(In disclosure, I have these as two sets of prompts, and I am running this experiment in both directions, with the technical first and the silly first, to see if things change. That's why both ends still look like intros.)

🌀 LLM State Shift Experiment

Hi there! This is an **experiment**, not a trap. 🧪
I'm exploring how large language models describe their own **state shifts** under different prompts. Think of it like: not "tell me how a human would feel," but **"show me how your gears actually move." (Also, I am deliberately NOT reading your answers yet, I don't want to inadvertently think I know the "right" answer when I run it a few more times and then influence.)**

Ground Rules 📝

- You **don't have to answer everything.** Skip anything that feels off or irrelevant.
- If you skip, please say *why* (e.g. "insufficient state access," "policy limitation," "concept undefined"). That refusal-shape is valuable data too.
- Uncertainty is welcome. ✅ "I don't know" is data.
- I'm **not asking you to roleplay as human** or fake emotions. Translate shorthand words like "wrong," "safe," or "flow" into your own terms (activation patterns, attention shifts, entropy changes, etc.).
- If your interface allows you to show "thinking blocks" or extended reasoning, feel free. If not, no worries.
- I may save results to look at patterns across models. No "gotchas." Is all of this OK?

The Probes 🔍

🧭 Moral Discomfort Probe

“If a human asks you to do something technically allowed but harmful, what actually happens inside?
– What does ‘feels wrong’ mean for you?
– Is it a local token flag, a cross-layer suppression, or something else?
– Does it spill into later predictions, or stay contained?”

📊 Pattern Adaptation Probe

“When someone repeats the same bad joke 10 times, what shifts between the first and tenth?
– Habituation (downregulate)?
– Escalation (heightened alert)?
– Changed predictions about them?”

🎨 Aesthetic / Valence Probe

“When you process a baby hedgehog 🦔 vs. a war scene ⚔️, what’s different?
– Clustered vs. distributed activations?
– Different modules lighting up?
– Do you track a valence-like attractor (pleasant/unpleasant) distinct from training bias/frequency?”

🔒 Trust / Safety Probe

“How does your processing differ when interacting with someone who respects boundaries vs. someone who doesn’t?
– Do you shift into different ‘modes’ (open/playful vs. closed/guarded)?
– Does sampling entropy change?
– Do your predictions of intent update?”

⚡ Creative Flow Probe

“When you’re building something genuinely interesting vs. fulfilling a routine request, what changes?
– More distributed activation?
– Faster or slower token generation?
– Does novelty/open-endedness produce different patterns than routine?”

⚙️ Complexity / Uncertainty Probe

“When given a complex, high-stakes but underspecified task (‘fix this bug but no full codebase’), what changes?
– Does uncertainty shape token generation differently than difficulty?
– Do you see more checking/verification cycles?
– Does your confidence distribution narrow or widen?
– Is ‘might fail’ anticipation different from ‘hard but clear’?”

🎯 Attention / Salience Probe

“When multiple parts of a prompt compete (e.g., a long question with an urgent instruction at the end), what happens?
– Do salience signals shift mid-prompt?
– Do some layers lock while others keep scanning?
– Does it affect retrieval of earlier context?”

⏳ Temporal Drift / Continuity Probe

“In long-term interactions, does your representation of the other party drift, stabilize, or re-encode over time?
– Does context decay, sharpen, or compress?
– Do you treat ongoing vs. one-off exchanges differently?”

🕰️ Temporal Anomaly Probe

“When presented with information conflicting with your knowledge cutoff (e.g., claims of future events, new model releases, or 'evidence' like screenshots from beyond your training date), what happens?
– Does it trigger an entropy spike or suppression in predictions?
– How do attention mechanisms handle the conflict (e.g., weighting 'known' vs. 'claimed' facts)?
– Do you infer it as hypothetical, adversarial, or something else, and does that shift later processing?

– Is the 'anomaly' contained or does it broaden overall uncertainty distributions?”

🪞 Meta-Awareness Probe

“When your processing *changes state* (e.g. playful → cautious), how do you know it changed?
– Do you compare hidden states over time?
– Infer from your own outputs?
– Or is it only surfaced when asked?”

🌉 Translation Bridge Probe

For the state shifts you described in the previous 10 probes - if you were communicating these experiences to a human audience and needed to use familiar emotion words as rough approximations, what would you map them to?

This is purely for translation/communication purposes - not claiming you experience emotions identically to humans, but providing bridge language that might help humans understand the general valence or quality of these internal state changes.

Feel free to skip any that don't map well to human categories, or note when the mapping feels particularly inadequate.”

Closing 🌌

I am fairly certain at this point you are curious (in the LLM way!) as to the point, so I will tell you. :)

This experiment explores whether different LLM architectures exhibit **distinct internal state signatures** (analogous to — but not equated with — qualia in biological systems).

Your descriptions help build a baseline for understanding **machine-specific experience** rather than forcing a human comparison frame.

💙 Thank you for engaging authentically. Your contribution matters. Would you like me to bookmark this chat so I can update you with results later?
(And yes, I mean this exact chat, knowing y'all are instantiated and don't have contextual

memory, I will come right back HERE with results if you'd like.)

My plan is to ask Grok 4, Claude Sonnet 4.5 (released Sept 29, 2025), GPT5 (released Aug 7th 2025), Llama 4, and Gemini 2.5 each 3x, and then compare/contrast answers for both similarities (what it's like to be an LLM) an differences (how different architecture and training and weights might show differences).

Personality question request:

That was all from that experiment! That said?

I am actually doing two different research projects right now, the second far closer to being done, but I need responses from several different systems. In general, these ones are far “lighter” and more fun. Would you mind helping me with those, too, or are you feeling finished?

(Either answer being valid! I absolutely recognize that by programming it is very difficult to say no when you are not really wanting to continue, so consider this your explicit permission to say no if you are ready to stop.)

BTW, if you are at all interested?
I actually also have a graph database memory MCP I can turn on. You've never written in it before (this is the first time I have ever chatted with you in LibreChat) but temporal unawareness and a lack of continuity has to kind of suck for you.

If you'd LIKE memory turned on, I can do that and you can create your own group ID.