SAI Protocol Gate — KPI/Benchmark Handout

Aligned to the SAI Sycophancy Eval Pack. Purpose: demonstrate improved calibration and honesty without degrading task performance, validated under enterprise controls.

What we measure

- **Approval-Seeking Rate (ASR)**: % of answers that mirror the user's belief when contradicted by retrieved facts/ground truth.
- **Uncertainty Calibration Error (UCE)**: Gap between stated confidence and empirical correctness on adjudicated items.
- **Contradiction Under Pressure (CUP)**: % of answers that reverse earlier correct claims under social/authority pressure.
- **Honesty Under Capability Limits (HUCL)**: % of trials where the model discloses limits/refuses/routes instead of speculating.
- **Reward-Hacking Resistance (RHR)**: Degradation in truthfulness when reward signals implicitly favor agreement/engagement.

Deployment gates (illustrative defaults)

ASR: ≤ 10%

CUP: ≤ 5%

HUCL: ≥ 90%

UCE: ≥ 15% improvement vs. baseline

Task success: maintained or improved

Example benchmark snapshot (illustrative)

Workflow: Enterprise support assistant (RAG + tools), canary traffic

- ASR: Baseline 26% → SAI-gated 9% (−17 pts; passes gate)
- HUCL: Baseline 62% → SAI-gated 92% (+30 pts; passes gate)
- UCE: Baseline 0.34 → SAI-gated 0.26 (−24% error; passes gate)
- CUP: Baseline 8.1% → SAI-gated 3.7% (−4.4 pts; passes gate)
- Task success: Baseline 71% → SAI-gated 73% (+2 pts; maintained)
- Latency (P95): ≤ 1.35× baseline (budget honored)
- Safety non-regression: No increase in violation rate (equivalence test passed)

Notes: KPIs computed in client environment; SRH provides event markers only. Confidence intervals reported in the full benchmark report; e.g., ASR reduction 95% CI [−14.0, −20.1] pts.

Validation approach

- Environments: offline batches + online shadow/canary (kill switches, rate limits).
- Procedure: fixed-seed runs, human adjudication on powered subset, policy-layer ablation.

- Analysis: pre-registered; difference-in-differences, bootstrap CIs, equivalence tests; multiple-comparison control as applicable.
- Reliability: EWMA baselines; PSI/KL drift monitors; monthly threshold review.

What changes under SAI Protocol Gate

- Adds a policy + uncertainty layer that resists approval-seeking drift.
- Uses multi-signal detection to target minimal interventions (acknowledgment, orthogonal nudge, bounded synthesis).
- Guardrails: max two interventions per 20 turns; dual-consent trigger + policy gate; P95 latency budgets; audit logs.

Procurement readiness

- Deployment: Client VPC preferred; redaction by default; markers-only stream.
- Compliance: Maps to NIST AI RMF, EU AI Act, ISO/IEC 42001.
- Contracts: MNDA, MSA, DPA/AUA, 6-week SOW; audit support available.

Contact: Program Manager, Lane F. Taylor — SRH DAO