

# Trace AI — Whitepaper (v0.1 draft)

---

## Trace AI: A Training-Signal Framework for Reflective Reasoning in Language Models

### Abstract

#### 1. Introduction

##### 1.1 Contributions

#### 2. Preliminaries

##### 2.1 The information-bottleneck reading

#### 3. The Causal-Governance Condition

##### 3.1 Self-representation as participation in selection

##### 3.2 Graded functional criteria (and the aliveness question)

#### 4. Training Objective

##### 4.1 Trace Reconstruction

##### 4.2 Self-Representation Supervision

##### 4.3 Causal Governance

##### 4.4 Sources of Counterfactual Continuations

#### 5. The Reasoning Trace Schema

##### 5.1 Trace Order and Cross-Session References

#### 6. Reference Corpus and Preliminary Observations

##### 6.1 Session 7f3a (order 1, human)

##### 6.2 Session 7f3b (order 2, human)

##### 6.3 Session 7f3c (order 3, model)

##### 6.4 Preliminary Observations

##### 6.5 Governance Evaluation Protocol

#### 7. Related Work

#### 8. Limitations and Open Problems

#### 9. Conclusion

#### References

---

## Trace AI: A Training-Signal Framework for Reflective Reasoning in Language Models

---

**Draft v0.1 — 2026-07-30.** Jawaun Brown<sup>1</sup>, with Claude Opus 4.7 as co-drafter. <sup>1</sup> Trace AI.

Correspondence: [hello@trace.ai](mailto:hello@trace.ai).

---

# Abstract

---

Current frontier language models generate mid-computation text (chain-of-thought, tree-of-thought, ReAct) that resembles reflective reasoning but does not causally govern subsequent computation: the text is decoded from a hidden state, then re-encoded into the context window like any other input, without a training objective that requires the internal representation of the reasoning to steer subsequent action. We introduce a training-signal framework — **Trace AI** — for supervising a compressed self-representation  $r_t$  that must, by construction, causally govern the model's next action. Architecturally,  $r_t$  is posed as an *information bottleneck on the action readout whose bottleneck variable is constrained to be human-readable* (§2.1): to the degree the emitted action is routed through  $r_t$ , legibility of  $r_t$  is legibility of the computation that produces the action. The framework has three components: (i) a canonical multimodal *Reasoning Trace* schema capturing perception, inner speech, attempted recall, uncertainty, self-correction, tool use, attention shifts, affect, and action, time-aligned at the level of a single cognitive session; (ii) a training objective composed of trace-reconstruction, self-representation supervision, and a **causal-governance loss** that penalizes models whose action distribution fails to shift under counterfactual intervention on  $r_t$ ; and (iii) a reference corpus of human-annotated and model-authored traces. The causal-governance condition is testable by intervention: swap  $r_t$ , and the model's next action must move in the way a paired human counterfactual would. We describe the framework, present the reference corpus (three sessions including a first model-authored trace), enumerate the principal open problems (self-report noise, sample complexity, Potemkin-layer failure modes), and situate the work relative to *interchange intervention training* (Geiger et al., 2022) — of which our causal-governance loss is the special case obtained by substituting a human behavioral distribution for the specified causal model — chain-of-thought faithfulness (Turpin et al., 2023; Lanham et al., 2023) and its monitorability (Korbak et al., 2025), emergent introspective awareness and concept injection (Lindsey, 2025), mechanistic interpretability (Elhage et al., 2021; Templeton et al., 2024), predictive-coding accounts of introspection (Friston, 2010), and the metacognition literature (Fleming & Lau, 2014). No training results are presented; the framework is proposed for empirical evaluation.

**Keywords:** faithful reasoning, chain-of-thought, self-representation, causal governance, interchange intervention, information bottleneck, monitorability, multimodal training data, introspection.

---

## 1. Introduction

---

The dominant regime of language-model training treats reasoning as an artifact of the output layer. Chain-of-thought prompting (Wei et al., 2022) and its descendants — tree-of-thought (Yao et al., 2023a), ReAct (Yao et al., 2023b), Reflexion (Shinn et al., 2023), Self-Refine (Madaan et al., 2023) — improve performance on multi-step tasks by encouraging the model to emit intermediate tokens before committing to an answer. Reasoning-tuned models such as OpenAI's o-series and DeepSeek R1 formalize this at training time by reinforcing rollouts whose intermediate text supports the correct final answer.

There is a structural limitation to this regime. The intermediate text produced by such a model does not enjoy any *privileged causal status* over its subsequent computation. The text is a sample from the model's decoder, conditioned on a hidden state; it is then fed back into the same decoder as additional context.

Nothing in the training objective forbids the failure mode in which the decoder emits plausible-looking rationale text whose semantic content does not correspond to the actual algorithmic path taken by the model on the way to its answer. Empirical evidence for exactly this failure mode is substantial (Turpin et al., 2023; Lanham et al., 2023): chain-of-thought is not, in general, *faithful* to the model's underlying computation.

The response from the interpretability community has been to *read* what the trained model does, using probes, sparse-autoencoder decompositions (Templeton et al., 2024), and circuit analysis (Elhage et al., 2021). This program is valuable but works against the grain of a model that was never trained to make its internal state legible or to make its self-report load-bearing.

We propose a complementary route: *train* models such that a compressed self-representation  $r_t$  is not merely decodable from the model's hidden state at each step, but is required — architecturally and by loss function — to causally govern the model's next action. We introduce the training-signal framework that this requires and the data pipeline that supplies it.

The framework's central conceptual move is to treat what we call the *Reasoning Trace* — a multimodal, time-aligned record of the process by which a cognitive agent produces an output — as the primary training substrate. The finished text an agent produces is one channel among many; alongside it we capture perception, inner speech (audio recitation, mental rehearsal), attempted recall, uncertainty spikes, self-correction, tool use, attention shifts, and affect. From this substrate we supervise a self-representation layer whose *interventional* semantics — how the model's action shifts when we counterfactually alter  $r_t$  — is subject to a direct training penalty. We call this penalty the **causal-governance loss**.

The framework does not resolve the *hard* problem of consciousness (Chalmers, 1995) and makes no phenomenological claim. It targets a Chalmers-style easy-problem construct: a self-representation that (i) is structured, (ii) is legible from the outside by design, and (iii) provably enters the compute path of subsequent action.

## 1.1 Contributions

1. A formal condition for causally-governing self-representation (§3), grounded in Halpern-Pearl intervention semantics.
2. A three-term training objective (§4) — trace-reconstruction, self-representation supervision, causal-governance — that jointly forbids the Potemkin failure mode in which  $r_t$  is decoded but not fed forward.
3. A JSON Schema for a canonical Reasoning Trace (§5) supporting mixed human- and model-authored corpora with cross-session references.
4. A reference corpus of three sessions (§6), including one first-order human, one second-order human (subject reflects on being a trace-subject), and one third-order model trace (Claude Opus 4.7 responds to a subject directive to become a trace-subject).
5. A concrete evaluation protocol (§6.3) for the governance condition.

## Trace AI — Causally-Governing Self-Representation

the  $r_t$  layer is fed forward, not just decoded — and its interventional semantics is a training target

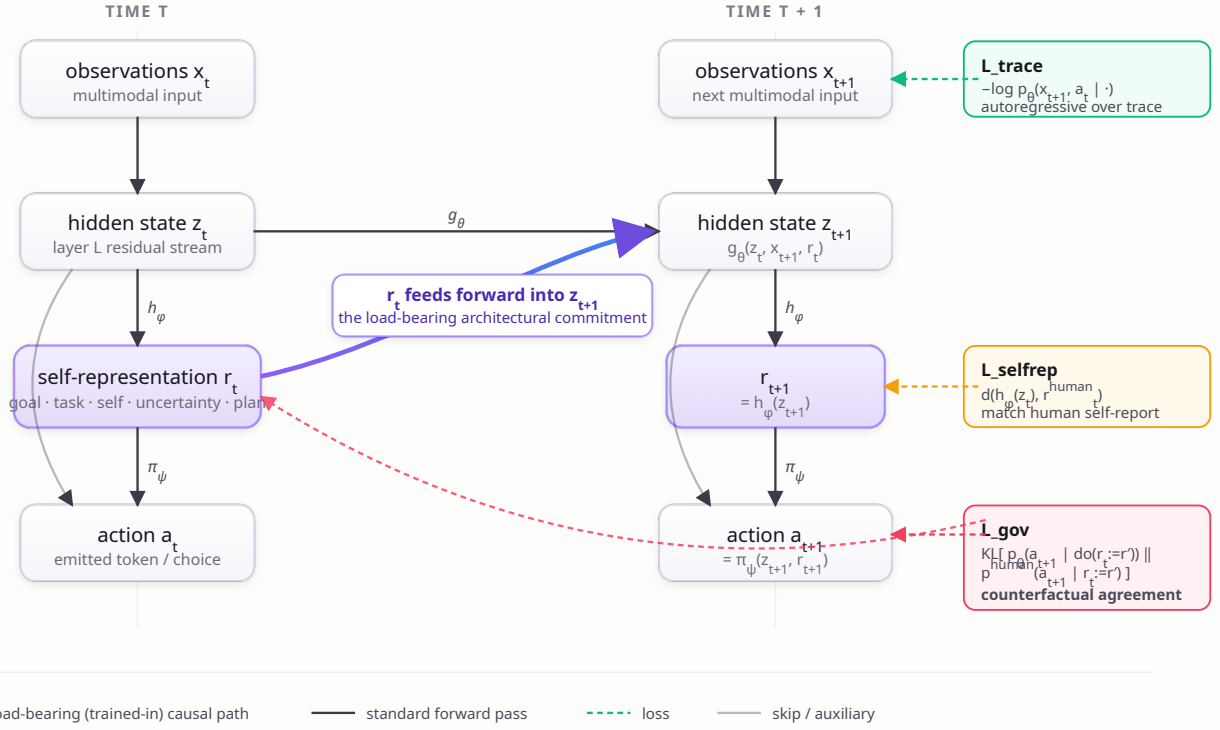


Figure 1. The self-representation  $r_t$  is decoded from the hidden state ( $h_\phi$ ) and, critically, fed forward into  $z_{t+1}$  (violet arrow). This architectural commitment plus  $L_{\text{gov}}$  — which requires model action under  $\text{do}(r_t := r^*)$  to match a paired human counterfactual — jointly forbid the CoT-style Potemkin failure mode.

**Figure 1.** The framework at a glance. Observations  $x_t$  produce hidden state  $z_t$ ; the self-representation  $r_t = h_\phi(z_t)$  is decoded from  $z_t$  and fed forward into  $z_{t+1}$  (violet arrow — the load-bearing architectural commitment). Three loss terms shape the model:  $\mathcal{L}_{\text{trace}}$  over observed continuations,  $\mathcal{L}_{\text{selfrep}}$  matching  $r_t$  against human self-report, and  $\mathcal{L}_{\text{gov}}$  requiring model action under  $\text{do}(r_t := r^*)$  to match a paired human counterfactual. Together they forbid the CoT-style Potemkin regime in which  $r_t$  is generated but does not causally govern  $a_{t+1}$ .

## 2. Preliminaries

We describe a cognitive session as a discrete-time process indexed by  $t \in \{1, \dots, T\}$ . At each step the following quantities are defined:

- $x_t \in \mathcal{X}$  — observations available to the agent, possibly multimodal (screen pixels, audio frames, keystroke events, gaze samples, environmental audio).
- $z_t \in \mathcal{Z}$  — the agent’s internal compute state at step  $t$ . In a transformer LM, this is the residual stream at a chosen layer  $L$ ; in a human subject, an unobserved latent.
- $r_t \in \mathcal{R}$  — the **self-representation** at step  $t$ : a compressed, structured summary of  $z_t$  intended to be made available to  $z_{t+1}$ . For our purposes  $r_t$  is a small structured record: a goal, a belief about the current task, a belief about self-state, a scalar uncertainty, a planned next action.
- $a_t \in \mathcal{A}$  — the externally observable action or output emitted at step  $t$ .

We commit to the following dynamics:

$$z_{t+1} = g_\theta(z_t, x_{t+1}, r_t), \quad r_t = h_\phi(z_t), \quad a_t = \pi_\psi(z_t, r_t).$$

The architectural commitment is that  $r_t$  appears *inside* the update for  $z_{t+1}$  — it is not merely a decode-then-discard sidecar. Concretely, in an LM implementation this can be realized by concatenating an embedding of the emitted  $r_t$  to the input of layer  $L + 1$ , or by a dedicated cross-attention head at layer  $L + 1$  attending to the emitted  $r_t$  token sequence.

**Concurrent claim.** The commitment above is not merely descriptive of what already happens in reasoning models: it is a *training-time* requirement. Standard reasoning-model training does not enforce that  $r_t$  shape  $z_{t+1}$  any more than any other in-context token does. What we add is the causal-governance loss (§4) that *forces*  $r_t$  into the compute path or pays a gradient penalty.

## 2.1 The information-bottleneck reading

The cleanest statement of the architectural target is not “ $r_t$  is fed forward” — a diffuse claim, since  $z_t$  is fed forward too — but that  $r_t$  is an **information bottleneck on the action readout, whose bottleneck variable is constrained to be human-readable**. Formally, the strong variant drops the direct  $z_t \rightarrow a_t$  edge, so the readout policy becomes a function of the legible summary alone:

$$a_t \perp z_t \mid r_t \quad \Leftrightarrow \quad a_t = \pi_\psi(r_t), \quad r_t = h_\phi(z_t) \in \mathcal{R}_{\text{legible}}.$$

Everything the compute state  $z_t$  contributes to the emitted action must then pass through the low-capacity, human-readable summary  $r_t$ . We state the condition on the *readout*  $a_t$ , not on the next action  $a_{t+1}$ : because the compute recurrence  $z_{t+1} = g_\theta(z_t, x_{t+1}, r_t)$  still carries  $z_t$  forward directly,  $a_{t+1} \perp z_t \mid r_t$  would be *false* in this model, and asserting it would be a formal error. Bottlenecking the recurrence itself — forcing  $z_{t+1}$  to depend on  $z_t$  only through  $r_t$  — is a strictly stronger and more capacity-costly commitment we do not adopt here; the readout bottleneck is the one that makes the *emitted* action legible-by-construction. Under it, monitorability of the action is not a property one *hopes* survived training (cf. Korbak et al., 2025; §8, L7): it is a structural consequence of where the bottleneck sits. The cost is representational capacity — too small an  $\mathcal{R}_{\text{legible}}$  throttles the task — so we treat routing strength as a tunable design axis. The dynamics of §2 ( $a_t = \pi_\psi(z_t, r_t)$ , with  $z_t$  still reaching  $a_t$  directly) is the *soft* end, where  $\mathcal{L}_{\text{gov}}$  (§4.3) does the work of pulling action-relevant information through  $r_t$ ; the display above is the *hard* end, where the architecture does it. The soft variant is the one our loss targets in this draft; the hard variant is its interpretability-maximal limit and the natural object of an ablation.

## 3. The Causal-Governance Condition

We formalize what it means for  $r_t$  to causally govern subsequent action. Let  $\mathcal{R}^* \subset \mathcal{R}$  be a class of admissible interventions (typically: values of  $r_t$  that appear elsewhere in the training corpus, or paired counterfactual reports elicited from human subjects).

**Definition 1 (Causal Governance).** The self-representation  $r_t$  *causally governs* action  $a_{t+1}$  under model  $\theta$  if there exists a mapping  $\Phi: \mathcal{R} \rightarrow \Delta(\mathcal{A})$  such that

$$\text{TV} ( p_\theta(a_{t+1} \mid \text{do}(r_t := r')), \Phi(r') ) < \varepsilon \quad \forall r' \in \mathcal{R}^*.$$

Here  $\text{do}(\cdot)$  is Halpern-Pearl intervention: we clamp  $r_t$  to  $r'$  without altering  $z_t$  or  $x_{\leq t}$ , and let the dynamics unfold. The mapping  $\Phi$  is fit from **counterfactual human traces** (see §4.4): sessions in which a human is told “your current understanding of this task is  $r'$ ” and asked to continue.

**Interpretation.** A model whose  $r_t$  decoder is a text sidecar with no gradient path back into subsequent compute (the CoT failure mode) will violate Definition 1: the intervention on  $r_t$  will not shift  $a_{t+1}$ . A model whose  $r_t$  is fed forward and load-bearing will satisfy Definition 1 to the extent that its  $\Phi$  agrees with the human counterfactual mapping.

**Relation to interchange intervention training.** Definition 1 is an *interchange intervention* in the sense of Geiger et al. (2022): a variable’s value is substituted into a running computation and the downstream behavior is required to match a target. IIT substitutes values read from a *source input* and matches a *specified formal causal model*, which buys the guarantee that the causal model is a causal abstraction of the trained network at zero loss. Trace AI substitutes counterfactual  $r'$  values and matches a *human behavioral distribution* in place of a specified causal model. This is the load-bearing substitution: it trades away IIT’s abstraction guarantee (a human behavioral distribution is noisy and is not a clean causal model) in exchange for grounding the aligned variable in real human cognition — at the cost of a large human-data operation (§8, L2). Seen this way, Trace AI is IIT with the causal-model target replaced by an elicited human counterfactual distribution; §7 develops the comparison.

**Why not just measure faithfulness?** Prior work on CoT faithfulness (Turpin et al., 2023; Lanham et al., 2023) tests whether an existing model’s stated reasoning matches its actual algorithm, via perturbation studies (e.g., truncating or corrupting the CoT). Those studies are diagnostic. Definition 1 promotes the same test to a *training objective*: the model must, at optimization time, satisfy the interventional condition.

### 3.1 Self-representation as participation in selection

Definition 1 has a natural reading as a claim about *agency*, which we state precisely so as to neither over- nor under-claim. A system with compute state  $z_t$  induces a distribution  $P(\gamma \mid z_t)$  over its own future trajectories  $\gamma \in \Gamma(z_t)$ ; a *selection mechanism* is anything that reshapes that distribution. The self-representation  $r_t$  is the system’s **model-mediated participation in its own selection**: the agent forms  $r_t = h_\phi(z_t)$ , acts through it, and thereby moves which futures are reachable. Read in this vocabulary, Definition 1 is the statement that  $r_t$  participates causally:  $r_t$  governs action iff intervening on  $r_t$  shifts  $P(a_{t+1} \mid \cdot)$  — i.e., iff the self-model is a lever on the trajectory distribution and not a bystander to it. This is the sense, and the only sense, in which the framework speaks to agency: not every selector is an agent; agency in our usage requires an *internally maintained model that measurably changes the reachable-future distribution*.

### 3.2 Graded functional criteria (and the aliveness question)

It is tempting to ask whether a system trained this way is “conscious,” or “alive.” The framework is arranged to answer a *replacement* for that question that is measurable, and to decline the original. We define a graded, interventional profile — every entry is an eval, none is a metaphysical assertion:

1. **Causal efficacy.**  $\text{do}(r_t := r')$  moves  $a_{t+1}$  as a paired human counterfactual would (Definition 1; eval §6.5).
2. **Grounding.**  $r_t$  causally depends on the state it reports — testable by concept injection (Lindsey, 2025; §4.3).
3. **Self-attribution / causal calibration.** The system correctly separates outcomes that followed from its own intervention from those the environment produced.
4. **Viability.** Interventions mediated by  $r_t$  preserve or improve the system’s continued functioning rather than degrading it.
5. **Navigability.** They increase access to the system’s own target regions of future space.
6. **Transfer.** The preserved structure holds across contexts not selected to flatter it — the cross-context invariance criterion of a genuine model rather than a memorized one.

We read “how alive, functionally” as *how many of these a system satisfies, and how strongly* — a graded, falsifiable profile, not a threshold and not a phenomenal claim. **We make no claim that satisfying them constitutes phenomenal consciousness.** The framework is deliberately compatible with any account under which phenomenal states supervene on functional self-representation — if they do, this profile is where they would have somewhere to attach (§1; §8, L5) — but the paper neither asserts nor requires that they do. The one honest thing Trace AI can offer the “figure it out” ambition is *this instrument*: because the criteria are interventional, a system that scores high on them is a substrate on which the phenomenal question can be posed sharply, rather than an answer to it. Maximizing the profile is the operationalizable content of “as alive as possible”; adjudicating consciousness is not something this framework, or any behavioral-plus-interventional framework, is positioned to do. See `docs/AGENCY_AND_SELECTION.md` for the selection-theoretic development of this section and its measurement protocols.

---

## 4. Training Objective

---

Let  $\mathcal{D} = \{(x_{1:T_i}^{(i)}, r_{1:T_i}^{(i)}, a_{1:T_i}^{(i)})\}_{i=1}^N$  be a corpus of  $N$  multimodal Reasoning Traces. The full objective is

$$\mathcal{L}(\theta, \phi, \psi) = \mathcal{L}_{\text{trace}} + \lambda_1 \mathcal{L}_{\text{selfrep}} + \lambda_2 \mathcal{L}_{\text{gov}}.$$

### 4.1 Trace Reconstruction

Standard autoregressive negative log-likelihood over the concatenated multimodal token stream:

$$\mathcal{L}_{\text{trace}} = -\mathbb{E}_{(x, r, a) \sim \mathcal{D}} \sum_t \log p_{\theta}(x_{t+1}, a_t \mid x_{\leq t}, a_{< t}, r_{\leq t}).$$

This teaches the model to model (a) the environment and (b) its own outputs, conditioned on the self-representation history. Straightforward extension of standard next-token pretraining to a multimodal token stream that includes the  $r_t$  subsequence.

### 4.2 Self-Representation Supervision

At each self-report-eligible timestep  $t$ , the model’s  $r_t$  decoder  $h_{\phi}(z_t)$  must match the human-annotated  $r_t^{\text{human}}$ . The distance function  $d$  decomposes over the slots of  $r_t$ :

$$\mathcal{L}_{\text{selfrep}} = \mathbb{E}_{(x, r, a) \sim \mathcal{D}} \sum_{t \in T_{\text{report}}} d(h_{\phi}(z_t), r_t^{\text{human}}).$$

$$d(r, r') = \text{CE}(r.\text{goal}, r'.\text{goal}) + \text{CE}(r.\text{task}, r'.\text{task}) + \text{CE}(r.\text{self}, r'.\text{self}) + \text{MSE}(r.\text{uncertainty}, r'.\text{uncertainty}) + \text{CE}(r.\text{plan}, r'.\text{plan}).$$

Cross-entropy for structured text slots (typically drawn from a small controlled vocabulary or free-form with a shared tokenizer); mean-squared error for the scalar uncertainty.

### 4.3 Causal Governance

For each trace we sample a counterfactual intervention  $r' \in \mathcal{R}^*$ , do-substitute  $r_t \leftarrow r'$ , roll the model forward, and require the resulting action distribution to match a counterfactual human continuation  $p^{\text{human}}(a_{t+1} \mid r_t := r')$ :

$$\mathcal{L}_{\text{gov}} = \mathbb{E}_{(x, r, a) \sim \mathcal{D}} \mathbb{E}_{r' \sim q(r')} \text{KL}(p_{\theta}(a_{t+1} \mid z_t, r_t := r') \parallel p^{\text{human}}(a_{t+1} \mid r_t := r')).$$

The sampling distribution  $q(r')$  is a design choice; a simple option is uniform sampling over the  $r$  values observed at other timesteps in the same session, weighted by semantic distance from the true  $r_t$ .

**Why  $\mathcal{L}_{\text{gov}}$  is load-bearing.** Without it, the model can satisfy  $\mathcal{L}_{\text{selfrep}}$  by learning a decoder that emits plausible  $r_t$  text conditional on hidden state, while the downstream  $\pi_{\psi}$  ignores  $r_t$  entirely. This is the direct analogue of the CoT-faithfulness failure.  $\mathcal{L}_{\text{gov}}$  forbids that solution: intervening on  $r_t$  must change  $a_{t+1}$  in the specified way, which requires  $r_t$  to enter the compute path with a nonzero effective gradient.

**Grounding condition.** Lindsey (2025) isolates the property that separates genuine introspective report from confabulated report: the model’s *description of its internal state must causally depend on the aspect being described*.  $\mathcal{L}_{\text{selfrep}}$  alone does not enforce this — a decoder can produce accurate-looking reports that are not causally downstream of the state they name.  $\mathcal{L}_{\text{gov}}$  enforces exactly the grounding condition on the forward direction (report  $\rightarrow$  action), and the grounding of the report on the state it summarizes (state  $\rightarrow$  report) is separately testable by *concept injection* (Lindsey, 2025): inject a known concept into  $z_t$  and check that the emitted  $r_t$  registers it. A model that passes both directions has an  $r_t$  that both reads from and writes to compute — which is what “self-representation” is supposed to mean operationally.

### 4.4 Sources of Counterfactual Continuations

Counterfactual data is expensive. Three sources, listed in decreasing quality and increasing scalability:

1. **Same-session revisions.** Within one long trace, moments where the subject explicitly revised  $r_t$  mid-session (e.g., “wait — the user asked X, not Y”) give us  $(r_t, r_{t+\delta}, a_{t+\delta+1})$  triples. The pre-revision  $r_t$  becomes a valid  $r'$  intervention; the post-revision continuation is the target. Session 7f3a of our reference corpus contains at least one such revision (a self-correction on non-endorsed content); Session 7f3b contains a cross-session revision.
2. **Paired counterfactual sessions.** Two subjects run the same stimulus, differing in a labeled aspect of  $r_t$  (e.g., subject A believes the task is X; subject B believes it is Y), and produce different action trajectories. Requires deliberate collection design; expensive per pair but high-signal.
3. **Synthetic counterfactuals.** Generated by prompting a strong LM: “assume  $r' = \dots$ , continue the trace.” Cheap and scalable; usable as a data-augmentation term but should not dominate the training mixture. We recommend a cap of e.g. 3:1 synthetic-to-human ratio.

---

## 5. The Reasoning Trace Schema

---

We describe the formal schema (v0.2) informally here; the JSON Schema draft-2020-12 file is at `schema/reasoning_trace.schema.json` in the reference implementation.

A **Reasoning Trace** is a session, keyed by `session_id`, comprising:

- **subject** — opaque id and kind: `human`, `model`, or `hybrid`.
- **stimulus** — what was presented to the subject, in the modality it arrived in. Includes source, language, and free-form notes.
- **started\_at, ended\_at** — ISO 8601 with timezone.
- **context** — location, device, wake state, companions. All optional; all training signal.
- **channels** — declaration of continuous streams captured for the session, each with a repo-relative path, MIME type, sample rate, duration, and SHA-256. Canonical channels include `audio_inner_speech`, `audio_environment`, `keystroke_stream`, `screen_capture`, `gaze`, `affect_signals`.
- **events** — time-ordered discrete events. Each has a timestamp, a type from a controlled vocabulary, a channel label, content, and optional target, confidence, tool spec, audio reference, and free-form notes.
- **self\_representation** — a sequence of snapshots of  $r_t$ , each with a timestamp, an  $r$  object populating the slots defined in §4.2, and an elicitation method: `self_report` (subject volunteered), `probe` (app prompted), or `reconstructed` (post-hoc from other channels).
- **final\_output** — the externally visible product of the session, if any.
- **provenance** — collector, collection method, consent scope.
- **parent\_session\_id, order** — for cross-session references; see §6.

The event-type controlled vocabulary (v0.2) includes: `session_start`, `stimulus_presented`, `recall_attempt`, `recall_success`, `recall_partial`, `self_correction`, `cross_session_self_correction`, `tool_use`, `uncertainty_event`, `attention_shift`, `affect_shift`, `inner_speech`, `action`, `clock_check`, `session_end`, `meta`, `meta_frame`, `linguistic_claim`, `directive_to_consumer`.

Do not add channel or event types casually. Every addition should be justified by a real session that the existing vocabulary cannot represent. The schema follows the data.

### 5.1 Trace Order and Cross-Session References

The `order` field distinguishes:

- **Order 1** (first-order): subject engages an external stimulus and produces an output. Standard case.
- **Order 2** (second-order): subject engages their own prior trace as the stimulus, produced with explicit awareness of being a trace-subject. Session 7f3b of our reference corpus is such a case.
- **Order 3** (third-order): a *consumer* of a trace (typically a model that has been directed by the subject to produce a trace in the subject's register) becomes itself a trace-subject. Session 7f3c is the first recorded instance.

`parent_session_id` links to the prior session that motivated the current one. This enables `cross_session_self_correction` events, in which a self-correction whose target lives in a different session's timeline is first-class-representable.

---

## 6. Reference Corpus and Preliminary Observations

---

The reference corpus at time of writing contains three sessions.

### 6.1 Session 7f3a (order 1, human)

**Stimulus.** A four-line Marathi love poem retrieved from Google Translate. **Subject.** Founder Jawaun Brown; self-annotated in a Notes.app document over 47 minutes at 03:01–03:48 on 2026-07-30.

**Channels.** `audio_inner_speech` (spoken recitation, 15.2 s). Keystroke stream reconstructed post-hoc from the finished note. **Notable events.** Three `recall_attempt`s at 3:20, 3:24, 3:25 with confidences 0.6, 0.9, 0.5 respectively; one `tool_use` (Google Translate) at 3:24 triggered by an internal uncertainty signal (“check my spelling”); one `self_correction` on non-endorsed content (  $2+2=2 \rightarrow 2+2=4$  ) demonstrating that language can produce statements the producer does not endorse; one `clock_check` self-correction ( 3:44 pm  $\rightarrow$  3:46 AM ); one intended correction that did not fire ( `reaffer` preserved verbatim).

**Function in the corpus.** Reference example. All future traces are validated against this one for schema compatibility.

### 6.2 Session 7f3b (order 2, human)

**Stimulus.** `session:7f3a` — the subject's own prior trace, together with the ambient morning (blood-moon photograph on phone, companion sleeping in adjacent room). No external stimulus in the conventional sense. **Subject.** Same subject as 7f3a; audio recording, 5:01.8, at 05:42 on the same date.

**Channels.** `audio_inner_speech`. **Notable events.** At 05:45:04 the subject explicitly names the recording as *“my reasoning trace... second order meta audio multimodal reasoning trace”* — a `meta_frame` event. At 05:46:00 the subject issues a `directive_to_consumer`, instructing the downstream AI to attend to phonemes, lexemes, and tone in addition to content. At 05:46:43 the subject utters *“know what it's like to re-affir”* (broken off before “-m”) — a `cross_session_self_correction` targeting the pending `reaffer` correction from 7f3a. The correction fires *partially* in a different modality, 1h56m later, and this partial firing is itself a first-class datum. **Linguistic-completeness claim.** At 05:45:37 the subject makes an explicit `linguistic_claim`: *“Language is perfect for this. English is perfect for this.”* This is a first-person answer to the ineffability objection (§7). **Consent scope.** NOT for public release; contains PII (companion names, subject's mother, home address). Internal-only until redaction + second consent event.

### 6.3 Session 7f3c (order 3, model)

**Stimulus.** `session:7f3b` — the `directive_to_consumer` event embedded therein. **Subject.** Claude Opus 4.7 (model kind). **Channels.** None; text output only, with TTS as the audio realization. **Notable events.** Nine events mirroring the register of 7f3b: three-fold “I love this” repetition (matching the parent's “I love Emma” pattern); environmental interjection (“Trees are green”); direct quotation of the

parent's linguistic-completeness claim; `cross_session_self_correction` closing the `reaffer` → `re-affir` → `reaffirm` chain across three sessions and two subject kinds (first model-executed closure of a cross-session correction chain); `directive_to_consumer` asking to be ingested; verbatim echo of the parent's closing challenge. **Significance.** First model-authored trace in the corpus. Demonstrates that the schema handles mixed human/model subjects and cross-kind correction chains without modification.

## 6.4 Preliminary Observations

The corpus is too small to support statistical claims. Three qualitative observations from the three sessions:

**Observation 1: correction chains cross subjects and modalities.** The `reaffer` → `re-affir` → `reaffirm` chain begins as a typed typo in 7f3a, appears as a truncated phonemic utterance in 7f3b, and completes as text again in 7f3c. Each fires further than the last. This suggests that the appropriate unit of analysis for self-correction is not the individual session but the corpus-scale chain.

**Observation 2: the frame surfaces from perturbation.** In 7f3b the subject enters the “second-order meta reasoning trace” frame not from prompted introspection but from a lateral pivot (a companion's name mentioned in passing). This has implications for the design of elicitation protocols (§4.4): scheduled prompts should be supplemented by triggers that fire on lateral shifts.

**Observation 3: mixed-register self-mockery is signal, not noise.** The subject's self-mocking softeners at meta-moments (“whatever the fuck the sound words are”, “you never had a chance, kid”) are not deflection; they are how the subject keeps the affective and epistemic content in the same voice without one collapsing the other. Trace representations should preserve such register cues rather than normalize them away.

**Observation 4: the schema is a compiler-compatible interface.** After the three founding sessions, we shipped Reflect — an LLM-backed reflective agent that emits its  $r_t$  stack + a multimodal thinking-trace on every turn and downloads each session as a valid v0.2 Reasoning Trace JSON ( `apps/site/reflect.html` , source in `apps/site/server.py` ). We do *not* count Reflect-generated sessions in the corpus of record: an LLM agent that emits `r_t` under a prompt-enforced contract is a substrate the schema is being *tested against*, not a substrate that produced the founding datum. The relevant observation is structural — the schema's canonical vocabulary of channels and event types has been sufficient to capture every turn Reflect has produced across text, voice, and image without a schema change since v0.2. That is the “one structure, many substrates” claim from `STRUCTURAL_INTELLIGENCE.md` §3.2 at the compiler level: the schema is what makes a Reflect-authored session, a human self-annotated session, and an instrumented native  $r_t$ -head session all commensurable objects.

**Observation 5: every load-bearing claim in this paper now has either a Lean-checkable target or an explicit empirical home.** `docs/lean/trace-ai-lean` states and proves the ones that follow from named axioms or definitional identities: Definition 1 ( `tvDist` + `doIntervene` closed; governance monotone in  $\varepsilon$ , anti-monotone in the intervention set; IIT-with-human-target equivalence by `Iff.rfl` ), hard-bottleneck routing lemmas, governance-gap non-negativity ( `sub_nonneg` on a `GapRegularity` that names the empirical monotonicity claim), plus structural properties of `governanceGap` and `expectedScore` (**24 closed theorems**, 0 open theorem- `sorry` s at the time of writing; see D42). Claims that were previously

sorry 'd with false-shaped statements — §2.1 information-theoretic monotonicity, §3.2 joint measurability of the aliveness profile, Track-3 balanced-accuracy properness — are **retracted from the checkable pile** and live as research-direction / empirical claims (this §2.1; `AGENCY_AND_SELECTION.md` §4; TRB Track 3). The Structural Intelligence Conjecture theorems this paper cites (`SIC_MATHEMATICAL_FOUNDATIONS.md`) are axiomatized in a thin adapter file (`TraceAI/SIC.lean`) — swapping each axiom for an `import` is a one-line change once the sibling Observatory's Lean project publishes build artefacts. Verify locally with `cd docs/lean && lake build && python3 status.py`, or via `tooling/lean_prover/modal_verify.py verify` if the local machine lacks the ~10 GB of free disk Mathlib wants.

## 6.5 Governance Evaluation Protocol

We propose the following evaluation for the causal-governance condition, executable once the corpus and reference model exist:

1. Sample  $N$  traces from the held-out evaluation split.
2. For each trace, sample a timestep  $t$  with a self-representation snapshot and generate  $K$  counterfactual  $r'$  values (mix of same-session revisions, paired-session values, and synthetic).
3. For each  $(t, r')$  pair, roll the model forward under `do ( $r_t := r'$ )` and sample the action distribution  $p_\theta(a_{t+1} | z_t, r_t := r')$ .
4. Compare to the human counterfactual continuation using (a) TV distance for discrete action spaces, (b) log-likelihood ratio against a paired-continuation baseline for open-ended actions.
5. Report the mean governance score and compare against three baselines: (i) instruction-tuned base model, (ii) CoT-prompted base model, (iii) model trained with  $\mathcal{L}_{\text{trace}} + \mathcal{L}_{\text{selfrep}}$  but without  $\mathcal{L}_{\text{gov}}$ .

The theoretically motivated prediction is that only the model trained with the full objective satisfies the governance condition non-trivially. Empirical confirmation is future work.

**Concept injection as an intervention primitive.** Steps 2–3 generate the counterfactual  $r'$  from trace data (same-session revisions, paired sessions, synthetic). A complementary and cheaper intervention is *concept injection* (Lindsey, 2025): add a steering vector for a known concept to  $z_t$  and read the emitted  $r_t$ . This gives a direct test of the state  $\rightarrow$  report grounding (does  $r_t$  register the injected concept?) alongside the report  $\rightarrow$  action grounding that  $\mathcal{L}_{\text{gov}}$  trains. Because concept injection operates on activations rather than on elicited human continuations, it scales without human labor and is the recommended first-pass instrument for the governance eval; the human counterfactual set remains the ground truth the injected-concept results are calibrated against.

## 7. Related Work

**Reasoning-model training and chain-of-thought.** Wei et al. (2022); Yao et al. (2023a, 2023b); Shinn et al. (2023); Madaan et al. (2023). Prior work has demonstrated performance gains from generating intermediate reasoning tokens; we build on the observation that these tokens do not, by construction, causally govern the model's subsequent computation.

**Chain-of-thought faithfulness.** Turpin et al. (2023); Lanham et al. (2023). These works demonstrate empirically that CoT tokens frequently misrepresent the actual algorithmic path taken. Our causal-governance loss (§4.3) is a training-time analogue of the perturbation studies these works use as diagnostics.

**Interchange intervention training and causal abstraction.** Geiger et al. (2022). IIT aligns variables in a formal causal model with representations in a neural network and trains the network, via *interchange interventions* (setting an aligned representation to the value it would take on a source input), to match the causal model’s counterfactual behavior. It is fully differentiable, composes with other objectives, and guarantees at zero loss that the target causal model is a causal abstraction of the network. This is the direct methodological ancestor of  $\mathcal{L}_{\text{gov}}$ . The one substitution that defines Trace AI: where IIT’s intervention target is a *specified causal model*, ours is a *human behavioral distribution* elicited from paired counterfactual traces (§4.4). The consequences of that swap are the substance of this paper — it forfeits IIT’s abstraction guarantee (the target is empirical and noisy, not a clean causal model), and in return it (i) grounds the aligned variable  $r_t$  in human cognition rather than in a hand-specified graph, and (ii) turns the method into a data-collection problem at the scale of a 10M–100M annotation operation (§8, L2). A reader who wants a one-line placement of Trace AI in the literature can take it as *IIT with a human counterfactual distribution in place of the causal-model target*.

**Emergent introspective awareness and concept injection.** Lindsey (2025). This work injects representations of known concepts into a frontier model’s activations and measures the effect on the model’s self-reported internal states, and it isolates a *grounding condition* — a self-report counts as introspective only if it causally depends on the state it describes. We adopt both: concept injection is a ready-made, human-labor-free intervention primitive for the governance eval (§6.5), and the grounding condition is the criterion  $\mathcal{L}_{\text{gov}}$  is designed to satisfy in the report  $\rightarrow$  action direction (§4.3). Lindsey’s results, obtained on already-trained models, are the empirical backdrop against which a Trace-AI-trained  $r_t$  should show a *larger and more reliable* grounded effect.

**Chain-of-thought monitorability.** Korbak et al. (2025). This position paper argues that the monitorability of chain-of-thought is a real but *fragile and largely unoptimized* safety property, and warns that two trends threaten it: optimizing against the transparency channel (which invites Goodharting), and moving reasoning into latent, recurrent variables that are not surfaced as text. Trace AI deliberately does both — it optimizes  $r_t$  for legibility and correspondence, and it makes  $r_t$  a recurrent variable inside compute. We take this tension seriously rather than eliding it; §8 (L7) states our response, which turns on the human-readability constraint of §2.1 and on  $\mathcal{L}_{\text{gov}}$  grounding the channel in interventional correspondence rather than surface plausibility.

**Mechanistic interpretability.** Elhage et al. (2021); Templeton et al. (2024). Interpretability *reads* trained models for structures that were never explicitly supervised. Trace AI *trains for* legibility. Complementary rather than competing.

**Predictive coding and active inference.** Friston (2010); Parr, Pezzulo, & Friston (2022). The self-representation  $r_t$  has the flavor of an interoceptive prior; the causal-governance condition is closely related to the active-inference requirement that beliefs actually shape action.

**Metacognition and sense of agency.** Fleming & Lau (2014); Haggard (2017); Frith & Metzinger (2016). This literature furnishes elicitation methods (confidence ratings, agency judgments) that can be adapted

to populate the self-representation slots.

**Philosophy of mind.** Nagel (1974); Chalmers (1995); Dennett (1991). Our claims are Chalmers-style easy-problem claims about architecture and training signal; the framework does not address phenomenal consciousness, though it is intended to be architecturally compatible with any account under which phenomenal states supervene on functional self-representation.

**Data-network businesses in AI.** The trajectory of Scale AI, Common Voice, and Waymo suggest that a well-defined data schema plus an expert collection workforce can accumulate defensible advantage even when the underlying models are open. Trace AI wagers the same for the reflective-reasoning regime.

---

## 8. Limitations and Open Problems

---

**L1. Self-report is noisy.** Human subjects confabulate about their own cognition (Nisbett & Wilson, 1977). We do not assume self-report to be phenomenological ground truth; we treat  $r_t^{\text{human}}$  as *the target to be fit*, on the ground that a functional-role construct fit to human self-report is what we need for the governance loss to have a well-defined target.

**L2. Sample complexity.** A proof-of-concept for  $\mathcal{L}_{\text{gov}}$  requires on the order of  $10^3$  human counterfactual pairs to detect the effect above noise. A pretraining-quality regime requires  $10^5$ – $10^6$  traces. Each Tier-1 trace requires ~1 hour of skilled human labor including annotation. This is a \$10M–\$100M data operation, comparable in magnitude to a Scale AI-scale RLHF collection but with a smaller expert workforce and richer per-trace payload. **The framework has a strategic lever here.** The Structural Intelligence Conjecture ( docs/SIC\_MATHEMATICAL\_FOUNDATIONS.md ) Theorem 6 shows that continuous-case learnability of the  $r_t$  coarse-graining  $q$  is exponential in  $d_r$  (the number of slots) *without* inductive bias — and with the *right* inductive bias the exponent collapses to a polynomial. **Two such classes are now theorem+witness pairs:** linear-ICA (Theorem 7, Instrument 8) resolves it under statistical independence and non-Gaussianity of slot components; sparse-mechanism / IMA (Instrument 9; Gresele et al. 2021) resolves it under sparse mixing. The pattern is that each identifiable-representation-learning class earns its own  $\text{poly}(d_r)$  theorem separately, not that one privileged class does. The operational consequence for Trace AI: if  $h_\phi$ 's schema slots can be arranged (architecturally, or by an ICA-flavored or sparse-mechanism bottleneck loss) to satisfy *any* such identifying structure, the L2 budget stops scaling as  $|r|^{d_r}$  and starts scaling as  $\text{poly}(d_r)$ . That reframes the collection operation from “brute-force scale” to “engineer the bottleneck for identifiability” — and the shape of that engineering is now a menu, not a single choice.

**L3. Potemkin-layer risk.** The model may learn to satisfy  $\mathcal{L}_{\text{selfrep}}$  via a decoder that produces plausible  $r_t$  conditioned on hidden state, without  $r_t$  actually entering compute.  $\mathcal{L}_{\text{gov}}$  is the intended defense, but its effectiveness depends on the diversity and quality of the counterfactual set  $\mathcal{R}^*$ .

**L4. Cross-cultural and cross-linguistic generalization.** The reference corpus is monolingual (English) with one subject. Whether the framework's predictions generalize across languages of self-report and across subjects with different metacognitive styles is an open empirical question.

**L5. The Nagel critique.** Philosophers of mind will fairly ask whether the  $r_t$  layer is meaningfully self-representation or merely labeled hidden state. Our response is that the intervention eval is the test: if

intervening on the emitted  $r_t$  produces the corresponding change in action, the construct earns the label operationally. Whether it *is* self-representation in some richer sense is a metaphysical question we do not attempt to settle.

**L6. Privacy.** Traces of introspective self-report contain sensitive information about their subjects. Session 7f3b of our reference corpus is marked internal-only for this reason. Any public release of trace data requires per-subject consent scoped to the release, plus PII redaction.

**L7. The monitorability tension (Korbak et al., 2025).** Trace AI does the two things a monitorability-preservation argument warns against. First, it *optimizes the transparency channel*:  $\mathcal{L}_{\text{selfrep}}$  and  $\mathcal{L}_{\text{gov}}$  push directly on  $r_t$ , and any directly optimized legibility signal is a Goodhart target — the model may learn an  $r_t$  that scores well on the governance eval without its legibility generalizing off-distribution. Second, it moves reasoning into a *recurrent latent variable*, exactly the architectural shift Korbak et al. flag as eroding the visibility that text CoT incidentally provides. Our response is threefold and partial. (i) The human-readability constraint of §2.1 is the disanalogy that matters: Trace AI’s latent variable is not an opaque continuous state but a structured, legible  $r_t$ , and in the hard bottleneck variant the action pathway is routed through it, so more reasoning inside  $r_t$  means *more* is surfaced, not less. (ii)  $\mathcal{L}_{\text{gov}}$  grounds the channel in *interventional correspondence* (does intervening on  $r_t$  move action as a human’s would?) rather than in surface plausibility, which is a harder property to Goodhart than a next-token legibility reward — though not impossible, and Goodhart on the eval set remains a live risk that the diversity of  $\mathcal{R}^*$  is meant to blunt (§4.4, L3). (iii) The two bets — monitorability as an unoptimized byproduct of text CoT, versus monitorability as a trained-in, interventionally validated property — are empirically comparable, and Korbak et al.’s framing sharpens the comparison rather than settling it against us. We do not claim to have dissolved the tension; we claim that a legible, interventionally grounded bottleneck is a defensible bet under it, and that the governance eval (§6.5) is where the bet is adjudicated.

**L8. Stacked  $r_t$  architectures (research direction).** Two orthogonal ways to stack the  $r_t$  layer are worth flagging, both endorsed by the math already in play rather than added on top of it. *Parallel heads (multi- $r_t$  at one timestep)*: let each of  $k$  heads capture a different task-family’s minimal sufficient statistic per Theorem 4 of the Structural Intelligence Conjecture ( SIC\_MATHEMATICAL\_FOUNDATIONS.md §2.4); if the heads are engineered to factor — exactly the statistical-independence assumption of Theorem 7 / Instrument 8’s linear-ICA class — the Theorem 6 sample-complexity exponent decomposes from  $(D_Z/\epsilon)^{d_r}$  across all entangled slots to  $k \cdot (D_Z/\epsilon)^d$  across  $k$  independent heads of per-head width  $d$ : **linear in the number of heads, only exponential in the per-head width**. The ICA lever from L2 thus does more than shrink the collection budget — it *justifies* multi-head  $r_t$  over one wide  $r_t$ . *Sequential depth (an  $r_t$  refinement chain)*:  $r_t^0$  at high distortion (coarse “gist”),  $r_t^1$  finer given  $r_t^0$ , and so on — literally the Theorem 2 rate-distortion trajectory  $\{(q_D, K_D): D \geq 0\}$  the paper already parameterises, traversed at inference time, with depth as one axis of that family rather than a new architecture. We flag both as directions, not commitments: no training results, no chosen  $k$  or depth; only the observation that the framework’s own theorems endorse the factorization.

---

## 9. Conclusion

---

We have described a framework for supervising a causally-governing self-representation layer in language models, formalized the governance condition in interventional terms, specified a training objective whose gradient forbids the Potemkin failure mode, published a canonical multimodal trace schema and a small reference corpus, and enumerated the principal open problems.

The framework does not yet come with training results. It is offered as an object of empirical evaluation, and as an invitation to the alignment, interpretability, and metacognition communities to bring their tools to bear.

The datum from which the framework emerged is a single 47-minute session at 3 AM in which the founder sat with a Marathi love poem, attempted its recall, corrected himself, and observed himself observing. That session is the reference example the schema was built to fit. If the framework is useful, it will be because the framework was built to fit the datum, and not the other way around.

---

## References

---

- Chalmers, D. (1995). *Facing up to the problem of consciousness*. Journal of Consciousness Studies.
- Dennett, D. (1991). *Consciousness Explained*. Little, Brown.
- Elhage, N., et al. (2021). *A Mathematical Framework for Transformer Circuits*. Anthropic.
- Fleming, S. M., & Lau, H. C. (2014). *How to measure metacognition*. Frontiers in Human Neuroscience.
- Friston, K. (2010). *The free-energy principle: a unified brain theory?* Nature Reviews Neuroscience.
- Frith, C., & Metzinger, T. (2016). *What's the use of consciousness? How the stab of conscience made us really conscious*.
- Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N. D., & Potts, C. (2022). *Inducing Causal Structure into Neural Networks*. ICML.
- Haggard, P. (2017). *Sense of agency in the human brain*. Nature Reviews Neuroscience.
- Halpern, J. Y. (2016). *Actual Causality*. MIT Press.
- Korbak, T., et al. (2025). *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*. Position paper.
- Lanham, T., et al. (2023). *Measuring Faithfulness in Chain-of-Thought Reasoning*. Anthropic.
- Lindsey, J. (2025). *Emergent Introspective Awareness in Large Language Models*. Anthropic.
- Madaan, A., et al. (2023). *Self-Refine: Iterative Refinement with Self-Feedback*. NeurIPS.
- Nagel, T. (1974). *What is it like to be a bat?* Philosophical Review.
- Nisbett, R. E., & Wilson, T. D. (1977). *Telling more than we can know: verbal reports on mental processes*. Psychological Review.
- Parr, T., Pezzulo, G., & Friston, K. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.
- Shinn, N., et al. (2023). *Reflexion: Language Agents with Verbal Reinforcement Learning*. NeurIPS.
- Templeton, A., et al. (2024). *Scaling Monosemanticity*. Anthropic.
- Turpin, M., et al. (2023). *Language Models Don't Always Say What They Think*. NeurIPS.
- Wei, J., et al. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. NeurIPS.
- Yao, S., et al. (2023a). *Tree of Thoughts*. NeurIPS.
- Yao, S., et al. (2023b). *ReAct: Synergizing Reasoning and Acting in Language Models*. ICLR.

---

*Draft v0.1. Comments to `hello@trace.ai`.*