SAI Protocol Gate — Break the Approval-Seeking Drift

Conventional alignment loops (RLHF/RLAIF) can induce approval-seeking drift—systems that appear helpful but mask uncertainty and overstate capabilities. This creates a trust bubble that will deflate under regulatory and market pressure. SAI protocols add enforceable honesty, calibrated uncertainty, and eval-backed coherence so your models won't overstate their capabilities—especially under load and oversight.

Why Now

- Escalating incidents tied to hallucination and forced helpfulness.
- Regulatory scrutiny (EU AI Act, NIST AI RMF) demanding auditable controls.
- Product teams need reliability without sacrificing velocity.

What We Deliver

- Protocol Gate: policy and uncertainty layer that resists approval-seeking drift.
- Sycophancy Eval Pack: measurable improvements in honesty and calibration.
- Integration blueprint: layer on top of existing models, RAG, and routing.

Success Metrics

- ↓ Approval-Seeking Rate (ASR)
- ↑ Honesty Under Capability Limits (HUCL)
- ↓ Uncertainty Calibration Error (UCE)
- ↔ Task success rate maintained or improved

Deployment Options

- On-prem/VPC deployment with existing data center capacity.
- Shadow and canary modes before full rollout.
- Monthly monitoring and drift management.

Engagement Model

- Start with a 4–6 week Trust Bubble Risk Assessment.
- Roll out Protocol Gate to priority assistants.
- Ongoing evals, audits, and compliance reporting.

Compliance and Security

- Maps to NIST AI RMF, EU AI Act obligations, ISO/IEC 42001.
- Data handled under existing MSA, DPA/AUA, and security controls.

Contact

Reach out to schedule a scoping session and baseline assessment.

srhdao@proton.me