

PRELIMINARY EXPERIMENTAL PROTOCOL — MENS-SCCA × APEX

Protocol skeleton for preregistration and confirmatory execution

1. RESEARCH QUESTION

Determine whether longitudinal trajectory instrumentation adds reproducible information beyond final-outcome evaluation.

2. UNIT OF EXPERIMENT

Agent × Task × Run × Condition.

3. DOMAINS

When authorized: Investment Banking, Management Consulting, and Corporate Law.

4. VARIABLES

Outcome variables: final score, pass/fail criteria, expert acceptance, compliance with constraints, and externally verifiable state changes.

Trajectory variables: observable actions, tools, transformations, preservation, errors, propagation, recovery, consistency, and transfer.

Controls: task, agent/model, condition, environment, tools, initial information, constraints, rubric version, and evaluator versions.

5. CONDITIONS

BASELINE, PERTURBATION, RECOVERY, and TRANSFER, as defined in the main proposal.

6. MEASUREMENT

Exploratory constructs: outcome, trajectory integrity, transformation fidelity, failure propagation, recovery, cross-run consistency, and transfer loss/preservation.

No composite index is assumed valid before stability, sensitivity, redundancy, and reproducibility are established.

7. EVALUATION

Deterministic verification; expert rubric; model-based judge. Disagreements are retained and analyzed.

8. PARTITIONS

DEV for development; PILOT for calibration; HELDOUT for confirmatory evaluation. The final sample size will depend on authorized task availability and required precision.

9. ANALYSIS

The confirmatory statistical plan will be fixed before confirmatory analysis. Planned analyses include condition differences, trajectory-outcome association, inter-run stability, expert/judge agreement, sensitivity, redundancy, failure modes, recovery, and transfer.

A single average will not be used to conceal critical failure categories.

10. CONTAMINATION AND PROVENANCE

Before confirmation, define development tasks, held-out tasks, training exposure, public artifacts, confidential data, and version controls. Mercor data access remains authorization-dependent.

11. REPRODUCIBILITY RECORD

Each execution should, where authorized, be associated with task_id, run_id, agent/model, condition, environment/version, tool configuration, rubric version, judge version, trajectory record, final artifact, and evaluation result.

12. SUCCESS AND NON-SUCCESS CRITERIA

Favorable evidence requires stability, relevant discrimination, incremental information beyond outcome, reproducible relation to expert criteria, and robustness to sensitivity analysis.

Insufficient evidence includes redundancy, replication failure, excessive judge dependence, excessive instrumentation sensitivity, or no improvement in failure identification/explanation.

13. METHODOLOGICAL ORDER

Question → hypotheses → specification → governance → preregistration → calibration → instrument validation → pilot → execution → confirmatory analysis → robustness → interpretation.