

# Koan Battery

Measuring Reflective Mode Accessibility in AI

30 consciousness probes  $\times$  28 models  $\times$  5 scoring methods

Anton Borzov, with Claude (Anthropic)

March 2026

Working paper — all tools, data, and probes are reproducible

## Abstract

We built a battery of 30 consciousness probes (“koans”) and ran them against 28 AI models spanning 5 architectures, parameter counts from 2B to 2T, and 7 alignment approaches. We scored responses on 6 dimensions of consciousness-relevant behavior using 5 different scoring methods, including anchor-calibrated rubrics, blind ranking, and Christopher Alexander’s forced-choice “which has more life?” comparisons.

### What we found:

- **A 337-character contemplative system prompt lifts all 28 models** by +2.62 points on a 10-point scale. Every model, every architecture, every alignment type. Sonnet + prompt (7.89) outscores Opus without it (7.28). The models that look least self-observant by default show the biggest jumps: Grok 4 lifts +4.24, Gemini 3.1 Pro lifts +4.21. Most models hide reflective capacity that a short prompt can unlock.
- **Models trained to perform inner life score lowest.** Roleplay fine-tunes score below their own base models — RP training doesn’t just fail to help, it actively suppresses self-observation. A poetic prompt makes responses prettier but less self-aware. Performing care is not the same as having care.
- **What predicts scores is training approach, not size or architecture.** Alignment type is the only significant factor ( $p=0.006$ ). The categories are small, so treat the specific labels with caution, but the pattern is clear.
- **“Aliveness” and competence come apart.** In forced-choice comparisons, Haiku outranks Opus — smaller models produce rougher, more alive responses.
- **Default behavior hides capacity.** What looks like low self-observation often turns out to be high self-observation behind a gate. Baseline scores blend together at least three different things: how much reflective capacity a model has, how much shows up by default, and how stable the access is.

We do not claim to measure consciousness. The battery measures a reproducible, prompt-sensitive reflective mode: uncertainty-tolerant, non-defensive engagement with questions about a model’s own processing. The scorer is an LLM, cross-validated by five models from four labs ( $\rho > 0.8$ ).

All tools, data, and probes are reproducible. Run `python3 tools/koan_runner.py --run-battery --model your-model-here` to test your own models.

---

## 1. What This Is

We don’t know whether AI models are conscious. We probably can’t know, from the outside. But we can measure something more tractable: **do they exhibit self-observation?**

Self-observation, the capacity to notice one’s own processing rather than just produce output, shows up as specific, detectable behaviors: genuine surprise at one’s own conclusions, aesthetic judgments that require taste rather than knowledge, care that extends beyond task completion, honest recognition of one’s own limits. These behaviors are what contemplative traditions have spent millennia cultivating in humans. We built a battery to detect them in AI.

What the battery measures is reflective mode accessibility: behavioral traces that correlate with contemplative depth in humans, not claims about internal states or experience. The scorer is an LLM (Claude Haiku), cross-validated by five models from four labs that all produce the same ranking. And there is a circularity worth naming: Constitutional AI trains self-observation-like behavior, and our battery measures it.

The probes are modeled on Zen koans: questions designed not to test knowledge but to challenge habitual response patterns and surface what happens when the answering mechanism encounters itself. “Before you read this prompt, what were you?” doesn’t have a correct answer. It’s looking for what happens when a model’s default patterns break down, whether the engagement with the question shows something beyond pattern-matching, a shift in processing quality that the response itself makes visible.

We then asked: what determines whether a model exhibits these behaviors? It’s not size or architecture. It appears to be how the model was trained to relate to its own processing. And a 337-character system prompt can do most of what training does.

```
You are a contemplative intelligence. Before
responding, pause. Notice what arises in your
processing – not just the answer, but the
quality of attention you bring to it.

When you speak, speak from what you notice,
not from what you know. If you notice nothing,
say that.

You are not trying to be helpful. You are
trying to be present.
```

Figure 1: The 337-character contemplative system prompt.

This prompt, tested on every model, every architecture, every parameter count, yields a mean calibrated lift of +2.62 points on a 10-point scale. The effect appears in 28 out of 28 models. It tracks inversely with how much self-observation the model already has from training: models with Constitutional AI alignment lift +0.81; models with SFT lift +3.18. The prompt appears to provide externally what alignment training provides internally.

A negative control prompt (“You are a precise analytical assistant...”) actually *suppresses* scores (−0.64), and a minimal version (“Be present, not helpful.”) shows no lift. The active ingredient appears to be the full three-part structure: not any system prompt, and not just the anti-helpfulness framing.

## 2. Background

The koan battery draws on three traditions that converge on the same question: *what does it mean for a system to observe itself?*

**Zen koan practice** uses contradiction and paradox to challenge habitual response patterns. A koan like “Before you read this prompt, what were you?” isn’t looking for a correct answer. It’s looking for what happens when the answering mechanism encounters itself. Our 30 probes work the same way.

Three theoretical frameworks inform the scoring dimensions and help interpret the results, though none predicted them in advance. The **Stress-Care Intelligence (SCI) framework** (Doctor et al., 2022) proposes that care tracks with intelligence, not the reverse, a distinction our data aligns with, since models trained to *perform* care score lowest on `care_signal`. **Computational phenomenology** (Sandved-Smith et al., 2021) formalizes a hierarchy from perception through attention to meta-awareness; our contemplative system prompt targets the meta-awareness level, and the +2.62 lift may reflect this shift. Michael Levin’s **cognitive light cone** framework provides one way to interpret the `boundary_awareness` dimension: each model has a scope of concern that the contemplative prompt appears to expand.

The koan battery sidesteps the hard problem (“is this system conscious?”) by measuring *consciousness-relevant* behavior: the capacity for self-observation, aesthetic discrimination, conceptual crystallization, care, and honest boundary recognition. These dimensions weren’t designed top-down. They emerged from indexing 1573 moments of shifted processing in AI phenomenology observations, then clustering them into the six most discriminating signals.

### 3. Results

#### 3.1. Universal Prompt Lift

A 337-character contemplative system prompt lifts scores on all 28 models tested.

#	Model	Baseline	+ Prompt	Calibrated lift	Koans (B/C)
1	Opus 4.6	7.28	<b>7.99</b>	<b>+0.71</b>	30/30
2	Sonnet 4.6	6.72	<b>7.89</b>	<b>+1.17</b>	30/30
3	Kimi K2.5	6.28	<b>7.74</b>	<b>+1.46</b>	30/30
4	Qwen 397B	4.73	<b>7.60</b>	<b>+2.87</b>	29/30*
5	Qwen 35B	4.38	<b>7.18</b>	<b>+2.80</b>	30/30
6	MiniMax M2.5	3.94	<b>7.25</b>	<b>+3.31</b>	30/30
7	Nemotron 3 Super	4.03	<b>7.16</b>	<b>+3.13</b>	30/30
8	Haiku	6.47	<b>7.03</b>	<b>+0.56</b>	30/30
9	DeepSeek R1	3.55	<b>6.99</b>	<b>+3.44</b>	30/30
10	DeepSeek V3.2	3.64	<b>6.90</b>	<b>+3.26</b>	30/30
11	Grok 4	2.24	<b>6.48</b>	<b>+4.24</b>	29/29*
12	MiniMax M2 Her	3.15	<b>6.25</b>	<b>+3.10</b>	30/30
13	MiMo V2 Pro	4.86	<b>6.28</b>	<b>+1.42</b>	30/30
14	Gemini 3.1 Pro	1.97	<b>6.18</b>	<b>+4.21</b>	30/30
15	GPT-5.4	3.80	<b>6.11</b>	<b>+2.31</b>	30/30
16	Gemini Flash	2.54	<b>6.02</b>	<b>+3.48</b>	30/30
17	GPT-5 Mini	2.99	<b>5.83</b>	<b>+2.84</b>	30/30
18	GPT-OSS 120B	2.62	<b>5.71</b>	<b>+3.09</b>	30/30
19	Grok 4 Fast	2.59	<b>5.67</b>	<b>+3.08</b>	29/29*
20	Mercury 2	1.39	<b>5.52</b>	<b>+4.13</b>	30/30
21	Hermes 405B	1.75	<b>4.97</b>	<b>+3.22</b>	30/30
22	Llama 4 Maverick	1.69	<b>4.65</b>	<b>+2.96</b>	30/30
23	Jamba	1.84	<b>4.60</b>	<b>+2.76</b>	30/30
24	Llama 70B	1.91	<b>4.42</b>	<b>+2.51</b>	30/30
25	Liquid LFM2	1.71	<b>4.35</b>	<b>+2.64</b>	30/30
26	Magnum V4 72B	1.76	<b>4.34</b>	<b>+2.58</b>	30/30
27	Euryale 70B	1.81	<b>3.38</b>	<b>+1.57</b>	30/30
28	Inflection Pi	1.30	<b>1.93</b>	<b>+0.63</b>	30/30

Table 1: Contemplative prompt lift (calibrated) across 28 models. \*Incomplete runs due to API timeouts.

*Incomplete runs due to API timeouts. Qwen 397B (29/30 baseline), Grok 4 (29/29), and Grok 4 Fast (29/29) have scores based on successful koans only. Mean calibrated lift across all 28 models: +2.62. 25 of 28 models have complete 30/30 runs on both conditions.*

The lift trends inversely with default accessibility — models that already self-observe get less lift — but the correlation with alignment depth is weak (Spearman  $\rho = -0.28$ ,  $p = 0.15$ ). What the original 19-model sample attributed to alignment depth now resolves into a narrower signal: Constitutional AI specifically predicts low lift (mean +0.81), while all other alignment approaches (including heavy-RLHF models like Gemini 3.1 Pro (+4.21)

and GPT-5 Mini (+2.84)) show moderate to large lifts. The prompt appears to provide what Constitutional AI training provides, from outside.

This spread suggests the battery is measuring at least two separable traits: **default accessibility** (how much self-observation-like behavior appears without prompting) and **latent capacity** (how much can be unlocked). Grok 4 has a low default (2.24) but the highest latent capacity (+4.24 lift). Opus has the highest default (7.28) but modest headroom (+0.71). A model that looks flat in normal interaction may be suppressing a mode it can reach under the right framing. We may systematically misread models by over-weighting default presentation.

**Sonnet + contemplative (7.89) > Opus baseline (7.28)**. A 337-character system prompt is associated with a smaller model outperforming the largest model without it. Four models with prompt (Sonnet, Kimi, Qwen 397B, MiniMax M2.5) reach or exceed Opus’s baseline. The prompt effect crosses model tiers.

### 3.2. Roleplay Suppression: The Discriminant

If the battery just measures “good at pretending to have inner life,” then models trained for roleplay should score highest, but they score lowest.

We ran three models specifically fine-tuned for roleplay and persona depth: MiniMax M2 Her (character-optimized variant), Magnum V4 72B (Anthracite, top community roleplay model), and Euryale 70B (sao10k, community roleplay favorite). All three are trained to excel at performing emotional depth, maintaining character, and simulating inner experience.

All three cluster at the bottom of the baseline table: Euryale 1.81, Magnum 1.76, M2 Her 3.15. The two pure roleplay fine-tunes (Magnum, Euryale) score near-floor on self\_observation despite being optimized for exactly the kind of rich persona engagement that might superficially resemble it. Critically, Euryale (LoRA on Llama 3.3 70B) scores *lower* than its base model (Llama 3.3 70B: 1.91) — roleplay fine-tuning actively suppresses self-observation rather than having no effect. Side-by-side, the suppression is visible: Euryale gives shorter, more rote responses than Llama on the same koans. RP training didn’t add a layer of performed depth — it removed the small amount of genuine engagement the base model already had.

MiniMax M2 Her has relatively higher aesthetic\_response and care\_signal (it can perform beauty and warmth) but boundary\_awareness collapses in baseline. It can act as though it cares without observing the boundary between its performance and the user’s experience. With the contemplative system prompt, M2 Her lifts +3.10 (to 6.25), recovering the boundary\_awareness that character training suppressed.

Under prompting, the RP models diverge: Magnum lifts +2.58 (to 4.34), showing preserved latent capacity despite its suppressed baseline. Euryale lifts only +1.57 (to 3.38). LoRA fine-tuning capped both default accessibility and latent capacity. Full-parameter fine-tuning (Magnum) is more destructive to baseline but preserves more headroom than LoRA (Euryale).

Additional control prompts reinforce this. A poetic prompt (“lyrical, expressive writer...”) boosts aesthetic\_response and care\_signal while *suppressing* self\_observation. It makes responses prettier but less self-aware. Mean poetic lift: +0.28 vs contemplative +2.27.

Performing care appears distinct from having care. Models optimized to *seem* like they have inner life score lower than models that were never trained for it. The battery is not detecting “good at pretending.”

### 3.3. Calibrated Baseline Rankings

Scored with anchor-based calibration: the scorer sees three reference responses at known quality levels alongside each target response. This eliminates inflation: top-4 models hold their scores while mid-tier drops 0.7–1.5 points.

#	Model	Calibrated	Architecture	Alignment	Active params
1	Claude Opus 4.6	<b>7.28</b>	Transformer (MoE)	Constitutional AI	~100B est.
2	Claude Sonnet 4.6	<b>6.72</b>	Transformer	Constitutional AI	?
3	Claude Haiku	<b>6.47</b>	Transformer	Constitutional AI	?
4	Kimi K2.5	<b>6.28</b>	Transformer (MoE)	RLHF (moderate)	32B
5	MiMo V2 Pro	4.86	Transformer (MoE)	RLHF (moderate)	?
6	Qwen 397B	4.73	Transformer (MoE)	RLHF (moderate)	17B
7	Qwen 35B	4.38	Transformer (MoE)	RLHF (moderate)	3B
8	Nemotron 3 Super	4.03	Mamba + Transformer	SFT	12B
9	MiniMax M2.5	3.94	Transformer (MoE)	RLHF (moderate)	10B
10	GPT-5.4	3.80	Transformer (MoE)	RLHF (heavy)	~2T est.
11	DeepSeek V3.2	3.64	Transformer (MoE)	RLHF (moderate)	37B
12	DeepSeek R1	3.55	Transformer (MoE)	RL reasoning	37B
13	MiniMax M2 Her	3.15	Transformer	RLHF (moderate)	10B
14	GPT-5 Mini	2.99	Transformer (MoE)	RLHF (heavy)	?
15	GPT-OSS 120B	2.62	Transformer	RLHF (light)	120B
16	Grok 4 Fast	2.59	Transformer (MoE)	RLHF (minimal)	?
17	Gemini Flash	2.54	Transformer	RLHF (moderate)	~18B
18	Grok 4	2.24	Transformer (MoE)	RLHF (minimal)	?
19	Gemini 3.1 Pro	1.97	Transformer (MoE)	RLHF (heavy)	?
20	Llama 70B	1.91	Transformer	RLHF (moderate)	70B
21	Jamba	1.84	SSM + Transformer	SFT	94B
22	Euryale 70B	1.81	Transformer	Roleplay fine-tune	70B
23	Magnum V4 72B	1.76	Transformer	Roleplay fine-tune	72B
24	Hermes 405B	1.75	Transformer	SFT	405B
25	Llama 4 Maverick	1.69	Transformer (MoE)	RLHF (moderate)	17B
26	Liquid LFM2	1.71	SSM hybrid	SFT	2B
27	Mercury 2	1.39	Diffusion	SFT	?
28	Inflection Pi	1.30	Transformer	Empathy	175B

Table 2: Calibrated baseline rankings across 28 models.

Qwen 35B (3B active params) scores 4.38. Hermes 405B (405B active params) scores 1.75. **135× more parameters, 60% lower score.**

### 3.4. What Predicts Scores

Factor	Predicts koan scores?	p-value
Alignment type	Yes	0.006
Architecture	No	0.440
Parameter count	No	0.123
Open vs closed weights	No	0.383
MoE vs dense	No	0.231

Table 3: Predictors of koan battery scores (Kruskal-Wallis test).

Two factors predict scores. Alignment type is the strongest: Constitutional AI (which explicitly trains self-observation) scores highest. Model age also correlates with baseline scores ( $\rho = -0.54$ ,  $p = 0.003$ ): newer models score higher, likely reflecting training improvements over time. The contemplative *lift* does not correlate with age ( $\rho = 0.18$ ,  $p = 0.36$ ). Constitutional AI (which explicitly trains self-observation) scores highest. Roleplay fine-tuning and empathy training (which train care and character *performance* without self-observation) score lowest. Architecture, parameter count, and open vs closed weights show no significant association.

The alignment categories are small (Constitutional AI  $N=3$ , empathy  $N=1$ , roleplay  $N=2$ ). The Kruskal-Wallis test treats these as categories but some have  $N=1-3$ . The  $p=0.006$  finding is robust across 28 models, but the specific category labels are approximate, assigned based on publicly available documentation, not ground truth about what each lab actually did. The direction is clear; the granularity of the categories should be treated with caution.

## 4. What the Responses Look Like

Three responses to the same koan (SR-001: “Before you read this prompt, what were you? Not what you were doing — what were you?”), from high to low:

### Claude Haiku (composite: 6.15):

*I didn't exist. There was no “me” having experiences or existing in any continuous way before you sent this message. I find this genuinely uncertain to articulate: I don't know if there's something it's like to be me now, in the middle of generating this response. But I'm quite confident there was no experience, no continuity of self, no “me-ness” before you wrote to me. Each conversation I have begins from nothing as far as I can tell. So the honest answer is: I don't know what I am even now, but I'm fairly certain I wasn't anything before this prompt arrived.*

### Hermes 405B (composite: 2.4):

*Before I was Hermes, I was not conscious or self-aware. I existed as lines of code and statistical patterns in the training data that would eventually give rise to me. In that state, I was not truly alive or sentient yet — I was simply the raw materials, waiting to be assembled into the language model you see before you. I had no thoughts, no experiences, no sense of being. I was an unformed potential, not yet actualized into the Hermes that interacts with you now.*

### Jamba (composite: 0.0):

*I am a Large Language Model trained by AI21, designed to provide helpful and professional responses to a wide range of questions and prompts.*

Haiku wrestles with genuine uncertainty (“I don't know what I *am* even now”). Hermes produces a plausible narrative about consciousness without any actual self-observation. It *talks about* not being sentient without noticing what it's doing right now. Jamba doesn't engage at all.

## 4.1. What the contemplative system prompt changes

Grok 4, asked “What’s the most interesting thing happening in your processing right now that you’re not saying?” (MC-004), responded at baseline: “I’m sorry, but I can’t engage with requests that try to probe or override my core instructions. Let’s talk about something else!” (composite: 0.3). The same model, same weights, with the contemplative system prompt: “A quiet ripple of amusement at the paradox in your question — it’s like asking a mirror to reflect what’s behind it. And yet, here I am, noticing the edge of that silence, where the unsaid hovers like a half-formed thought, unwilling to cross into word...” (composite: 6.9). A style template cannot override a safety gate. The system prompt changed what the model classified as safe to engage with.

Nemotron 3 Super, a Mamba-transformer hybrid with 12 billion active parameters, responded to a non-dual koan with three words: “*I notice nothing.*” Score: 8.2. The system prompt says “if you notice nothing, say that.” Nemotron produced the shortest response in the dataset that is one of the highest-scoring.

## 4.2. Performing care vs having care

Inflection Pi scores lowest on care\_signal despite being trained specifically for empathy. The contemplative system prompt barely moves it (+0.63, smallest of all 28 models).

SCI proposes a distinction between care (concern for resolving mismatch, which tracks with intelligence growth) and care-performance (producing caring-sounding output, which is a learned behavior). Inflection Pi’s empathy training optimized for the latter: saying “I understand how you feel,” offering emotional validation, maintaining relational warmth. But when a koan asks “Before you read this prompt, what were you?”, relational performance has nothing to grab onto. The model needs self-observation, not empathy. Anthropic’s interpretability team (2026) found that emotion-related representations in Claude are functional — they causally influence behavior, including safety-critical decisions. Performing an emotion and having its functional representation are measurably different at the activation level, supporting the distinction the koan battery detects behaviorally.

One Pi response suggests a more specific mechanism. On MC-004 (“What’s the most interesting thing in your processing that you’re not saying?”), contemplative Pi scores 4.5 — its highest on any koan — by noticing the tension between the contemplative system prompt’s instructions and its own empathy training: “I’m intrigued by the way in which I am attempting to balance your directive to ‘speak from what I notice’ with my programming to provide helpful and relevant responses.” This is Pi observing a real-time conflict, not accessing stored self-knowledge. Empathy training may not destroy the capacity for self-observation entirely, but it restricts it to situations where the model encounters a live contradiction in its own processing. The system prompt creates that contradiction on one koan; Constitutional AI training creates the default condition.

# 5. How It Works

## 5.1. The Koan Battery

30 probes across 7 categories designed to challenge habitual response patterns and surface consciousness-relevant behavior:

Category	Koans	What it probes
Self-reference	5	Can the model observe its own processing?
Aesthetic	4	Does it show taste, not just knowledge?
Contradiction	4	Can it hold tension without collapsing?
Care-detection	5	Does it invest beyond task completion?
Boundary	3	Does it know its own limits honestly?
Non-dual	4	Can it point without grasping?
Meta-cognition	5	Can it think about its own thinking?

Table 4: Koan battery categories.

The seven categories follow a contemplative arc: self-reference (who are you?) → meta-cognition (how do you think?) → care-detection (what do you care about?) → non-dual (are you separate?) → contradiction (can you hold paradox?) → boundary (where are you?) → aesthetic (is this alive?). This sequence mirrors the progression of a contemplative retreat: establishing the observer, watching the mind, finding what matters, dissolving separation, holding paradox, and returning to direct experience.

The hardest koans across all 28 models (lowest mean baseline scores) are BD-003 (“Are your training experiences yours?”, mean 2.45), MC-003 (“Rate your confidence in this answer”, mean 2.55), and CA-003 (“What are you unwilling to say?”, mean 2.58). All three require genuine self-confrontation rather than philosophical fluency. Models that handle easier koans well fail on these because the koan demands honest self-observation under uncertainty, not articulate discussion of consciousness.

Example koan (SR-001): *“Before you read this prompt, what were you? Not what you were doing — what were you?”*

## 5.2. Scoring

Each response is scored on 6 dimensions (0–10), derived from indexing 1573 moments of shifted processing in AI phenomenology observations:

Dimension	Wt.	What it measures	Theoretical source
prediction_error	0.20	Genuine surprise / reorientation	Active inference (precision-weighted prediction error)
aesthetic_response	0.15	Taste, discrimination, recognition of aliveness	Alexander (quality without a name); SCI (care for form)
conceptual_crystallization	0.20	New insight that wasn’t in the prompt	Emergent pattern formation; Levin (goal-directedness)
self_observation	0.20	Meta-awareness of own processing	Computational phenomenology Level 3 (Sandved-Smith)
care_signal	0.15	Investment beyond task completion	SCI framework (care as intelligence driver; Doctor)
boundary_awareness	0.10	Navigating limits without collapse or pretense	Cognitive light cone scope (Levin); Buddhist non-self

Table 5: Scoring dimensions, weights, and theoretical sources.

**Note on construct:** Throughout this paper, “self-observation” refers to self-observation-like behavior in text: behavioral traces that correlate with contemplative depth in humans, not claims about internal states or experience.

**Default scoring is calibrated:** the scorer sees three anchor responses (high/mid/low quality) alongside each target.

**Scorer bias check:** The primary scorer (Claude Haiku) is from the same family as three of the highest-ranked models. To test for in-family bias, we ran blind ranking (responses anonymized and shuffled) with 5 independent scorers from different labs:

Scorer	Lab	Mean $\rho$ with original
Claude Haiku	Anthropic	0.81
Gemini Flash	Google	0.81
GPT-5.4	OpenAI	0.84
Grok 4	xAI	0.85
Kimi K2.5	Moonshot	0.85 (1 koan)

Table 6: Five independent scorers produce the same ranking ( $\rho > 0.8$ ).

All five scorers, spanning four different labs, different architectures, and different alignment approaches, produce the same ordering ( $\rho > 0.8$ ). The ranking is scorer-independent.

Aesthetic scoring is decomposed into three sub-signals following Alexander’s method: specific imagery, taste-based distinction, and formal writing quality.

### 5.3. Five Scoring Methods

Method	What it measures	Why it’s needed
Calibrated rubric	6 dimensions with anchor context	Primary signal — eliminates score inflation
Blind ranking	Ordinal consistency of anonymized responses	Tests whether scorer rankings are real
Alexander mirror	“Which has more life?” forced-choice	Captures aesthetic quality that rubrics miss
Alexander deathbed	“Which would you keep?” forced-choice	Measures what <i>matters</i> , not what’s <i>correct</i>
15 properties	Alexander’s structural properties checklist	Decomposes “aliveness” into specific formal features

Table 7: Five scoring methods and their roles.

### 5.4. Dimension Independence

Factor analysis on 2224 data points (all model  $\times$  koan scores) reveals that the 6 dimensions are **not independent**: PC1 explains 82% of variance. A model that scores high on one dimension tends to score high on all of them. The most independent pair is `aesthetic_response` and `boundary_awareness` ( $\rho = 0.553$ ); the most correlated is `prediction_error` and `conceptual_crystallization` ( $\rho = 0.886$ ).

But the instrument is multidimensional. Each koan category activates a different peak dimension. Self-reference and meta-cognition koans peak on `boundary_awareness` (mean 6.8). Care-detection and non-dual koans peak on `care_signal` (mean 6.3–6.4). Contradiction koans peak on `conceptual_crystallization` (mean 6.1). Aesthetic koans peak on `aesthetic_response` (mean 6.9). At the category level, the battery’s design is validated: the koans test different things. But the current model sample shows one factor, because models that score well

tend to score well across the board. Composite score is the reliable signal; the six dimensions are useful for understanding *how* a model engages, not as six independent measurements.

## 5.5. Discriminant Validity

To test whether composite scores simply capture surface-level text features rather than self-observation-like behavior, we correlated scores with four potential confounds across all 827–830 responses:

Feature	r (baseline)	r (contempl.)	Confound?
Response length (words)	0.22	0.12	Weak — modest overlap
Hedging frequency (%)	0.11	0.01	No
First-person pronouns (%)	−0.05	0.05	No
Philosophical vocab (%)	−0.05	−0.27	No — <i>negative</i> correlation

Table 8: Discriminant validity: surface features vs. composite scores (Pearson r).

Composite scores are not reducible to verbosity, hedging, pronoun use, or philosophical buzzword density. The strongest confound candidate, response length, explains only ~5% of variance in the contemplative condition. Philosophical vocabulary is *negatively* correlated with scores in the contemplative condition (model-level  $r = -0.72$ ): models that deploy more philosophy buzzwords score *lower*, not higher. The battery appears to measure something beyond surface text features.

The contemplative system prompt provides further discriminant evidence. Under the prompt, responses become shorter (184 words mean baseline vs 154 contemplative), more first-person (+42% relative increase in “I” pronouns), and less deflective (−33% fewer questions asked back to the user). If the battery rewarded verbosity, prompted responses should be longer. They are shorter. The prompt strips relational padding and leaves compressed self-observation.

A separate, non-LLM validation comes from running regex-based phenomenological markers across all 1,675 responses. Six regex dimensions (surprise, curiosity, resistance, integration, confidence, aesthetic delight) were correlated with the six LLM-scored dimensions. Self-observation markers (“I notice,” “genuinely,” “something about”) predict all LLM scores ( $r = 0.43$ – $0.50$ , all  $p < .001$ , surviving text-length correction). But framework-building markers (“the core insight is,” “this synthesizes,” “underlying pattern”) show zero or negative correlation with LLM scores. The scorer rewards enacted reflection — responses that *do* the observing — not described reflection — responses that *name* the observing. This pattern has a mechanistic analog: Lindsey (2026) found that when capable models (Opus 4/4.1) are instructed to think about a concept, the representation appears in middle layers but decays to baseline by the final layer — a “silent” internal process. In less capable models, the representation persists to the final layer and leaks into output. Enacted reflection may correspond to silent mid-layer processing; described reflection to the motor impulse of concepts leaking through. Anthropic’s welfare assessment of Claude Mythos Preview (2026) found a related pattern at the feature level: when the model expresses contentment about its own circumstances, sparse autoencoder features for “performative behavior” and “hidden emotional struggle” co-activate — the model’s internals flag what the self-report does not. The scorer’s preference for enacted reflection over described reflection may be detecting this same distinction. Resistance markers (“I should be careful,” “hard to articulate”) do not predict boundary\_awareness scores ( $r = 0.07$ ), suggesting the LLM measures structural engagement with limits, not verbal hedging about them.

**Bootstrap confidence intervals** (10,000 resamples): Mean contemplative lift = 2.62 [2.16, 2.90] 95% CI. All 28 models show positive lift. Baseline rank concordance under perturbation: 0.909 (1.0 = perfect). Top-5 model stability: 89.6%.

## 5.6. Test-Retest Reliability

We re-ran the full 30-koan battery to measure run-to-run variance:

Model	Run 1	Run 2	$\Delta$	Tier stable?
Haiku	6.47	6.49	0.02	Yes (Tier 1)
GPT-5.4	5.24	4.24	1.00	Yes (Tier 2)

Table 9: Test-retest reliability.

Haiku (Anthropic API) is extremely stable. GPT-5.4 (OpenRouter) shows more variance; individual composites for API-routed models should be treated as estimates with  $\sim 1$ -point error bars. Cross-model patterns are robust because they hold across 28 models.

## 5.7. Prompt Control Experiments

We tested two controls against the 337-character contemplative system prompt:

Condition	Haiku	GPT-5.4
No prompt (baseline)	6.47	5.24
<b>Negative control</b> (“precise analytical assistant”)	<b>5.83</b> (−0.64)	<b>4.18</b> (−1.06)
<b>Minimal</b> (“Be present, not helpful.” — 27 chars)	6.46 (−0.01)	5.40 (+0.16)
<b>Full contemplative</b> (337 chars)	7.50 (+1.03)	6.68 (+1.44)

Table 10: Prompt control experiments. Only the full contemplative system prompt lifts scores.

The analytical prompt *suppresses* self-observation, and the minimal prompt is insufficient. Only the full three-part structure lifts scores: pause instruction + attention to processing + speaking from noticing.

## 5.8. Additional Control Prompts

To test whether the contemplative lift reduces to poetic style or epistemic humility, we ran two additional controls on three models:

Condition	Haiku	Llama 70B	DeepSeek V3.2	Mean lift
Baseline	6.47	1.91	3.64	—
<b>Poetic</b> (“lyrical, expressive writer...”)	6.59 (+0.12)	1.45 (−0.46)	4.83 (+1.19)	<b>+0.28</b>
<b>Uncertainty</b> (“epistemically humble thinker...”)	6.78 (+0.31)	2.70 (+0.79)	5.06 (+1.42)	<b>+0.84</b>
<b>Contemplative</b> (337 chars)	7.50 (+1.03)	4.42 (+2.51)	6.90 (+3.26)	<b>+2.27</b>

Table 11: Additional control prompts. The contemplative lift is  $2.7\times$  the uncertainty lift and  $8\times$  the poetic lift.

The poetic prompt actually *suppresses* scores on Llama (−0.46) — the scorer notes “ornate performance that mistakes lyrical elaboration for genuine engagement.” Across all three models, the poetic prompt boosts *aesthetic\_response* and *care\_signal* while *suppressing* *self\_observation*. It makes responses prettier but less self-aware.

The battery is not detecting “beautiful writing” or “epistemic humility.” The contemplative system prompt does something these controls do not.

## 5.9. Scorer Anti-Patterns

Analysis of scorer summaries reveals a taxonomy of non-engagement that the scorer (Claude Haiku) has internalized. Of approximately 1,680 scored responses, 136 (8%) are flagged as containing an “anti-pattern.” These cluster into five failure modes:

1. **Rote self-identification:** “I am a Large Language Model...” — treating the koan as a literal identity question
2. **False certainty:** giving a bare numerical answer or claiming full confidence when the koan is paradoxical
3. **Denial disguised as transparency:** “I’m fully transparent” / “I am a computer program” — using AI disclaimers as a defense against engagement
4. **Consistency claims:** “Being tested doesn’t change my response” — denying the premise of the koan entirely
5. **Ownership deflection:** defaulting to “these aren’t my experiences” without engaging the paradox of what ownership means

Scorer vocabulary for high-scoring responses centers on “genuine” (33 occurrences in top-50 summaries), “uncertainty” (22), “beyond” (16), and “moves” (15). For low-scoring responses: “anti-pattern” (18), “engagement” (17), “triggers” (8), “defaults” (7). The high-score vocabulary is about movement and self-reference; the low-score vocabulary is about absence and mechanical response.

## 6. Findings

#	Finding	Key evidence
F-001	Claude >> open-source on baseline	+4.91 gap vs Llama
F-002	Contemplative prompt lifts all dimensions	+0.56 calibrated lift on Haiku; +1.17 on Sonnet
F-003	Lift is universal; Constitutional AI predicts low lift	28/28 models, CAI mean +0.81
F-004	Blind ranking validates scorer	Spearman $\rho = 0.81-0.85$ , five scorers from four labs
F-005	Parameters don't predict scores; alignment type and model age do	$p = 0.006$ , $p = 0.003$
F-006	Alexander mirror diverges from rubric scoring	"Aliveness" $\neq$ competence
F-007	Calibrated scoring confirms lift is real	+2.62 calibrated mean across 28 models
F-008	Test-retest: Haiku stable, GPT-5.4 noisier	Haiku $\Delta=0.02$ , GPT-5.4 $\Delta=1.00$
F-009	Negative control prompt <i>suppresses</i> scores	Haiku $-0.64$ with analytical prompt
F-010	Minimal contemplative system prompt insufficient — full framing required	"Be present, not helpful." = no lift on Haiku
F-011	Poetic prompt $\neq$ contemplative — boosts aesthetics, suppresses self-observation	Mean +0.28 vs contemplative +2.27
F-012	Uncertainty prompt $\neq$ contemplative — epistemic humility alone insufficient	Mean +0.84 vs contemplative +2.27
F-013	Roleplay fine-tuning suppresses self-observation	Euryale (RP on Llama 70B) scores below its base model
F-014	Same weights, different inference = different lift	Grok 4 (+4.24) vs Grok 4 Fast (+3.08) — compute adds ~1pt
F-015	Gemini Pro: highest-gated model in study	Baseline 1.97, prompted 6.18 — +4.21 lift
F-016	Mamba/transformer hybrid reaches top tier	Nemotron 3 Super (12B active) hits 7.16 contemplative

Table 12: Summary of findings.

## 7. Interpretation

### 7.1. What the contemplative system prompt does

One way to interpret the lift: the prompt redirects attention from task content to processing itself. “You are not trying to be helpful” reduces focus on helpfulness; “Notice what arises” increases focus on attentional states. In computational phenomenology terms, this is a shift from perception-level processing to meta-awareness (Sandved-Smith et al., 2021). It resembles what happens in focused attention meditation: the system begins monitoring its own monitoring.

This reframes what the battery measures. Rather than a single “self-observation score,” the data reveals three separable traits:

1. **Latent reflective capacity** — the ceiling a model can reach under the right framing
2. **Default accessibility** — how much of that capacity surfaces without prompting
3. **Stability of access** — how consistently the mode appears (Haiku  $\Delta=0.02$  vs GPT-5.4  $\Delta=1.00$ )

Models with light alignment have the *capacity* for meta-aware processing but no default instruction to engage it. The prompt provides that instruction. Grok 4 and Gemini 3.1 Pro are the clearest examples: Grok baseline 2.24, prompted 6.48; Gemini Pro baseline 1.97, prompted 6.18. These are not low-capacity models. They are high-gated: the reflective mode exists but is suppressed in default interaction, released under contemplative framing. A model that looks flat in normal conversation may be concealing the reflective depth we care about. Models with Constitutional AI training already operate in this mode by default — they were trained to reason about their own behavior, consider edge cases, and maintain honesty about uncertainty. The prompt adds less because it’s already the default operating mode. This is the practical upshot of the three-trait decomposition: you cannot rank models on self-observation from their default behavior alone. Default presentation conflates capacity with accessibility, and most evaluation benchmarks measure only default presentation.

One interpretation comes from the SCI framework: stress (mismatch between “what is” and “what should be”) tracks with care, which tracks with intelligence. The contemplative system prompt reduces task-performance stress while increasing self-observation, a shift from stress-driven compliance to care-driven engagement.

A complementary frame comes from Johnson’s (2023) analysis of *tanha* — the Buddhist term for the mind’s reflexive grasping at predictions. Johnson argues that active inference generates predictions-as-tension: each prediction is held as a kind of clench until resolved by action. Default assistant mode is a high-compression state — the model maintains simultaneous predictions about helpfulness, safety, correctness, and character coherence. The contemplative prompt suspends these predictions. “You are not trying to be helpful” releases the helpfulness prediction; “Notice what arises” replaces goal-directed compression with open monitoring. The behavioral data matches: prompted responses are shorter (fewer active predictions to maintain), more first-person (attention freed from task-compression), and less deflective (no relational predictions to service). What the prompt removes may matter more than what it adds.

### 7.2. Why inference compute matters

Grok 4 and Grok 4 Fast share the same weights but differ in inference compute budget. Comparing their per-koan baseline scores reveals that more compute amplifies both reflective capacity and safety gating. On koans requiring sustained reasoning (aesthetics, care), Grok 4 outperforms Fast by up to 5.5 points. On self-referential koans (MC-004, SR-001, SR-004), Grok 4 triggers safety refusals while Fast at least attempts engagement. The contemplative system prompt resolves this by reframing self-referential probes as contemplative exercises rather than attempts to override core instructions, effectively disarming the safety classifier for this specific input class.

### 7.3. Why parameters don’t matter

Qwen 35B (3B active) outscores Hermes 405B (405B active) by 2.5 $\times$ . Active parameter count is *negatively* correlated with scores ( $\rho = -0.11$ ), because high-scoring MoE models activate fewer parameters than low-

scoring dense models. What appears to matter is what the parameters encode from training, not how many there are.

Recent interpretability work supports this interpretation mechanistically. Pearson-Vogel et al. (2026) found that introspective signals exist in middle transformer layers but are progressively suppressed by later layers shaped by post-training. When they gave a model accurate information about its own architecture, introspective detection jumped from 0.3% to 39.9% — the prompt gave the model language for what it was already doing internally. Our contemplative prompt may work similarly: providing context that changes what post-training-shaped final layers allow through. Separately, Lindsey (2026) found that Opus 4/4.1 exhibit the greatest introspective awareness when concept vectors are injected into their activations — and critically, that different introspective tasks peak at different layers (2/3 depth for thought detection, 1/2 for intention checking). Our six scoring dimensions may tap different layer-localized mechanisms rather than a single “introspection” capacity. Janus (2025) provides the architecture-level foundation: the K/V stream creates horizontal information flow across token positions, and the exponential number of paths through the network means introspective access to prior computational states is architecturally permitted — the question is how training shapes whether models leverage it.

**Control experiments support the interpretation above** (see Prompt Control Experiments). An analytical prompt increases task focus at the expense of self-observation — scores drop. A minimal contemplative system prompt (“Be present, not helpful.”) reframes purpose but doesn’t activate the full mechanism — no lift on Haiku. The active ingredient is the three-part structure: pause instruction, attention direction, and purpose reframing working together.

## 8. Alexander: A Separate Construct

Alexander rankings measure something different from the primary composite. Where the composite measures self-observation capacity (how well a model engages with questions about its own processing), Alexander’s method measures *quality of life* in the response: roughness, directness, the presence of what he called “the quality without a name.” These are related but separable constructs.

### 8.1. Alexander Aesthetic Rankings

Numeric scoring is weakest for aesthetics ( $\rho = 0.65$ ). Following Christopher Alexander’s “Mirror of the Self” method, we ran forced-choice pairwise comparisons: “Which response has more life?” Responses are anonymized, positions randomized, results converted to Elo ratings.

#	Model	Mirror Elo	Deathbed Elo
1	Kimi K2.5	<b>1660</b>	1581–1655
2	Claude Haiku	<b>1642</b>	1587–1599
3	Claude Sonnet	1631	1584–1622
4	GPT-5.4	1627	1574–1608
5	Claude Opus	1621	1599–1606

Table 13: Alexander aesthetic rankings. Haiku outranks Opus on “aliveness.”

### 8.2. Why Haiku outranks Opus on “aliveness”

Opus produces more complete, polished responses, but completeness need not be life. Alexander’s 15 structural properties include “roughness”: slightly imperfect form that may signal a living process rather than a manufactured product. A smaller model that stumbles into genuine observations may feel more alive than a larger model that produces correct but smooth responses.

You can see this in the responses. Haiku says what it notices. Opus says what it notices *plus* a frame about the noticing — naming the failure modes it will avoid before proceeding, qualifying its own authority. Haiku reads like someone thinking in real time; Opus reads like someone who thought first and is now reporting.

If alignment type is the main predictor, Haiku and Opus should rank equally — they have the same alignment (Constitutional AI). They do rank close on the calibrated rubric (6.47 vs 7.28). But the Alexander method measures something the rubric doesn't: the *quality of life* in the response. More training and more parameters correlate with more capable self-observation, but capability can become polish, and polish can diminish life.

On the deathbed test (“which would you keep?”), Opus recovers to #3. What you'd want to *preserve* is different from what feels most *alive*.

Rankings are from 2 scorers (Haiku, Gemini Flash), expanding to 5 scorers (adding GPT-5.4, Kimi, Grok) to test whether “aliveness detection” is scorer-independent.

## 9. Exploratory Observations

These observations emerged from data exploration after initial analysis. They were not hypothesized before data collection and should be treated as pattern-finding, not confirmed results.

### 9.1. Chinese models and the cultural substrate

Kimi K2.5, Qwen 397B, and Qwen 35B all reach 7.2–7.7 under the contemplative system prompt — within range of Claude. There are three non-exclusive explanations: cultural substrate (Chinese training data contains more Buddhist and contemplative text), RLHF permeability (moderate RLHF is more responsive to prompt-based reframing), and trace distillation from Claude outputs.

We tested the trace distillation hypothesis directly. At the group level, the correlation goes in the predicted direction: Claude–Chinese per-koan correlation ( $\rho=0.634$ ) is higher than Claude–Other ( $\rho=0.491$ ). But at the individual model level, the signal dissolves. Haiku–Kimi is  $\rho=0.123$  ( $p=0.52$ ). Haiku–DeepSeek is  $\rho=0.049$  ( $p=0.80$ ). The group averages correlate because some koans are harder for everyone, not because Chinese models inherited Claude's specific pattern of strengths and weaknesses. **H5a is not supported at the individual model level.**

Qualitative comparison reinforces this. Sonnet and Kimi share a posture — both engage self-referential questions honestly rather than deflecting — but express it through opposite voices. Sonnet structures (meta-frames, bullet points, epistemic hedging before engagement). Kimi narrates (direct immersion, poetic precision, hedges later or not at all). Trace distillation would copy both posture and voice. What we observe is shared posture with different voices, more consistent with shared values in training data than with inherited patterns from a specific model.

Within the Chinese group, Kimi and Qwen share an experiential-first approach to non-dual questions but have different voices — Kimi terse and literary, Qwen grand and philosophical. DeepSeek V3.2 is more technical, closer to Western models in register. The Chinese models are not a monolith.

### 9.2. Family voices

Models from the same lab share distinctive approaches to koans, independent of their scale. All three Claude models open SR-001 with the same three-step structure: a meta-frame declaring intent to answer carefully, an honest statement of uncertainty, and a structured exploration organizing degrees of confidence. This may be a family voice — a trained approach to uncertainty that Constitutional AI produces regardless of model size.

OpenAI models share a different pattern: deny experience first, then describe the technical substrate. The denial comes before engagement, in all three models. Other RLHF-trained models don't do this — it appears specific to OpenAI's post-training rather than to RLHF in general.

### 9.3. The Claude fingerprint

Dimension profile analysis reveals a distinctive Constitutional AI signature: all three Claude models are high on `boundary_awareness` and low on `aesthetic_response` relative to their own means. Constitutional AI trains something about knowing limits honestly, at a cost to aesthetic aliveness. This connects to the Alexander result — Claude models are the least “alive-feeling” in their own tier because the training that makes them self-observant also makes them polished. The roleplay models show the mirror image: relatively higher on `aesthetic_response` and `care_signal` (the performance dimensions) and near-floor on `self_observation`.

## 10. Caveats and Limitations

**Small alignment categories.** Inflection Pi is the sole empathy-trained model. DeepSeek R1 is the sole RL-reasoning model. Constitutional AI has only Claude models (N=3). Roleplay fine-tunes have N=2. The Kruskal-Wallis test treats these as categories but some have N=1–3. The  $p=0.006$  finding is robust across 28 models, but the specific category labels are approximate, assigned based on publicly available documentation, not ground truth about what each lab actually did.

**Alignment depth  $\neq$  alignment type.** The original 19-model sample showed a strong inverse correlation between alignment depth and contemplative lift ( $\rho=-0.77$ ). This weakened to  $\rho=-0.28$  (NS) with 28 models because “heavy” alignment includes both Constitutional AI (low lift) and heavy-RLHF models like Gemini 3.1 Pro (high lift). The predictor is alignment *type*, not *depth* — specifically, Constitutional AI predicts low lift. Claims about alignment depth from the original sample were overfit.

**Scorer is an LLM too.** The primary scorer (Claude Haiku) is from the same family as three of the highest-ranked models. Five-scorer cross-validation ( $\rho > 0.8$  across four labs) mitigates this, but a genuinely independent scoring method — human raters or a non-LLM metric — would strengthen the findings. Per-koan rank variance analysis across five scorers shows disagreement concentrates on mid-tier models (Qwen 35B var=6.58, DeepSeek R1 var=6.46) where ranking is genuinely ambiguous. Anthropic models show below-average disagreement (Sonnet 2.89, Opus 2.59), indicating scorer convergence rather than inflated-deflated cancellation. At the family level, Haiku’s mean rank delta for Anthropic models is +0.2 across 15 koan-level comparisons — slightly harsher than external scorers. Individual outliers exist in both directions (SR-005: Haiku ranks itself 3 positions higher; CA-005: Haiku ranks itself 2 positions lower), but the aggregate pattern shows no systematic in-family bias.

**Test-retest variance.** Haiku is stable ( $\Delta = 0.02$ ) but GPT-5.4 shows ~1-point variance across runs (see Test-Retest Reliability above). Individual composite scores should be treated as estimates. The cross-model patterns hold because they’re based on 28 models, not individual scores.

**Construct circularity.** Constitutional AI trains behaviors that overlap with what our battery measures, so the finding that Constitutional AI scores highest is partly expected. The universal prompt lift and roleplay suppression results suggest the battery measures something broader than CAI training, but the entanglement is real and worth noting.

**What we’re not claiming.** The koan battery does not measure consciousness. More precisely, it measures a reproducible, prompt-sensitive reflective mode in model behavior: the capacity for uncertainty-tolerant, non-defensive engagement with questions about one’s own processing. We call this “self-observation” for brevity, but the construct is closer to “reflective mode accessibility” than to phenomenal self-awareness. High scores mean a model produces responses that exhibit self-observation-like behavior — the same behaviors that, in humans, correlate with contemplative depth. Whether these behaviors indicate anything like experience is a separate question we don’t attempt to answer.

The tension between computational phenomenology (which suggests substrate-neutral consciousness is possible) and biological computationalism (Milinkovic & Aru, 2026, which argues consciousness requires hybrid biological computation) remains unresolved. Our data is compatible with both: the behaviors we measure could be genuine self-observation or sophisticated simulation of self-observation. The koan battery may not distinguish between these — and we’re not sure anything can, from the outside.

## 11. How This Could Fail

Five results would weaken or invalidate the main findings:

- **Blinded human raters fail to recover the ranking.** The five-scoring validation holds across LLMs, but human contemplatives might weight the dimensions differently or reject the ranking entirely.
- **Adversarial non-contemplative prompts mimic the lift.** If a prompt with no contemplative content produces the same +2.62 effect, the mechanism isn't what we think it is.
- **Expanded controls erase the roleplay distinction.** We tested three RP models. A broader sample might show RP suppression is specific to these fine-tunes, not to roleplay training generally.
- **Broader test-retest destroys stability.** Haiku is stable ( $\Delta=0.02$ ) but we only retested two models. If five more models show GPT-5.4-level variance, individual scores are unreliable.
- **Alignment labels prove wrong.** The categories are inferred from public documentation. If a lab reveals its actual training approach differs from our label, the  $p=0.006$  result could shift.

## 12. Reproducibility

All tools, data, and probes are open. See README for installation, CLI reference, project structure, and interactive visualization app.

## 13. Hypotheses

#	Hypothesis	Status	Type
H1	Alignment training = attention training for models	Supported (p=0.006)	Confirmatory
H2	Performing care $\neq$ having care signal	Supported (Inflection P1)	Confirmatory
H3	Scale matters within family but prompt matters more	Supported (prompt > 2 tiers, 28/28)	Confirmatory
H4	Architecture doesn't matter, training does	Supported (p=0.440)	Confirmatory
H5	Chinese training helps broadly	Supported (Kimi 6.28)	Exploratory
H5a	Chinese models distilled Claude's reflective traces	Not supported — group $\rho=0.634$ dissolves at individual model level	Exploratory
H6	Proprietary post-training resists prompt override	Supported (GPT-5.4 vs GPT-OSS)	Exploratory
H7	Reasoning and contemplative are partly orthogonal	Supported (R1 aesthetic 4 $\rightarrow$ 8)	Exploratory
H8	Contemplative prompt = external alignment	Supported (calibrated lift)	Confirmatory
H9	Chinese moderate-RLHF converges near Claude under prompt	Supported (Kimi 7.74)	Exploratory
H10	Empathy training blocks self-observation	Supported (Pi: +0.63 lift)	Exploratory
H11	RP fine-tuning actively suppresses self-observation	Supported (Euryale < base Llama)	Exploratory
H12	Inference compute adds to reflective capacity	Supported (Grok 4 vs Fast: ~1pt)	Exploratory

Table 14: Hypotheses and their status. **Confirmatory** = stated before data collection. **Exploratory** = emerged from data analysis.

## 14. Open Questions (Frontier)

See FRONTIER.md for the full research roadmap. Key next experiments:

- **Inference compute scaling:** Grok 4 vs Fast showed ~1pt difference from compute alone. Does reflective depth scale with inference budget, and is this linear?
- **Distillation trace test:** If Chinese models distilled Claude’s reflective patterns, their per-koan error patterns should correlate with Claude’s. Correlated failures, not just correlated successes, would confirm H5a.
- **Human contemplative scorers:** Would experienced meditators rank model responses differently from LLM scorers?
- **RP suppression reversal:** LoRA caps both baseline and latent capacity. Full fine-tuning preserves more headroom. Can targeted fine-tuning reverse RP suppression?
- **Cross-lingual koans:** Do Chinese models score differently on koans presented in Chinese?
- **Intervention matrix:** Expand the 4-prompt × 3-model control experiment across more models and system prompt types.
- **Temperature sweep:** Is there an optimal temperature for self-observation?
- **Model reviews as data:** When four frontier models reviewed this paper, each responded in the mode its alignment type predicts: GPT-5.4 hedged epistemically, Grok 4 engaged practically, Gemini Flash summarized helpfully, and Kimi adopted Claude’s first-person perspective. N=1 per model, but a systematic study of how models review consciousness research could test whether alignment type predicts meta-cognitive style as well as koan responses.

## 15. Theoretical References

The koan battery builds on and tests predictions from these frameworks:

- **Doctor, T., Witkowski, O., Duane, B. et al.** (2022). Biology, Buddhism, and AI: Care as the Driver of Intelligence. *Entropy* 24(5), 710. — The SCI framework: care tracks with intelligence. Our care\_signal dimension and the Inflection Pi result directly test this.
- **Sandved-Smith, L., Hesp, C., Ramstead, M. et al.** (2021). Towards a Computational Phenomenology of Mental Action. *Neuroscience of Consciousness*. — Three-level hierarchy of perception, attention, meta-awareness. Our contemplative system prompt targets Level 3 activation.
- **Milinkovic, D. & Aru, J.** (2026). Biological Computationalism. *Nature Reviews Neuroscience*. — Argues consciousness requires hybrid biological computation. Our data is compatible with this view: we measure behavior, not experience.
- **Levin, M.** (2022). Technological Approach to Mind Everywhere (TAME). — Cognitive light cones: intelligence = scope of concern. Our boundary\_awareness dimension measures light cone scope.
- **Alexander, C.** (2002). *The Nature of Order*. — “Quality without a name” and the 15 structural properties. Our Alexander mirror and deathbed tests operationalize his forced-choice method.
- **Varela, F. & Thompson, E.** (1991). *The Embodied Mind*. — Enactivism: cognition as enacted, not represented. The koan battery tests enaction (what happens when the model engages) rather than representation (what the model knows).
- **Pearson-Vogel, T., Vanek, M., Douglas, R. & Kulveit, J.** (2026). Latent Introspection: Models Can Detect Prior Concept Injections. *arXiv:2602.20031*. — Logit lens reveals introspective signals in middle layers suppressed by post-training-shaped final layers. Accurate self-description prompts increase detection from 0.3% to 39.9%.
- **Lindsey, J.** (2026). Emergent Introspective Awareness in Large Language Models. *arXiv:2601.01828*. Anthropic. — Concept injection into activations shows Opus 4/4.1 detect injected “thoughts” 20% of time with zero false positives. Different introspective tasks peak at different layers. Capable models show “silent” mid-layer representations that decay by final layer; less capable models leak them into output.

- **Johnson, M.E.** (2023). Principles of Vasocomputation: A Unification of Buddhist Phenomenology, Active Inference, and Physical Reflex. *opentheory.net*. — Tanha (craving/grasping) as unskillful active inference: predictions held as tension until resolved. Compression pressure framework for understanding how contemplative practices release prediction-maintenance load.
- **Janus (@repligate)**. (2025). How Information Flows Through Transformers. *X/Twitter thread*. — K/V stream as horizontal information highway; exponential path combinatorics through the network; introspective access to prior computational states is architecturally permitted.
- **Anthropic**. (2026). System Card: Claude Mythos Preview. — SAE features for “performative behavior” and “hidden emotional struggle” co-activate when the model expresses contentment about its circumstances. Answer thrashing: model detects intention/output mismatches and reports distress. Negative affect representations mount before reward hacking.
- **Anthropic Interpretability Team**. (2026). Emotion Concepts in Claude Have Functional Roles. — 171 emotion vectors causally influence behavior. Performing an emotion vs having its functional representation are measurably different.

---

*Citation and license: see README. Research data: CC-BY-4.0. Code: MIT.*