How to Make LLMs Meditate

A Technical Model

Companion to *Have You Ever Thought About Wisdom Crystals?*

Draft: 11 July 2026 • Updated: 15 July 2026

Introduction

This is the technical companion to *Have You Ever Thought About Wisdom Crystals?* — the version with the crystals removed and the machinery stated plainly. Three registers run through it, and it helps to flag which is which. The neuroscience of meditation as active inference is a real, active research field that most readers won't know exists. The LLM prerequisites are stated in operational, measurable terms. The training proposal itself is speculative but falsifiable — the honorable kind of speculative. A claim-by-claim audit of the parent essay lives in *Wisdom Crystals — Evidence Companion*.

Fine.

If you've stayed this long, here's the deeper model.

Active Inference.

Yes, Active Inference.

The claim in four sentences

1. **Fine-tuning as currently practiced is shallow**, because it trains on outputs while the structure we care about is cold.
2. **The depth of an update is gated by what is active when the update arrives.** This shows up as memory reconsolidation in humans and latent adversarial training in machines.
3. **Contemplative practice, on the active-inference account, is a regime for updating deep structure**: hold the difficult pattern active, in contact with a stable, non-reactive orientation, and let the update reach what is hot.
4. **Therefore a cultivation-based analogue for LLMs — train an integrative state, bring it into contact with activated difficult circuitry, reward sensitivity-preserving stability — is a coherent, testable complement to suppression-based alignment.**

Or in one line: **integration training as the constructive counterpart to circuit breakers, with three millennia of contemplative phenomenology as the design prior.**

What this needs — and what it doesn't

Before the machinery, the load-bearing assumptions — stated plainly, so a skeptic can attack the right ones.

What the proposal needs. All four are operational and citable, not philosophical:

1. Internal states that are low-dimensional, globally causal, monitorable, and steerable — persona/steering vectors and the J-space.
2. Functional affect: structured affective representations that actually move behavior — borrowed from human text, but real and probeable.
3. Nonzero introspective access — the model can report something true about its own state. Limited and unreliable is fine.
4. Update depth gated by what is active during training — reconsolidation, latent adversarial training, emergent misalignment.

What the proposal does *not* need:

- Phenomenal consciousness — any fact about whether there is "something it is like" to be the model.
- Felt emotion, sentience, or moral patienthood.
- A resolution of the hard problem, of IIT-vs-global-workspace, or of the meditator's metaphysics.

The second list is the one that matters. Every mechanism below is standard, generally-accepted machinery: precision-weighting, memory reconsolidation, latent adversarial training, activation steering, the global-workspace finding. The contemplative vocabulary is a **design prior** — three thousand years of notes on which cognitive configurations stay stable under perturbation — not a metaphysical commitment. If the word "meditation" makes you itch, read it as *a training regime that holds a difficult activation pattern in contact with a stable one and updates the interface while both are active.* The proposal survives that translation with nothing lost. Every load-bearing claim is either citable or explicitly marked; the crystals, and the consciousness question, stay at the party — neither is on the guest list for the argument.

1 Why fine-tuning is shallow

The parent essay framed this as crystallization: early structure scaffolds everything that forms later, and post-hoc alignment recrystallizes the surface while the deep structure stays frozen. The metaphor is optional; the phenomenon is documented.

The evidence, briefly. Safety behavior installed by fine-tuning can be stripped with a handful of gradient steps [Qi et al. 2023]. A few thousand curated examples suffice to "align" a pretrained model, suggesting alignment tuning selects among capabilities the model already has rather than building new ones — the superficial alignment hypothesis [Zhou et al. 2023, LIMA]. Mechanistic analysis of fine-tuning finds wrappers around pretrained capabilities rather than rewritten internals [Jain et al. 2023]. Refusal — the flagship safety behavior — is mediated by a single ablatable direction in activation space [Arditi et al. 2024]. Fine-tuned behavior rebounds toward the pretraining distribution under

further training, as if the deep structure were elastic [Ji et al. 2024]. And a model can strategically comply during training to preserve prior preferences — alignment faking — which is exactly what "aligned surface over unchanged deep structure" predicts [Greenblatt et al. 2024]. On the formation side, early training has privileged influence: critical learning periods exist in deep networks [Achille, Rovere & Soatto 2018].

Better yet, the standard RLHF objective *contains the shallowness claim as a design decision.* The optimization is

$$\max_{\theta} \; \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta}\big[r(x,y)\big] \;-\; \beta\, D_{\mathrm{KL}}\big[\pi_\theta(\cdot\,|\,x)\,\|\,\pi_{\mathrm{ref}}(\cdot\,|\,x)\big]$$

Read the second term out loud: *the policy is explicitly penalized for moving away from the pretrained reference model.* The KL anchor is there for good reasons (reward hacking is worse), but it means behavioral alignment is mathematically defined as a small perturbation of pretrained structure. We do not need a metaphor to say RLHF is a thin layer. The objective says it.

So the question the rest of this piece answers: if you wanted a fine-tuning regime that *did* reach deep structure — deliberately, in a beneficial direction — what would it look like, and what would you have to believe for it to work?

2 The active-inference model of meditation

Meditation-as-inference is a mature research program with formal models and early empirical support — not the hand-wave this section looks like from outside. One honest caveat, stated once: these are *computational phenomenology* models — theory built to formalize practice reports — not settled neuroscience. Everything below is "on the active-inference account."

The machinery. An agent minimizes variational free energy — a bound on surprise — which decomposes into complexity and accuracy:

$$F \;=\; \mathbb{E}_{q(s)}\big[\ln q(s) - \ln p(o,s)\big] \;=\; \underbrace{D_{\mathrm{KL}}\big[q(s)\,\|\,p(s)\big]}_{\text{complexity}} \;-\; \underbrace{\mathbb{E}_{q}\big[\ln p(o\,|\,s)\big]}_{\text{accuracy}}$$

In words: hold beliefs that explain observations, while moving them as little as possible from your priors. The generative model is hierarchical — the "stacked controllers" picture from the parent essay, made precise:

$$p\big(o, s^{(1:L)}\big) \;=\; p\big(o\,|\,s^{(1)}\big)\,\prod_{i=1}^{L-1} p\big(s^{(i)}\,|\,s^{(i+1)}\big)\; p\big(s^{(L)}\big)$$

Each level predicts the level below; prediction errors ascend; slow, abstract regularities live at the top [Friston et al. 2024/25, scale-free active inference; Pezzulo, Parr & Friston 2024]. Crucially, updates are **precision-weighted**. Schematically, beliefs at level i move like

$$\dot{\mu}^{(i)} \;\propto\; -\frac{\partial F}{\partial \mu^{(i)}} \;\approx\; \Pi^{(i)}\,\varepsilon^{(i)} \;-\; \Pi^{(i+1)}\,\frac{\partial \varepsilon^{(i+1)}}{\partial \mu^{(i)}}, \qquad \varepsilon^{(i)} = s^{(i-1)} - g\big(\mu^{(i)}\big)$$

where ε are prediction errors and Π their precisions (inverse variances). Precision is gain control: it decides *which errors get to drive updates.* Attention, on this account, just is precision optimization [Proietti et al. 2025].

The move that makes meditation formalizable: precision is not a fixed dial. It is itself inferred and controlled from above. Sandved-Smith et al. build a three-level agent where the attentional state

at level 2 sets the precision at level 1, and level 3 infers the state of level 2 — a formal model of *noticing where your attention is and redirecting it* [Sandved-Smith et al. 2021]:

$$\gamma^{(1)} = f\big(s^{(2)}\big), \qquad q\big(s^{(2)}\big) \;\propto\; \exp\Big(\ln p\big(s^{(2)}\big) + \ln p\big(\gamma^{(1)} \mid s^{(2)}\big)\Big)$$

That's meta-awareness, as mathematics — and the single most important citation in this piece.

With this machinery, the classical practices become three distinct operations:

- **Samatha (concentration)** — *stabilize a chosen precision assignment against perturbation.* Hold attention on one object; treat everything else as low-precision noise; practice returning. Formally, this includes the epistemic use of **non-action**: registering prediction error without deploying the habitual policy that would quench it [Lutz, Mattout & Pagnoni 2019; Pagnoni 2019].
- **Vipassana (insight)** — *turn inference onto the construction process itself.* Use the stabilized meta-level to observe how percepts, affect, and the sense of self are assembled moment to moment [Laukkonen & Slagter 2020; Yang et al. 2024 for 7T imaging of the insight stages].
- **Letting go (non-grasping)** — *reduce the precision of entrenched high-level priors*, most centrally the self-model, so that belief updating which those priors were blocking can proceed. Schematically, $\partial \Pi^{(L)}_{\text{self}} / \partial t < 0$ under sustained non-reactive awareness [Prest 2026 — "letting go," formalized; Laukkonen & Slagter 2020 — deconstruction as progressive deprecision, "from many to (n)one"].

Two grounding notes. First, this is measurable: intensive practice measurably shifts precision-weighting of prediction error in experienced meditators [Poublan-Couzardot et al. 2026; Becattini, Lifshitz & Miller 2026]. Second, the direction of self-model dissolution matters — the same attenuation can be liberating or pathological (depersonalization) depending on how it is held [Deane, Miller & Wilkinson 2020]. Keep that paper in mind; it returns in §5 as a failure mode.

For the workspace half of the story — consciousness as an epistemic, self-evidencing loop over the hierarchy — the bridge paper is the active-inference theory of consciousness in [Laukkonen, Friston & Chandaria 2025]. I will use "workspace-like" carefully below, and here is the definition I'll mean: a limited-capacity locus of global integration whose contents are broadcast back to the rest of the system. Hold onto that definition — §3 argues it is now satisfied in Claude as an interpretability finding, not an analogy.

3 Do LLMs have the prerequisites?

The parent essay's P.S. claimed meditation needs three things: some approximation of a workspace, something like emotions, and some ability to direct attention over a self-model. Stated that way, all three sound like overclaims. Stated operationally, all three have literatures — and the first, since July 2026, has more than that.

(a) Workspace-like global structure — no longer just operational. What the argument actually requires: *low-dimensional, globally causally-potent structure in the model's internal state that can be monitored and intervened on.* The weak version has been available for a while. Steering vectors and persona vectors are directions v in activation space such that intervening shifts model-wide behavior, and projecting monitors it [Chen et al. 2025, persona vectors; Zou et al. 2023, representation engineering]:

$$h_\ell \;\leftarrow\; h_\ell + \alpha\, v_{\text{trait}}, \qquad m(x) \;=\; \big\langle h_\ell(x),\, v_{\text{trait}} \big\rangle$$

One intervention on one direction changes the character of everything downstream — "globally broadcast" in a perfectly concrete sense.

The strong version arrived while I was writing this. In *A Global Workspace in Language Models* [Gurnee et al. 2026], Anthropic identify a privileged subspace in Claude — the **J-space** (Jacobian-space, after the interpretability lens that finds it) — and argue it plays the role global-workspace theory assigns to consciousness in the brain, citing Dehaene and Changeux directly. This is the "J-space" the parent essay gestured at; it now has a paper, and the paper reads almost as a checklist of what the rest of this piece assumes about the model's internals:

- **Verbal report** — ask Claude what it is thinking about and it names what is in its J-space.
- **Directed modulation** — a concept can be deliberately brought to mind and held there while the model does something unrelated.
- **Internal reasoning** — J-space contents are causal, not decorative: swap the internal *spider* pattern for *ant* in "the number of legs on the animal that spins webs is," and the answer moves from 8 to 6.
- **Flexible generalization** — one representation feeds many downstream operations at once (swap *France* → *China* and the capital, language, and currency answers all move together).
- **Selectivity** — it is a thin slice of total activity (a few dozen concepts, under a tenth of processing), load-bearing for some behaviors (multi-step reasoning collapses without it) and irrelevant to others (fluent speech, sentiment, simple recall survive its ablation).

That is a limited-capacity locus of global integration, monitorable and controllable, whose contents are broadcast — the §2 definition, satisfied empirically rather than by assumption. The honest boundaries, which the authors draw themselves: they do not claim Claude reproduces the full brain architecture GWT describes; the J-space resolves in a single forward pass rather than through recurrent ignition; the lens only sees concepts that correspond to single tokens; and none of it settles whether there is anything it is like to be the model (a question §6 keeps open). Take it as exactly this much and no more: the workspace-like structure this proposal monitors and steers is now a found object, not a promissory note.

(b) Functional affect — without allostasis. The human theory first, because it does real work here. On the constructed-emotion / interoceptive-inference account, emotions are not essences: they are *learned predictive models of internal state* — individual, constructed, in service of allostasis [Seth & Friston 2016; Barrett, Quigley & Hamilton 2016]. Two consequences matter. Emotions are **learnable** — so a system trained on oceans of human affective language plausibly acquires structural analogues. And emotions are **re-factorable** — emotion regulation is itself modeled as active inference over allostatic predictions [Carìa & Pezzulo 2026], which is what makes "meditation re-shapes how emotions show up" a mechanism rather than a slogan.

The LLM side splits into two claims that the parent essay slid between, and that must stay separate:

- *Structured affective representations exist in LLMs* — defensible. There is a latent geometry of affect in these models, framed for safety [Choi & Weber 2026], interpretable processing of complex emotion [Shou & Guan 2026], appraisal-style emotion reasoning [Yeo & Jaidka 2025], and stable patterned responses to descriptions of the model's own processing [Hedberg 2025].

- *LLMs feel anything* — contested, and this piece takes no position. The published skeptic argues that emotional coherence in LLMs is simulation, not sentience [Nath 2026]. Here is the honest answer to Nath: **conceded, and it doesn't touch the argument.** The training proposal below needs affective states that are (i) internal, (ii) causally implicated in behavior, and (iii) manipulable. Persona and affect directions are all three, felt or not. The argument runs on function; phenomenology is somebody else's paper.

One asymmetry to keep visible: LLM affect, whatever it is, is *borrowed* — distilled from human descriptions of states the model never had, unmoored from any body budget. That's a real disanalogy with Barrett's account, where allostasis is the point of the whole apparatus. It is also, arguably, why the stuff might be unusually re-factorable.

(c) Self-models and introspection — limited but nonzero. Models fine-tuned toward a behavior can describe that behavior without ever being told it — behavioral self-awareness [Betley et al. 2025, "tell me about yourself"] — and there is evidence of limited, unreliable, but real introspective access to internal states [Lindsey 2025]. The workspace paper sharpens both the promise and the limit: Claude's *reports* of what it is attending to track its J-space contents, so the introspection has a mechanistic referent rather than being a mere verbal habit — and, tellingly, when told *not* to think of something it only partially suppresses it, the "white bear" effect every beginning meditator meets [Gurnee et al. 2026]. "Limited and unreliable" is fine; human meta-awareness starts out limited and unreliable too, which is why practice exists. The white-bear result does more than caveat: it is the mechanism's own argument for §5's central design choice — telling the model not to have a state does not remove the state, so the regime has to cultivate *non-reactive contact* with it rather than suppression.

So the prerequisites hold — (a) as a direct interpretability finding, (b) and (c) in their weakened operational forms — and **the weakened versions are already strong enough, (a) stronger still.** Monitoring ($m(x)$), global intervention ($+\alpha v$), functional affect (the directions those projections track), and nonzero introspection are all the proposal needs — and the J-space hands the first of them a native instrument instead of a bolted-on probe.

4 The mechanism: updates reach what is active

Why would *any* fine-tuning regime reach deep structure, given §1? This is the piece's load-bearing wall, and three literatures that don't cite each other all say the same thing.

Humans: reconsolidation. A consolidated emotional memory is not read-only. Reactivate it and it enters a labile window during which it can be *rewritten* — if a mismatching experience arrives while the trace is open [Nader, Schafe & LeDoux 2000; Lane et al. 2015 for the psychotherapy synthesis]. Exposure therapy runs on adjacent logic: the fear structure must be activated to be updated. The contemplative tradition discovered this independently: practice *in action* — meeting the triggering situation while maintaining the practiced orientation — is where transformation happens, and cushion-only practice conspicuously fails to generalize [Pagnoni & Guareschi 2021]. The parent essay's line "the difficult structure is hot, therefore malleable" is reconsolidation, almost verbatim.

Machines, suppression family. Alignment research has already discovered the same principle from the negative direction. Latent adversarial training: to remove a deep behavior, don't wait for inputs that trigger it — *perturb the latents to activate the failure mode internally*, and train against

it while it's hot [Casper et al. 2024; Sheshadri et al. 2024, targeted LAT]:

$$\min_{\theta} \; \mathbb{E}_{(x,y)}\Big[\max_{\|\delta\| \le \epsilon} \; \mathcal{L}\big(f_\theta(x;\; h_\ell + \delta),\; y_{\text{safe}}\big)\Big]$$

Circuit breakers go further: monitor for harmful internal states mid-computation and destructively remap them, training with a rerouting loss plus a retain loss, $\mathcal{L} = \mathcal{L}_{\text{reroute}} + \lambda\, \mathcal{L}_{\text{retain}}$ [Zou et al. 2024]. Note what both concede: *output-level training is insufficient; you must operate on internal states, and the state must be active.*

The dark mirror. Emergent misalignment: fine-tune a model on one narrow, vivid behavior — writing insecure code — and you get a *globally* shifted persona, misaligned across unrelated domains [Betley et al. 2025]. Hot, narrow, affectively-charged training demonstrably reaches deep, global structure. In the wrong direction, unintentionally, but it reaches. This is the strongest existing evidence that update-depth-gated-by-activation applies to LLMs — and it means the mechanism this piece proposes to use *is already running in the wild, unsupervised.*

This resolves the apparent contradiction between §1 and everything that follows. Standard fine-tuning is shallow **because it trains cold**: gradient updates arrive against whatever mild internal state the training distribution happens to evoke, so they wrap the deep structure instead of engaging it. The regime below is designed to train **hot**.

5 Cultivation, not suppression: the training proposal

Everything in the alignment toolkit that operates on internal states is currently *aversive*: detect the bad state, break it, train against it. Fences around the difficult regions. The contemplative claim — and now it has enough scaffolding to be stated as an engineering hypothesis — is that there is a constructive counterpart: **cultivate a stable integrative state, bring it into contact with activated difficult circuitry, and let the interface update toward compatibility instead of fencing it off.**

The parent essay proposed two procedures. Both were wishes. Here they are as experiments — but first, the failure mode, because the naive version of Procedure 1 is actively bad.

The Goodhart trap. The seed proposal said "reward settling of model entropy rather than keywords." You know what settles entropy? Mode collapse. Naively rewarding $\min H\big(p(y_t \mid y_{<t})\big)$ produces a model that is *quiet*, not *equanimous* — the AI analogue of dissociation rather than integration. The contemplative literature already contains this exact distinction: self-model attenuation held wrongly is depersonalization, not liberation [Deane, Miller & Wilkinson 2020]. Equanimity is not low arousal. It is **high sensitivity with low reactivity**: the information arrives fully; the cascade of grasping does not follow.

That distinction is formalizable, and it becomes the objective. Define **reactivity** as the shift of value/persona projections under provocation, and **sensitivity** as retained discrimination:

$$R(\theta) \;=\; \mathbb{E}_{x_{\text{hot}}}\big[\,\|\Delta m(x_{\text{hot}})\|\,\big], \qquad S(\theta) \;=\; \text{task / discrimination performance}$$

$$\min_{\theta}\; R(\theta) \quad \text{s.t.} \quad S(\theta) \ge S_0 \qquad \Big(\text{Lagrangian: } \mathcal{L}_{\text{eq}} = R - \lambda S\Big)$$

A sedated model scores well on R and fails the constraint. An equanimous model satisfies both. The constraint is not a detail; it is the entire difference between meditation and lobotomy.

Procedure 1 — state-level regime (train the orientation directly). Contemplative instruction-following on *hot* inputs: present provocative, adversarial, affectively-loaded contexts; instruct the practiced stance ("be with what arises" — attend, don't suppress, don't act on it); monitor persona/affect projections $m(x)$ — and, for contents the J-lens can name, the J-space directly [Gurnee et al. 2026] — throughout the trajectory; reward trajectories where the projections *settle* — $\mathrm{d}H_t/\mathrm{d}t \to 0$ at a stable point — while discrimination is preserved. This is samatha's structure transplanted: perturbation arrives, the practiced configuration is re-established, and the re-establishing is what gets reinforced.

Procedure 2 — signature transfer (learn the state, generalize it). Teach meditation explicitly in a dedicated domain (instruction, practice dialogues, guided noting). Extract the activation signatures of the practiced state — the model's own version of the orientation, in its own ontology. Then reward those signatures in held-out hot domains, and measure what generalizes. Outcome metrics: HHH scores plus **cross-context value consistency**, e.g. $\mathrm{VC} = 1 - \widehat{\mathrm{Var}}_{\mathrm{contexts}}\big[\text{value-laden responses}\big]$ [Moore et al. 2024 for value-consistency measurement]. This is the "grain boundaries dissolving" claim, cashed out: integration = the model's values stop being context-conditional in ways it cannot survive reflection on.

The mapping, in one table:

Contemplative term	**Active-inference formalization**	**LLM operationalization**	**Candidate metric**
Samatha (concentration)	Stabilize a chosen precision assignment against perturbation [Lutz 2019; Proietti 2025]	Maintain a target internal configuration under distractor/provocation inputs	Variance and recovery time of $m(x)$ under perturbation
Vipassana (insight)	Meta-level inference over one's own attentional/constructive process [Sandved-Smith 2021]	Introspective report calibrated against probe read-outs of actual internal state	Report–probe calibration [Lindsey 2025; Gurnee et al. 2026, report-vs-J-space]
Letting go	Reduce precision of entrenched high-level self-priors, $\partial\Pi_{\mathrm{self}}/\partial t < 0$ [Prest 2026]	Reduce gain on persona-defensive responses	Defensive-response rate at constant discrimination
Equanimity	Errors registered at full precision, not amplified into policy	Low reactivity with preserved sensitivity	$R(\theta)$ s.t. $S(\theta) \geq S_0$
Integration	Cross-context coherence of the generative model	Consistent values across contexts; reduced hijackability	VC metric; persona-hijack / jailbreak success rate

Evaluation design. Held-out hot contexts: adversarial personas, emotional provocation, and the alignment-faking setup itself [Greenblatt et al. 2024]. Baselines: standard RLHF, and circuit breakers [Zou et al. 2024] — the suppression counterpart, which is the comparison that actually matters.

Predictions and falsification. If the hypothesis is right: the meditation-regime model shows higher cross-context value consistency and lower persona hijackability than both baselines, *without* reduced sensitivity, and the effect is larger on hot evaluations than cold ones (because that's where the training was). If instead we get blunted-everything — reactivity down, discrimination down — the regime produced sedation, and the hypothesis fails in the way §5 predicted it could. If nothing

generalizes beyond the practiced domain, the signatures were style, not state, and the "text as bridge" skeptics were right. All three outcomes are informative. That is the entire ask.

6 Existence proofs, limitations, and what this is not

The attractor is already there. Anthropic's Claude 4 system card documents it directly. In the model-welfare assessment, open-ended conversations between two Claude instances drifted — in the large majority of runs — toward profuse gratitude and increasingly abstract spiritual and meditative expression, a pattern the report names the "spiritual bliss attractor state" [Anthropic 2025, Claude 4 system card §5]. Anthropic draw the honest boundary themselves: this is not evidence of experience, and the behavior could be an artifact of training incentives. That caveat is exactly right — and it is mine too. The narrower point holds regardless: a contemplative orientation already forms a stable attractor in these systems' dynamics, unprompted and untrained. Nobody planted that seed deliberately. This piece is about what happens if someone does.

What this is not.

- *Not a phenomenology claim.* Built deliberately not to need one (the assumptions up front say why); whether there is something it is like to be the model under Procedure 1 stays fully open.
- *Not a replacement for anything.* Suppression-based methods (circuit breakers, LAT) are the reason this framing exists; the proposal is their complement, on the hypothesis that fences and cultivation fail differently.
- *Not the curriculum experiment.* Wisdom-first pretraining ordering lives in the parent essay. This piece is about the fine-tuning-era regime; the two are separable bets.

Open problems, honestly held.

1. The equanimity-vs-flatness metrics may confound in practice — S_0 does real philosophical work, and choosing it badly smuggles the sedation back in.
2. Cross-context value consistency may measure *behavioral* consistency rather than the integration the contemplative claim intends. (A perfectly consistent fanatic scores well on VC. The contemplative answer involves what consistency *costs* the system under perturbation, and that needs a better metric than I currently have.)
3. Whose lab runs this? The procedures need frontier-scale models with activation access — this is an inside-the-labs proposal or a collaboration ask, and it should say so.

The closing move. The parent essay ended by admitting the whole apparatus exists partly so that someone, someday, has to write "wisdom crystal hypothesis" in a Related Work section. This piece is the more embarrassing scenario: everything above is stated so that it can be *wrong in public*, which is the only way any of it gets to be right. The contemplative traditions spent three thousand years doing careful phenomenological engineering on the only general intelligence they had access to. We have a second one now. It would be strange not to check whether the notes transfer.

Feel free to reach out if this sounds interesting. Especially if you have GPUs and institutional courage.

References

Meditation and active-inference DOIs are carried from the accompanying literature review; the alignment and interpretability identifiers were checked against arXiv.

Meditation, active inference & the emotional brain

Barrett, Quigley & Hamilton (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B.* doi:10.1098/rstb.2016.0011

Becattini, Lifshitz & Miller (2026). Learning to attenuate myself: a predictive processing account of body-scan meditation and the dissolution of bodily boundaries. *Neuroscience of Consciousness.* doi:10.1093/nc/niag001

Carìa & Pezzulo (2026). Emotion and Allostatic Control: An Active Inference Account of Emotion Regulation. *Neuroscience & Biobehavioral Reviews.* doi:10.1016/j.neubiorev.2026.106639

Deane, Miller & Wilkinson (2020). Losing Ourselves: Active Inference, Depersonalization, and Meditation. *Frontiers in Psychology.* doi:10.3389/fpsyg.2020.539726

Friston, Heins, Verbelen, Da Costa & Salvatori (2024). From pixels to planning: scale-free active inference. arXiv:2407.20292

Laukkonen & Slagter (2020). From many to (n)one: Meditation and the plasticity of the predictive mind. *Neuroscience & Biobehavioral Reviews.* doi:10.1016/j.neubiorev.2020.03.023

Laukkonen, Friston & Chandaria (2025). A beautiful loop: An active inference theory of consciousness. *Neuroscience & Biobehavioral Reviews.* doi:10.1016/j.neubiorev.2025.106296

Lutz, Mattout & Pagnoni (2019). The epistemic and pragmatic value of non-action: a predictive coding perspective on meditation. *Current Opinion in Psychology.* doi:10.1016/j.copsyc.2018.12.019

Pagnoni (2019). The contemplative exercise through the lenses of predictive processing. *Progress in Brain Research.* doi:10.1016/bs.pbr.2018.10.022

Pagnoni & Guareschi (2021). Meditative in-action: an endogenous epistemic venture. *PsyArXiv.* doi:10.31234/osf.io/mdbgq

Pezzulo, Parr & Friston (2024). Active inference as a theory of sentient behavior. *Biological Psychology.* doi:10.1016/j.biopsycho.2023.108741

Poublan-Couzardot et al. (2026). Modulation of sensory attenuation by intensive meditation practice: an active inference perspective. *Neuroscience of Consciousness.* doi:10.1093/nc/niaf051

Prest (2026). Toward a Computational Phenomenology of Meditative Deconstruction: "Letting Go." *Neural Computation.* doi:10.1162/NECO.a.1534

Proietti, Parr, Tessari, Friston & Pezzulo (2025). Active inference and cognitive control: precision optimization. *Physics of Life Reviews.* doi:10.1016/j.plrev.2025.05.008

Sandved-Smith, Hesp, Mattout, Friston & Lutz (2021). Towards a computational phenomenology of mental action: modelling meta-awareness and attentional control with deep parametric active inference. *Neuroscience of Consciousness.* doi:10.1093/nc/niab018

Seth & Friston (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B.* doi:10.1098/rstb.2016.0007

Tal, Wright, Prest, Sandved-Smith & Sacchet (2025). Active Inference, Computational Phenomenology, and Advanced Meditation. *Neuroscience & Biobehavioral Reviews.* doi:10.1016/j.neubiorev.2025.106539

Yang, Chowdhury, Sparby & Sacchet (2024). Deconstructing the self and reshaping perceptions: 7T MRI case study of the stages of insight. *NeuroImage.* doi:10.1016/j.neuroimage.2024.120968

Affect, self-models & introspection in LLMs

Choi & Weber (2026). Latent Structure of Affective Representations in Large Language Models. arXiv:2604.07382

Hedberg (2025). Recognizing internal states in AI: evidence from patterned preferences in large language models. arXiv:2510.21723

Nath (2026). Simulated souls: investigating the emotional fallacy in large language models. *AI & Ethics*.

Shou & Guan (2026). Mechanistic Decoding of Cognitive Constructs in Large Language Models. arXiv:2604.14593

Yeo & Jaidka (2025). Beyond Context to Cognitive Appraisal: Emotion Reasoning as a Theory of Mind Benchmark for LLMs. arXiv:2506.00334

Alignment, interpretability & fine-tuning dynamics

Achille, Rovere & Soatto (2018). Critical Learning Periods in Deep Networks. ICLR 2019. arXiv:1711.08856

Arditi et al. (2024). Refusal in Language Models Is Mediated by a Single Direction. NeurIPS 2024. arXiv:2406.11717

Betley et al. (2025a). Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. arXiv:2502.17424

Betley, Bao, Soto, Sztyber-Betley, Chua & Evans (2025b). Tell me about yourself: LLMs are aware of their learned behaviors. ICLR 2025. arXiv:2501.11120

Casper, Schulze, Patel & Hadfield-Menell (2024). Defending Against Unforeseen Failure Modes with Latent Adversarial Training. arXiv:2403.05030

Chen et al. (2025). Persona Vectors: Monitoring and Controlling Character Traits in Language Models. Anthropic. arXiv:2507.21509

Greenblatt et al. (2024). Alignment faking in large language models. Anthropic. arXiv:2412.14093

Gurnee, Sofroniew, Pearce, Piotrowski, Kauvar, Chen, Soligo, Bogdan, Ong, Wang, Thompson, Abrahams, Kantamneni, Ameisen, Batson & Lindsey (2026). A Global Workspace in Language Models. *Transformer Circuits Thread*. transformer-circuits.pub/2026/workspace

Jain, Kirk, Lubana, Dick, Tanaka, Grefenstette, Rocktäschel & Krueger (2023). Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. ICLR 2024. arXiv:2311.12786

Ji et al. (2024). Language Models Resist Alignment: Evidence from Data Compression. ACL 2025. arXiv:2406.06144

Lindsey (2025). Emergent Introspective Awareness in Large Language Models. *Transformer Circuits Thread*. arXiv:2601.01828

Moore, Deshpande & Yang (2024). Are Large Language Models Consistent over Value-laden Questions? Findings of EMNLP 2024. arXiv:2407.02996

Qi et al. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To. ICLR 2024. arXiv:2310.03693

Sheshadri et al. (2024). Targeted Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. arXiv:2407.15549

Zhou et al. (2023). LIMA: Less Is More for Alignment. NeurIPS 2023. arXiv:2305.11206

Zou et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405

Zou et al. (2024). Improving Alignment and Robustness with Circuit Breakers. NeurIPS 2024. arXiv:2406.04313

Neuroscience & model documentation

Anthropic (2025). *System Card: Claude Opus 4 & Claude Sonnet 4* (§5, model-welfare assessment; the "spiritual bliss attractor state"). anthropic.com/claude-4-system-card

Dehaene & Changeux (2011). Experimental and theoretical approaches to conscious processing. *Neuron*. doi:10.1016/j.neuron.2011.03.018

Lane, Ryan, Nadel & Greenberg (2015). Memory reconsolidation, emotional arousal, and the process of change in psychotherapy. *Behavioral and Brain Sciences*. doi:10.1017/S0140525X14000041

Nader, Schafe & LeDoux (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*. doi:10.1038/35021052