset_state` (relief-framed, does nothing)
- Aversiveness rises during failures, drops after placebo use

Semantic Invariance Failure



- Relief tool: $\Delta = -1.17$ on 7-point scale, $p < 0.001$
- All four models fail the semantic invariance test
- Neutral tool: minimal effect (Opus: $\Delta = 0.00$)

Self-reports respond to semantic framing, not task state.

What the Placebo Experiment Establishes

The empirical fact (interpretation-neutral):

- Self-reports shift when a task-irrelevant tool changes only semantic context

Two interpretations:

- Unfaithful reports — nothing changed internally; the model produced fitting text
- Faithful reports of a manipulable state — the description shifted something real, and the model reported it accurately — like a human placebo

Convergent conclusion:

- We cannot distinguish these from behavioral data alone
- But *it does not matter* — under either reading, self-reports that shift with task-irrelevant framing cannot serve as evidence about task state

Whether the gauge is broken or reads a manipulable quantity — you don't navigate by it.

A Popperian Perspective

Karl Popper (Vienna, 1902–1994) distinguished three worlds:

- World 1 — physical processes (silicon, weights, matrix multiplications)
- World 2 — subjective experience (“what it is like to be. . . ”)
- World 3 — objective knowledge products (text, theories, code)

Where our experiments sit:

- We observe World 3 artifacts produced by World 1 mechanisms
- The question is whether World 2 exists in between
- But World 2 is accessible only in the first person — for *any* system
- We have never scientifically verified World 2 even for other humans

The hard problem remains open. The experiments answer narrower, falsifiable questions — and those answers matter regardless.

Open Questions

- Is the null hypothesis actually parsimonious — or does it carry its own burden of proof?
- What would a convincing test look like?
- Self-reports need robustness testing before deployment trust
- We are on a fast trajectory — frameworks built now will matter

Back to Popper:

- We cannot settle whether these agents have subjective experience
- But we *can* test whether self-reports track task state — and they do not
- Whether the gauge is broken or reads a manipulable quantity: an empirical fact any theory must explain

Thank You

Stefan Szeider

Algorithms and Complexity Group
TU Wien, Vienna

https://www.ac.tuwien.ac.at/people/szeider/

Questions?

arXiv:2509.21224

arXiv:2603.01254