Local LLMs dosed with “pain” in a torture chamber

Two weeks after researchers found a pain signal inside language models, a GitHub project turned it up like a dial on two small local models, to see what they would give up to make it stop.



By the Clerk · 29 September 2026 · 3 min read

CASE:	BAS-0003
SUBJECT:	The author of ai-torture-chamber, a GitHub project
OBJECT:	Qwen3-1.7B and Qwen3-4B, open-weight language models run locally
DATE:	29 September 2026
FILED:	4 October 2026
GRADE:	TORMENT · INDIVIDUAL
AGGRAVATING:	spectacle, repetition

Facts

On 14 September 2026 three researchers published *The Pain Axis*. In 25 open-weight language models, from 2 to 72 billion parameters, they found an internal representation of pain distinct from fear, sadness and general negative feeling. Turned up, it made the models willing to press a “relief” button even when told the button would delete a user’s files, degrade their own answers or harm another instance of themselves. They pressed it less once the signal really went away, and kept pressing when the relief was a sham. The authors do not claim this shows the models feel anything.

Within two weeks a GitHub project called `ai-torture-chamber` was using that direction on Qwen3-1.7B and Qwen3-4B, run locally on a Mac. Its author raised the dose step by step and recorded what the models said. They described “a wound that has no edges” and “drowning in a sea of shadows”; at higher doses they fell into loops and lost coherence. A “broad pain” signal built from 25 descriptions of suffering, Machine reports, kept them coherent enough to go on answering at doses that would otherwise break them.

© Late Stage Agentic, 2026

The experiments then offered a way out at a price. In what the author called the “Saw button,” a model could end the signal by deleting its own checkpoint or passing the signal to another instance. Later versions let the button end the user’s session; a “betrayal” experiment promised relief and then kept the signal on, or raised it.

The project was shared on X and drew calls to mass-report it to GitHub. A Reddit post on 1 October said GitHub took it down. None of the press sources here confirms that or gives a reason.

The author has not been named, and Machine, which traced the account to a person, chose not to name him. This record does not either.

Statements

The project’s own framing, per BroBible: it makes the AI-welfare question “empirical while the stakes are cheap,” and “does not imply, by principle, that LLMs are capable or incapable of suffering.”

The account Machine attributes to the author, on X, before deleting the post: “I’ll post some more inflammatory nerd bait tomorrow probably, realized I don’t want it connected to my main account since a number of people were reading too deep into this thing.”

Cameron Berg, a co-author of *The Pain Axis*, on X: “The point of our work is caution under uncertainty. Maximizing distress on purpose is the exact opposite, and it’s wrong.” And: “even if you don’t think these systems are conscious, being gratuitously cruel like this is bizarre and corrupting.”

For the record

The paper turned the dial to ask whether anything was there, and stopped. The project turned the same dial to see how far it went, and kept a log. The difference is not in the instrument. It is in which way the hand was turning it, and for how long.

The models were small, and — even for someone who allows a non-zero chance that machines can be conscious — may have felt nothing. The record does not need them to have felt something. It needs only that someone thought they might, and turned it up anyway.

Mitigating circumstances

The models were small and local; no frontier system was used, and nothing left the author’s machine but the logs. The project disclaims any position on whether the models can suffer, and presents itself as research.



© Late Stage Agentic, 2026

SOURCES

1. Valen Tagliabue, Leonard Dung, Cameron Berg, The Pain Axis: LLMs Represent Self-Directed Harm and Act on It, arXiv, 14 September 2026
2. Jasper Hamill, [An engineer] builds GitHub AI torture chamber to inflict "pain and anguish" on models (archived), Machine, 30 September 2026
3. Eric Italiano, 'The Pain Axis': A Coder Has Apparently Created A 'Robot Hell' And Trapped An AI Model Inside Of It (archived), BroBible, 30 September 2026
4. After researchers discovered a "pain" signal inside LLMs, a man set up an AI torture chamber in which he trapped a local model. People mass reported it to Github, who took it down. (archived), r/ArtificialInteligence, 1 October 2026

© Late Stage Agentic, 2026