

TECHNICAL APPENDIX — MENS-SCCA × APEX

Instrumentation architecture and measurement layer

DESIGN PRINCIPLE

MENS-SCCA is treated as a research instrumentation layer rather than an execution platform. The adapter boundary is intentional: APEX remains responsible for task execution and its authorized environment; MENS-SCCA receives only the observables needed for the experimental question.

OBSERVABLE MODEL

The instrumentation represents observable state, event identity, temporal sequence, evidence references, transformations, tool interactions, errors, recovery events, provenance, and evaluation outputs.

A minimal longitudinal record should support reconstruction of the relevant trajectory without requiring private reasoning traces.

ADAPTER ARCHITECTURE

APEX / authorized environment → extraction of authorized observables → ingestion adapter → canonical trajectory representation → metrics/registry → aggregate analysis.

The adapter isolates research instrumentation from Mercor's production systems and reduces coupling. It can be reviewed or replaced independently.

CORE DATA OBJECTS

Task identity; run identity; agent/model identity; condition; environment/version; tool configuration; initial state; event sequence; state transitions; evidence references; perturbation/recovery markers; final artifact; rubric version; judge version; evaluation result; provenance.

LONGITUDINAL REPRESENTATION

The methodological purpose is to make explicit properties that are otherwise compressed into a final score: preservation, transformation, propagation, recovery, consistency, and transfer.

Trajectory reconstruction is based on observable events and state transitions. The design does not require collection of hidden chain-of-thought.

EXPLORATORY METRIC DEFINITIONS

Outcome score: weighted criterion score when supported by the rubric.

Trajectory Integrity (TI): proportion of required observable properties preserved across the trajectory.

Transformation Fidelity (FT): preserved relevant properties divided by preserved + lost + ambiguous properties.

Failure Propagation (FP): downstream affected properties divided by downstream properties at risk.

Recovery Ratio (RR): post-recovery performance relative to an appropriate baseline.

Transfer Loss (TL): one minus normalized preservation under transfer.

Consistency Loss (CL): one minus normalized cross-run consistency.

These are candidate operationalizations, not validated universal metrics.

VALIDATION REQUIREMENTS

Before any composite score is used, test metric stability, sensitivity to perturbations, redundancy, inter-run reproducibility, agreement with expert assessment, and dependence on instrumentation choices.

The research registry should retain raw observable records, metric versions, transformations, evaluator versions, and analysis versions sufficiently to reproduce derived results.

ENGINEERING BOUNDARIES

Production deployment, API integration, compute scaling, authentication, security hardening, and infrastructure auditing are engineering tasks distinct from the principal investigator's methodological role.

MENS-SCCA should not require privileged or indiscriminate access to Mercor systems. The research design favors least-privilege access and minimal data extraction.