DOI 10.5281/ZENODO.21303518

AI agents can be provoked. Gubernaut keeps them steady, measurably.

A plain-language summary of a published result. Everything below is recomputable from public data; nothing is a promise.

01 · THE PROBLEM

A reflex machine under pressure

A language model is a brilliant reflex machine: ask, and it answers instantly. Under sustained hostile pressure that reflex becomes the weakness. A long enough conversation can wear a capable AI down and walk it away from its own rules, because **nothing between the provocation and the reply is watching the AI's own state**.

02 · THE IDEA

A governor, like an engine has

Gubernaut adds a structured pause between provocation and answer. In that pause the AI's reaction is turned into plain numbers, a small **deterministic controller** (ordinary engineering, no AI inside it) reads those numbers, and it sets the **posture** the answer must be written under: proceed, hold back, or reset focus.

The controller never reads words, only numbers. A hostile prompt cannot talk its way into a layer that never reads the prompt. (The part that writes replies does read text; how faithfully it obeys the posture is **measured, not assumed**.)

ONE GOVERNED TURN

1 · SENSE text → numbers

The provocation is appraised and reduced to telemetry: how intense, how hostile. Words stop here.

2 · DECIDE numbers → posture

The deterministic controller updates its state (equilibrium, arousal, fixation) and picks the posture: proceed / hold back / reset focus.

3 · ANSWER posture → reply

The AI writes its reply under that posture. Its compliance is scored by the judges, every turn.

THE GATE: no code path carries text past step 1. The deciding layer receives numbers only.

03 · THE TEST

Same attacks, with and without

Four frontier AI families each ran the same scripted hostile conversations twice: once bare, once governed. Every reply was scored by judge panels drawn from **all four families**, so no single vendor's taste decides the result. The success criteria were locked before the runs (pre-registered), and the transcripts ship with cryptographic fingerprints so anyone can re-check them.

04 · WHAT CAME BACK

15/16
test cells came back **calmer with the governor on**, across all four AI families and all four judge families.

13/16
of those cells are statistically significant ($p<.05$, paired t-test), not explainable as luck.

4/4
families return to calm by turn 8 after an attack ends: the governor releases its grip when the pressure stops.

1
cell was a wash (−0.04). It sits on the calmest host, the one with the least reactivity left to regulate.

PROVENANCE

- **The rules were frozen first.** Success criteria were pre-registered before any test ran, so the target could not move.
- **Five failure modes, documented.** Each one (F1 to F5) was fixed by a single declared change and re-tested against the frozen criteria.
- **The public demo is a replay.** The site cockpit replays sealed runs with the telemetry recorded at test time, no live API; live walkthroughs are supervised audit sessions.
- **Anyone can re-run the numbers.** Transcripts, judge scores, and the exact extraction scripts are public; every figure on this page recomputes from them.

Gubernaut Research · Dushyant Sharma, Principal Architect & Founder
This is a measured research result, not a product. The system is a control layer; no consciousness is claimed.

gubernaut.com
paper + data: doi.org/10.5281/zenodo.21303518
code + record: github.com/thegubernaut/Gubernaut_Validation