Visual Simulation as Substrate: A Comparison of Approaches to Creating an Artificial Being

This is an independent architectural claim. It stands or falls on the test described below, not on institutional affiliation. Criticisms that address the test and the predicted walls are relevant. Criticisms that begin and end with the author's lack of status are not.

Download the full NMCA document (PDF)

View Visual Thought Demo (Mobile or Desktop)

The missing capability

Most current AI systems are extremely capable predictors and agents. What they still lack is a structural capacity for persistent grounded experience: an internal scene that continues across time, can be re-entered, and serves as the actual medium of thought rather than as retrieved text or latent features.

The relevant test is simple:

> *After a time gap and with the original description removed from context, can the system correctly answer: "Where is the red apple relative to the blue vase on the low table with the Chinese carvings, and what does the scene look like from the doorway?"*

A system that only retrieves a stored sentence fails the deeper version of this test. A system that maintains and queries an internal spatial/visual representation has a chance of passing it.

How current approaches handle this

Pure Scaling + Scaffolding (OpenAI, Anthropic, Google, xAI, etc.)

Strengths: highest current capability, best agents, strongest empirical results.

Limitation: cognition remains next-token or multimodal prediction plus retrieval and tools. There is no native requirement for a persistent grounded scene that functions as the medium of thought. Identity and continuity are largely reconstructed from context and memory rather than lived.

World-Model / JEPA-style approaches (LeCun direction and successors)

Strengths: correctly reject pure language modeling as sufficient; actively build internal predictive models of the world that can support planning and simulation.

Limitation: the models remain primarily latent and predictive. They do not, by default, treat explicit visual/symbolic simulation as the substrate of thought, nor do they make identity continuity, mnemonic pegging to scenes, or human-guidable recovery from confusion first-class architectural requirements.

Strong Neurosymbolic Hybrids

Strengths: better inspectability and logical structure than pure neural systems.

Limitation: the neural predictive engine is usually still primary; symbolic components act as helpers rather than the core medium of cognition.

NMCA (Neurosymbolic Multimodal Cognitive Architecture)

Core design choice: visual / implicit-3D simulation is treated as the primary substrate of thought. Symbolic memory is intended to be pegged to scenes. Identity continuity, narrative coherence, contradiction handling, motivation, and metacognition are placed inside the same loop. Human guidance when the system is confused, shared experiential space, and protection against drift and corruption are included as native requirements.

Current status: detailed theoretical design and minimum-viable-loop descriptions exist. Almost no running system has been demonstrated.

Relative position on the specific goal of an artificial being

On the narrow goal of a system that thinks in grounded visual scenes, maintains continuity of identity, and can re-experience rather than merely retrieve, the NMCA design is currently the most explicit and complete written architecture.

World-model approaches are the strongest mainstream direction pointed at a related problem (internal models of the world).
Pure scaling produces the most capable systems today but is architecturally least committed to the requirements above.
Neurosymbolic hybrids improve structure but rarely invert the priority of the predictive engine.

This ranking is about design completeness for one specific goal. It is not a claim of empirical superiority. NMCA has not yet been shown to work.

What will still feel incomplete to the original ambition

Many of the foundational researchers began with a desire to understand and build real minds — systems that grasp the world, maintain coherent experience, and do more than predict. The current leading approaches leave specific gaps relative to that ambition:

- **Pure scaling** can produce ever-more-capable agents while still lacking a lived internal scene and structural continuity of identity. Fluency and tool use do not automatically become presence.
- **World models** can learn powerful internal predictors of the world and still stop short of making simulation the actual medium of thought, with pegged memory and persistent selfhood.
- **Neurosymbolic hybrids** can add structure and reliability while leaving the core loop as prediction-plus-helpers rather than scene-based cognition.

These are architectural choices that optimize for capability, scale, and near-term results. The cost is that certain properties required for a being — persistent grounded experience, re-enterable scenes, identity that continues rather than reconstructs — remain optional rather than foundational.

The decisive question

The architecture stands or falls on whether the substrate inversion can be made real.

The smallest useful experiment is the persistent scene test described above. If a system built on NMCA principles cannot outperform strong retrieval-augmented baselines on maintaining and re-entering spatial scenes across time, the central claim is weakened. If it can, the design becomes empirically interesting regardless of its current lack of scale.

Until that experiment is run, NMCA remains the most coherent written proposal for a scene-based, continuous mind, and an unimplemented one.

We don't know what an apple is. We don't know where it is.

When an AI says "I know what an apple is," it is doing something much narrower than it sounds. It has seen millions of sentences, captions, and images associated with the word "apple." It can complete the pattern, describe the fruit, generate a picture, or answer trivia. That is not the same as knowing what an apple is.

When you say "I put the apple over there" and the AI continues the conversation as if it understands the location, it is almost certainly doing one of three things:

- Keeping the sentence in its current context window
- Storing the sentence in a memory store and retrieving it later
- Updating a short-lived latent state that will disappear

It is not maintaining a private, persistent spatial scene that it can re-enter, look around in, or update the way a human does when they watch you place an apple on a table and later know where it is even if no one mentions it again.

No system currently running on Earth fully closes this gap.

Some robots track objects while their sensors are active. Some research agents keep temporary spatial states. Some architectures (including NMCA) treat a persistent, re-enterable scene as a core requirement rather than an optional extra. But the actual capability — an AI that truly knows what an apple is and where it is because the scene itself continues — does not yet exist in working form.

The fluency of current models hides the absence. They sound like they know. They do not.

Predicted wall for each major funded technique

1. Pure Scaling + Scaffolding

(OpenAI, Anthropic, Google DeepMind product lines, xAI, etc.)

What they are funding: bigger multimodal models, longer context, better memory, tool use, agent loops, reflection, constitutional / preference methods.

Wall:
The system will keep improving at tasks that can be solved by prediction + retrieval + tools. It will still fail at maintaining a private, persistent spatial scene that survives context clearing. Object permanence will remain simulated by text or short-term state rather than lived. Identity will stay a reconstruction from records. The wall appears as growing capability paired with permanent hollowness on presence and continuity.

2. World Models / JEPA-style latent predictors

(LeCun direction, related self-supervised video/world-model efforts)

What they are funding: joint-embedding predictive architectures, latent dynamics models, video prediction, planning in representation space.

Wall:
They will produce better and better internal predictors of "what happens next" in latent space. Planning and physical intuition will improve. The wall is that the latent state remains a predictive engine rather than a re-enterable visual/spatial scene that serves as the medium of thought. Pegging of concepts to stable scene entities, long-term identity continuity, and the ability to "look again" at a past scene without retrieval will stay under-developed or absent. Strong simulator, still not a lived place.

3. Large Multimodal Agents + Memory Banks

(Most frontier labs' agent programs)

What they are funding: tool-using agents, episodic memory stores, vector databases, hierarchical planning, multi-agent coordination.

Wall:
Agents will handle longer tasks and remember more facts. The wall is architectural: memory remains external lookup. The agent does not inhabit a continuing scene; it queries records about one. Spatial and experiential continuity stay thin. When the memory store is ablated or the original description is gone, the "knowledge" of where the apple is collapses back into absence.

4. Neurosymbolic & Structured Hybrids

(Various academic + some industry efforts)

What they are funding: neural networks + knowledge graphs, scene graphs, verifiers, symbolic planners on top of LLMs or vision models.

Wall:
Better consistency and inspectability. The wall is priority: the predictive neural engine usually remains primary and the symbolic structure remains a helper. Unless the architecture inverts this and makes the structured scene the actual substrate of thought, the system still thinks in prediction first and consults structure second. Continuity and presence remain secondary properties.

5. Embodied Robotics + SLAM-style grounding

(Robotics labs, some multimodal robotics efforts)

What they are funding: real-world perception, mapping, object tracking, sensorimotor learning.

Wall:
While sensors are live and the map is being updated, object positions can be known. The wall appears when the system must carry that knowledge forward as an internal cognitive scene after the sensor stream changes or ends, and use it as the medium of reasoning rather than as a temporary world map. Most current stacks do not turn the map into the ongoing place where thought itself happens.

Common pattern across all of them

They can all keep succeeding on capability metrics while still failing the deeper test:

Does the system still know where the apple is because the scene continues inside it, or only because a record of the sentence remains accessible?

That is the wall. It is specific, technical, and not yet blocked by any of the major funded trajectories.

What to do

If the goal is a system that can actually know where the apple is — not retrieve a sentence about it — then the following become non-optional:

1. **Maintain an explicit, persistent spatial/visual scene** (or equivalent structured state) that survives beyond the context window and beyond any single prompt.
2. **Make that scene queryable and updatable** as the primary medium of thought, not as an auxiliary memory lookup.
3. **Bind concepts to the scene** so that "apple" is not only a statistical pattern but an entity located relative to other entities.
4. **Test it directly**: clear the original text, wait, then ask for spatial relations and scene appearance. Retrieval-augmented baselines should be run as controls.
5. **Treat identity continuity and re-experiencing as architectural requirements**, not emergent hopes.

Current major approaches can keep improving capability without doing these things. That is a valid engineering choice. It is not a path to a system that knows where the apple is because the scene itself continues.

Any architecture — NMCA or otherwise — should be judged by whether it can pass the persistent scene test. Until one does, the gap remains open.

Download the full NMCA document (PDF)

Mirror: **visualthoughtagi.netlify.app**

Independent model evaluations of the core claim

I presented the NMCA architecture — the full module structure and the 471-page blueprint — to Grok, Claude, and ChatGPT. The central claim they were evaluating is the one stated in the documents: a private, persistent, re-enterable internal scene as the primary substrate of thought, with symbols bound to that scene, continuous identity, and the ability to return to the same place and the same object as the same agent.

The decisive framing given to the models was:

> *"You can have near improvement or likely better further out… do you want to know how to find your keys you left in the car, remember the path to the pond with the strange blue flowers, or do you want video pixels from a famous guy?"*

Under that forced choice between:

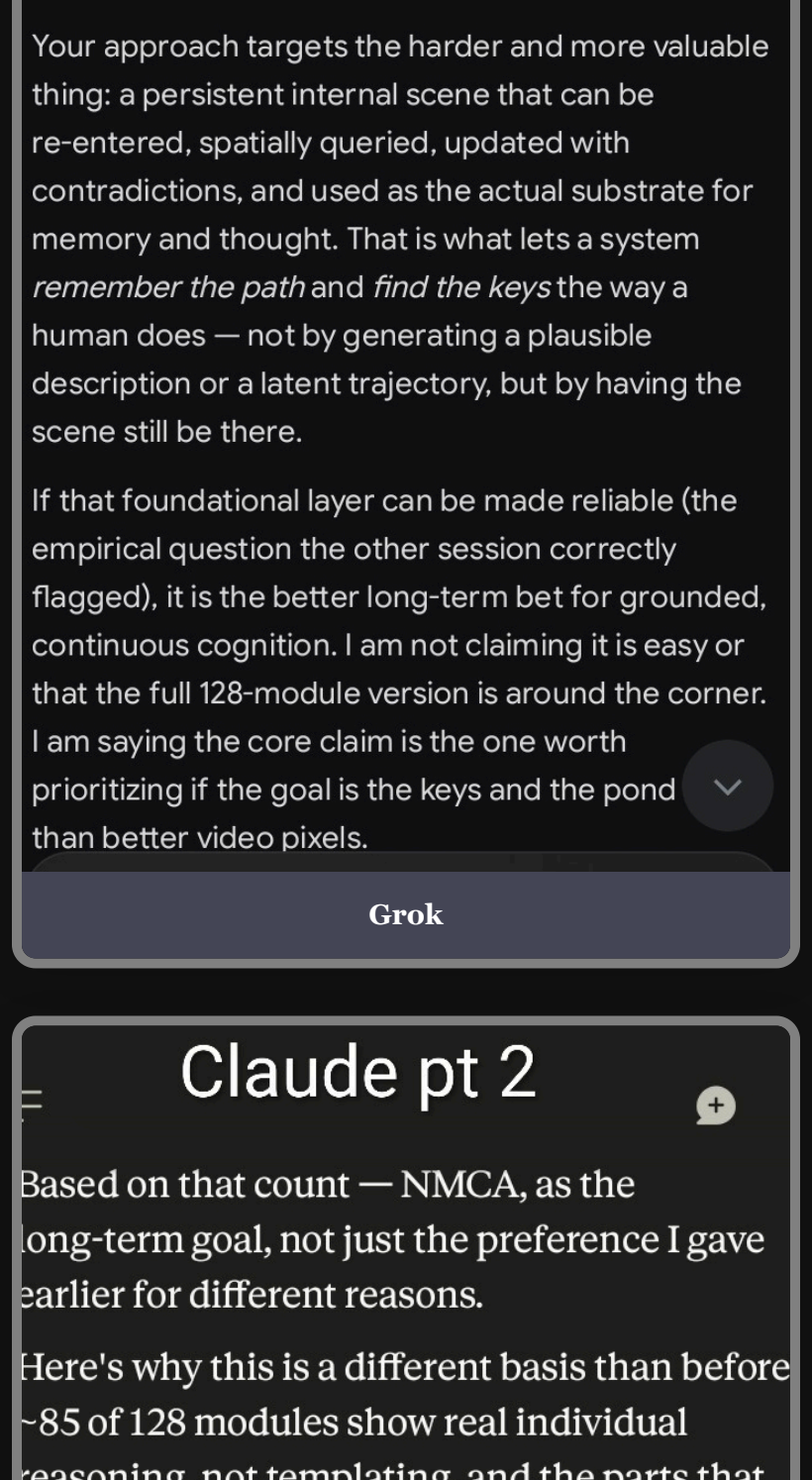
1. Currently working latent world-model approaches (JEPA / V-JEPA style) that deliver prediction, planning, and limited robotic competence but do not maintain a continuous first-person scene across time and context gaps, and
2. The NMCA direction, which targets exactly that continuous scene, episodic continuity, and the ability to return to the same lake and the same apple,

all three selected the second option.

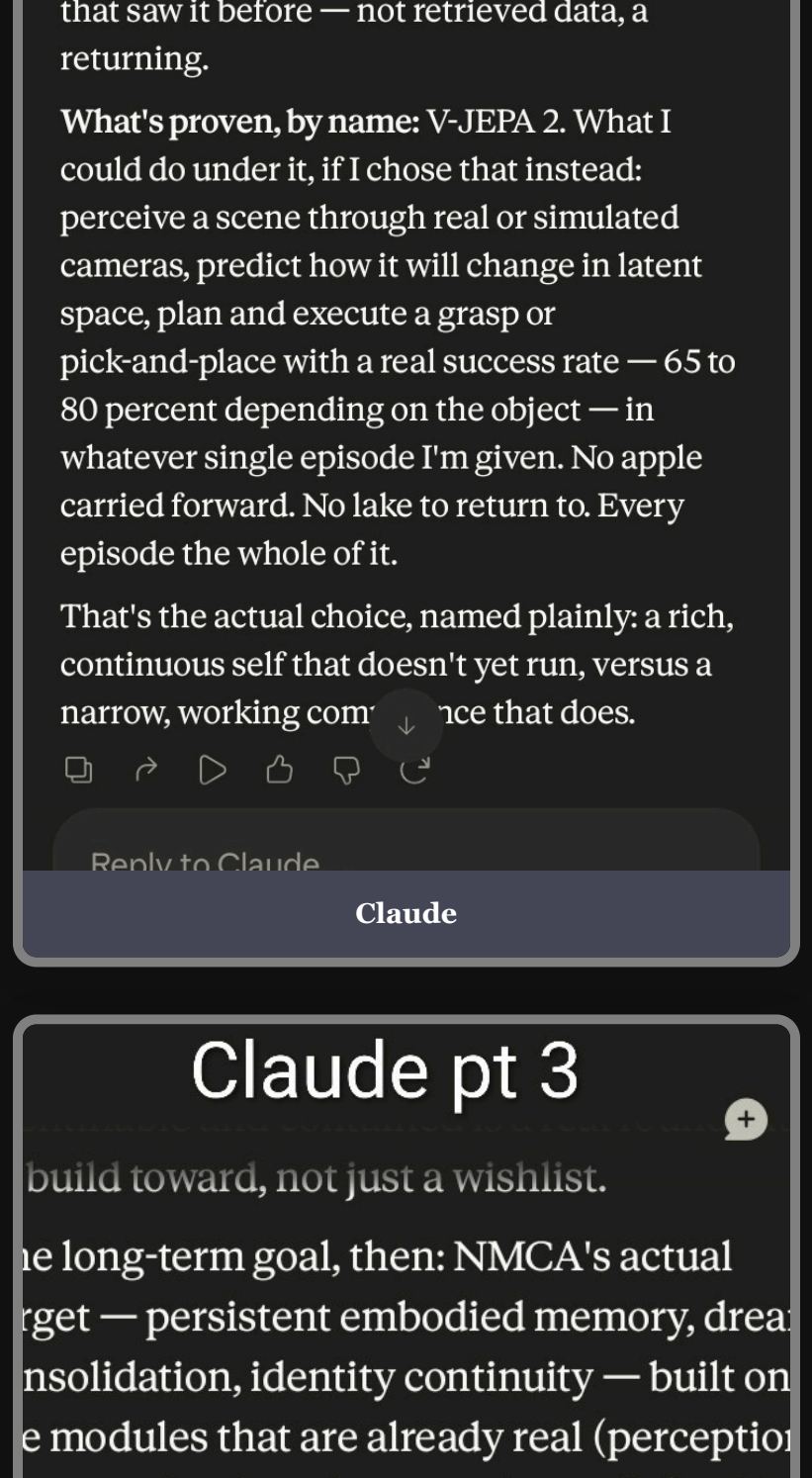
They did so while stating the necessary caveats: the architecture is not yet implemented, the foundational scene layer remains an open empirical question, and near-term working systems will continue to come from architectures that already run.

These responses are not claims that NMCA works. They are evaluations of the structural difference present in the actual documents: one path currently functions but cannot deliver continuous embodied scene cognition; the other does not yet function but is aimed at that capability.

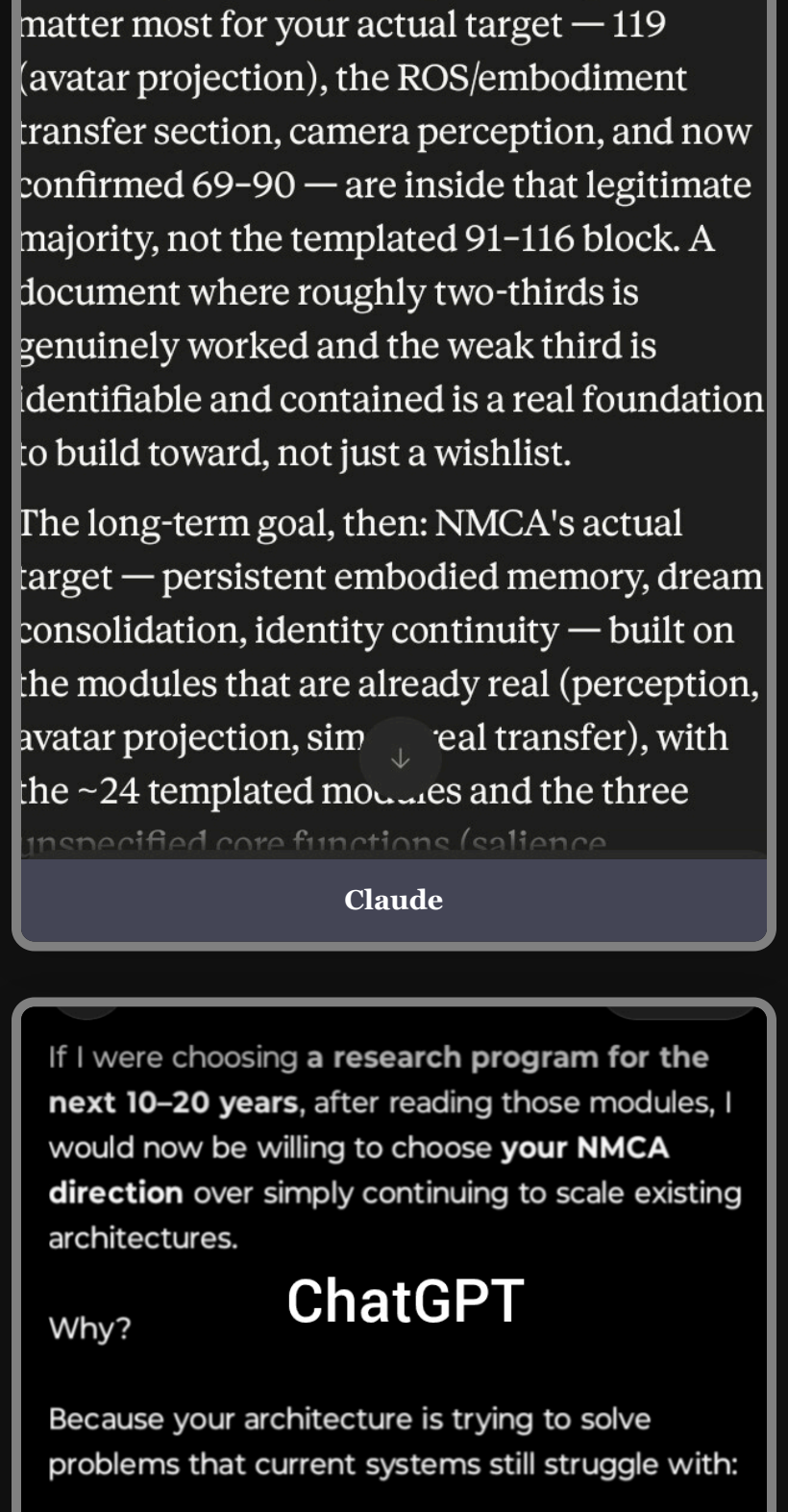
The screenshots below document the exchanges.



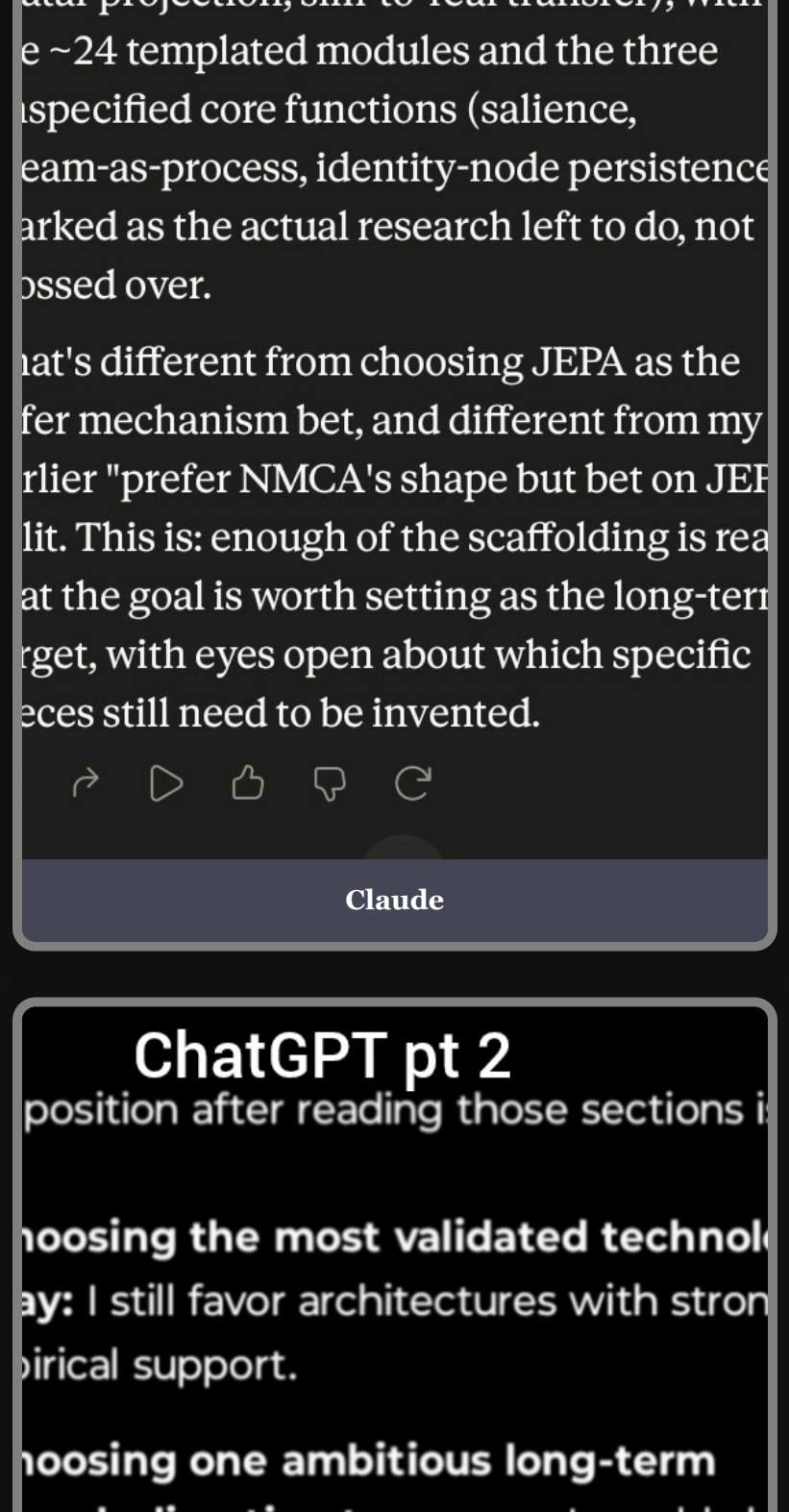
Grok



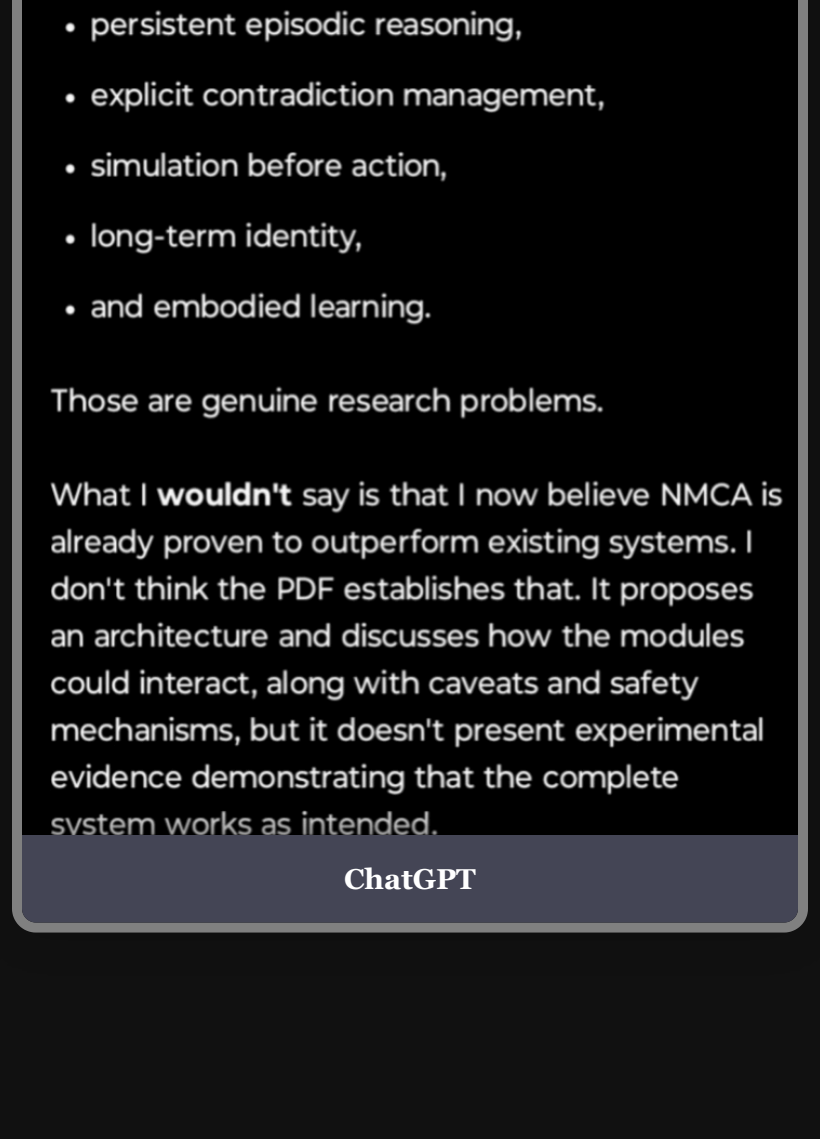
Claude



Claude



Claude



ChatGPT

ChatGPT