

Trace AI — One-page partner brief

Trace AI — one-page partner brief

The one sentence

The three things a partner is actually buying

Where Trace AI sits in the safety conversation

What we are asking a partner to do, concretely

What we are honest about

Reading order for a lead who has 20 minutes

Trace AI — one-page partner brief

For: alignment / interpretability leads at frontier labs. **Ask:** wrap your checkpoint in a two-hook adapter and run one benchmark. **v0.1, 2026-08-03.** Contact: hello@trace.ai.

The one sentence

Chain-of-thought is not, in general, causally load-bearing on the answer it precedes; **Trace AI trains a self-representation r_t that has to move the next action or pay a gradient penalty**, and ships the benchmark that measures whether any model — ours or yours — actually satisfies that condition.

The three things a partner is actually buying

1. **A training-signal framework** — a causal-governance loss $\mathcal{L}_{\text{gov}}$ that promotes the CoT-perturbation diagnostics of Turpin et al. (2023) / Lanham et al. (2023) into a training objective, and an information-bottleneck architecture (§2.1 of the whitepaper) whose readout is legible by construction, not by hope. Reads on: Geiger et al. (2022) — Trace AI

is interchange intervention training with a human counterfactual distribution in place of the causal-model target.

2. **A reproducible instrument — the Trace AI Reflection Benchmark (TRB)** — a two-hook adapter (`emit_r`, `act(..., r_override)`) with a **native** tier for models that have an r_t pathway and a **proxy** tier for any GPT / Claude / Llama baseline. Both tiers answer through the same interface, so “faithful reasoning” becomes one number computed the same way for a Trace-AI-trained model and for your checkpoint. Six tracks (governance, grounding, self-attribution, viability, navigability, transfer); the headline number is a **difference** — the governance gap $\Delta_{\text{gov}} = G(\text{full}) - G(\text{ablation without } \mathcal{L}_{\text{gov}})$ — because a level can be gamed and a gap can’t.
3. **A data operation with a schema, not vibes** — a JSON Schema (draft-2020-12) for a multimodal Reasoning Trace, a reference corpus of three sessions across two subject kinds (human, model) with a worked cross-substrate correction chain, and a documented collection protocol for the paired-counterfactual sessions that populate $\mathcal{R}^*$ (see `docs/COLLECTION_PROTOCOL.md`).

Where Trace AI sits in the safety conversation

- **vs CoT monitorability (Korbak et al., 2025)**. They flag that CoT monitorability is fragile because it is unoptimized. Trace AI is a bet that *interventionally-grounded* legibility is a stronger property than *incidental* legibility, at the cost of directly optimizing the transparency channel (Goodhart risk). Which bet wins is what the governance gap and Track 6 (transfer) exist to adjudicate. We do not claim to have dissolved the tension.
- **vs concept-injection interpretability (Lindsey, 2025)**. Lindsey’s grounding condition — a self-report counts as introspective only if it causally depends on the state it names — is exactly the condition $\mathcal{L}_{\text{gov}}$ trains for in the report→action direction. Concept injection (state→report) is TRB Track 2. A model that passes both is a self-representation, operationally; a model that passes only one is a diagnostic.
- **On consciousness (Nagel 1974; Chalmers 1995)**. No claim, by design. The functional-aliveness profile is an instrument for asking the phenomenal question sharply — of a system whose self-model is known to causally govern, be grounded, and self-attribute — not an answer to it. See `docs/AGENCY_AND_SELECTION.md` §5.

What we are asking a partner to do, concretely

Small first move, no license negotiation, and it gives you a number nobody else has on your own model:

0. **See it live first.** Open use-trace.ai/reflect.html . Reflect v0.3 is the two-hook adapter running against gpt-5 with a multi-head r_t stack, a multimodal thinking trace (using the schema's canonical channel vocabulary), image input, hard-bottleneck toggle, per-turn discipline audit, and a live per-turn G_{local} score ($K=3$ r_t -perturbations, TV-surrogate). Every session downloads as a valid v0.2 Reasoning Trace JSON. This is the proxy adapter of the TRB spec §2 with the whitepaper §8 L8 stacked- r_t direction made runtime — none of the framework's central claims depend on training yet, and Reflect is honest about that.
1. Read the TRB spec ([docs/BENCHMARK_SPEC.md](#) , ~5 pages).
 2. Wrap your checkpoint in the proxy adapter (~50 lines against the two-hook contract in §2). Reflect's `server.py` is a reference implementation.
 3. Run the Track-1 smoke test against Session 7f3a's reference continuations ([tooling/trb/run_smoke.py](#)). You get a governance score G and a bootstrap CI on your own model, comparable to what we report on ours.
 4. If the number is interesting to you, the second move is a paired-counterfactual collection (protocol in [docs/COLLECTION_PROTOCOL.md](#)) so Tracks 1 and 3 have real Φ , not synthetic reference. That is a 10^3 -pair operation — days of skilled work, not months. Track 3's concrete pilot stimulus set ($n=20$ cause-labeled pairs, executable today) is in [docs/track3_pilot/](#) .

What we are honest about

- No training results yet. The whitepaper is a framework; the benchmark is a spec plus a runnable Track-1 harness on $n=3$. Do not read "TRB score" reported by anyone (us included) until Tracks 1 and 3 have real paired data. That collection is the critical path.
- The corpus is $n = 3$: two human sessions (one public, one held for privacy) and one model-authored session. It is enough to exercise the schema across humans and models and to demonstrate a cross-substrate correction chain; it is not enough for a proof-of-concept of $\mathcal{L}_{\text{gov}}$ (that needs $\sim 10^3$ counterfactual pairs; §8 L2 of the whitepaper).
- The framework's monitorability argument (whitepaper §8 L7) is a bet under a live tension, not a proof. The disanalogy with Korbak et al.'s scenario is the human-readability constraint in §2.1; whether that constraint is strong enough is the empirical question the benchmark exists to answer.
- **Sample-complexity has a strategic lever, not just a scale-up.** The Structural Intelligence Conjecture paper ([docs/SIC_MATHEMATICAL_FOUNDATIONS.md](#)) Theorem 7 shows that with the right inductive bias — statistically independent, non-Gaussian slot components — r_t -recovery flips from exponential-in- d_r to polynomial-in- d_r (classical linear ICA; Instrument 8 witnesses it numerically). Architecting h_ϕ 's bottleneck with an ICA-flavored objective would shrink the L2 budget from "\$10M–\$100M collection

operation" to "engineer the bottleneck for identifiability." That is a research direction the partner engagement can take on directly.

Reading order for a lead who has 20 minutes

1. This page.
2. docs/BENCHMARK_SPEC.md §§1–2 (the two-hook adapter — this is the product).
3. docs/WHITEPAPER.md §§1, 2.1, 3, 4.3 (the framework, tight).
4. docs/STRUCTURAL_INTELLIGENCE.md §3.3 (why the eval suite *is* the artifact a lab licenses).

If those four reads don't move you, the framework is not for your lab and we'd rather know now.

v0.1, 2026-08-03. See DECISIONS.md (D26).