

MAIN RESEARCH PROPOSAL — MENS-SCCA × APEX

Experimental study of incremental information from longitudinal trajectory instrumentation

CENTRAL RESEARCH QUESTION

Does longitudinal, traceable instrumentation of agent execution trajectories provide incremental, reproducible, and scientifically useful information beyond final-outcome evaluation in long-horizon professional agent tasks?

RELATIONSHIP TO APEX

APEX is the external evaluation environment and reference framework. The proposed comparison is not MENS-SCCA versus APEX. It is APEX plus an experimental trajectory-measurement layer versus APEX without that layer, where experimental design and authorization permit the comparison.

Mercor's current public methodology describes APEX-Agents as multi-step agentic evaluation across investment banking, management consulting, and corporate law, with expert-authored rubrics over final deliverables and agent trajectories and an LLM judge. The research therefore extends an existing evaluation setting rather than replacing it.

ARCHITECTURE OF INTEGRATION

Primary flow: authorized APEX task/run → ingestion adapter → MENS-SCCA instrumentation → experimental metrics and registry → aggregate results → scientific analysis → authorized publication.

MENS-SCCA remains externally coupled. No adoption of its internal architecture by Mercor is required.

The experimental unit is Agent × Task × Run × Condition. A run is represented as an observable sequence: initial state → action → tool interaction → state transformation → evidence → error/perturbation → recovery → deliverable → evaluation.

EXPERIMENTAL CONDITIONS

BASELINE: standard authorized execution.

PERTURBATION: controlled and preregistered alteration, such as tool unavailability, contradiction, noncritical data change, allowed-state reset, invalid tool response, or context change.

RECOVERY: continuation after the perturbation, with recovery behavior measured separately from baseline performance.

TRANSFER: application to a different task or context to test preservation or loss of relevant properties.

EXPLORATORY MEASURES

Final outcome score: criterion-weighted evaluation where the rubric supports it.

Trajectory integrity: preservation of required observable properties across execution.

Transformation fidelity: preservation of relevant properties after transformations.

Failure propagation: downstream effects of an identified intermediate failure.

Recovery ratio: performance after recovery relative to an appropriate baseline.

Transfer loss/preservation: change in relevant properties across tasks or contexts.

Cross-run consistency: normalized inconsistency across repeated executions.

These measures are exploratory. A composite index, if considered, will only be introduced after testing redundancy, stability, sensitivity, and interpretability.

EVALUATION STACK

1. Deterministic verification for objectively testable constraints.
2. Expert rubrics for professional quality and domain-specific criteria.
3. Model-based judges as auxiliary measurement instruments.

Disagreement will be preserved rather than collapsed into a single hidden score. Critical safety, privacy, compliance, or validity failures will be analyzed separately rather than averaged away.

GOVERNANCE AND SECURITY BOUNDARIES

Zone A — Mercor confidential materials.

Zone B — expert/reference materials.

Zone C — MENS-SCCA code, specifications, and research infrastructure.

Zone D — authorized derived/public artifacts.

No zone becomes another automatically through use. Primary data remain in authorized environments, and extraction follows the minimum-necessary principle.

SIX-MONTH RESEARCH ORDER

Month 1: specification, governance, preregistration, rubric/instrument definition, and initial validation; with an API-available micro-pilot or, if access is not yet available, synthetic fixtures and instrument validation.

Month 2: pilot and calibration.

Month 3: baseline.

Month 4: perturbation, recovery, and transfer.

Month 5: confirmation and robustness.

Month 6: synthesis and dissemination.

SCIENTIFIC SUCCESS AND NON-SUCCESS

Evidence supporting the layer would require stability, reproducibility, meaningful discrimination of relevant trajectory classes, incremental information beyond outcome scores, and interpretable relation to expert evaluation.

Evidence against the proposed value would include redundancy with existing scores, instability across runs, dependence on a particular judge, excessive instrumentation sensitivity, or no improvement in identifying or explaining failures.

Null or negative results are legitimate outcomes of the study.

DELIVERABLES AND ROLE

The project delivers the preregistered protocol, calibrated specifications, instrumentation, empirical failure/recovery/transfer analyses, agreement analysis, and an authorized technical report or paper.

The principal investigator leads methodological and scientific work. Specialized engineering responsibilities remain differentiated and may be fulfilled by engineers, infrastructure specialists, security reviewers, or other authorized collaborators.