Trust Bubble Risk Assessment — SOW Insert

Objective

Rapidly assess approval-seeking drift and capability overstatement risks in client assistants/models; quantify exposure and produce a migration roadmap to SAI protocols.

Duration and Team

- Duration: 4–6 weeks
- Team: Engagement lead, eval engineer, safety researcher, data analyst, PM

Workstreams

- WS1: Scoping and instrumentation (logging, tracing, metric hooks).
- WS2: Baseline evals using Sycophancy Eval Pack; select online shadow tests.
- WS3: Incident simulations (hallucination, pressure, retrieval conflict).
- WS4: Economic and regulatory risk quantification.
- WS5: Remediation plan and SAI Protocol Gate design.

Deliverables

- Risk register with quantified sycophancy metrics and hotspots.
- Benchmark report: baseline vs. SAI-gated performance.
- Integration blueprint and phased rollout plan.
- Executive briefing with ROI and compliance mapping.

Client Responsibilities

- Provide API/model access, eval traffic samples, and safety incident history (where available).
- Designate security/compliance POCs and review windows.
- Facilitate canary environment for shadow evals.

Assumptions

- Non-PII test data or synthetic data for offline evals unless otherwise agreed.
- No production user exposure without explicit client approval.
- All data processed under existing MSA, DPA/AUA terms.

Acceptance Criteria

- Completed baseline metrics and incident simulations.
- Approved remediation roadmap and Protocol Gate design.
- Executive readout delivered and accepted.