

RESEARCH PITCH — MERCOR RESEARCH FELLOWSHIP · APEX

Traceable Evaluation of Frontier Agents on Long-Horizon Professional Tasks

PURPOSE

APEX-Agents evaluates frontier AI agents on long-horizon, cross-application professional tasks in investment banking, management consulting, and corporate law. Mercor's public methodology describes expert-authored rubrics applied to final deliverables and agent trajectories.

This research asks a complementary question: how much additional, reproducible scientific information can be obtained when execution trajectories are instrumented longitudinally in addition to evaluating final outcomes?

The proposal does not seek to replace APEX or turn MENS-SCCA into an evaluation product. MENS-SCCA is used as decoupled methodological infrastructure for studying trajectories, states, transformations, and evidence.

MEASUREMENT GAP

Two agents can reach similar final outcomes through materially different execution paths. One may preserve relevant information throughout; another may introduce an early error, lose information, and recover later. A final score can compress these differences.

The study therefore examines preservation and transformation of relevant information, error propagation, recovery after controlled perturbations, cross-run consistency, transfer loss or preservation, and the relationship between trajectory and final outcome.

HYPOTHESES AND FALSIFIABILITY

H1 — Trajectory integrity: executions with equivalent final outcomes can show systematic differences in observable trajectory integrity and error propagation.

H2 — Intermediate failures: event-level instrumentation can identify failures that are not distinguishable from the aggregate outcome alone.

H3 — Transfer: changes across tasks or contexts can reveal information loss or preservation before complete outcome degradation becomes visible.

H4 — Hybrid evaluation: deterministic checks, expert rubrics, and model-based judges can provide more interpretable evidence when disagreements are retained and analyzed.

H5 — Recovery: controlled perturbations can distinguish simple task completion from recovery capacity under changes or failures.

The proposal is explicitly falsifiable. If the added measures are redundant, unstable, poorly reproducible, or do not add information beyond conventional evaluation, that result will count against the proposed incremental value.

EXPERIMENTAL DESIGN

The study will use authorized long-horizon professional tasks in the APEX-Agents domains. Each task will specify scenario, objective, source materials, tools and constraints, initial conditions, expected deliverable, observable evidence, failure/recovery conditions, and rubric.

Where infrastructure and authorization permit, conditions will include baseline, controlled perturbation, recovery, and transfer.

The unit of analysis is Agent × Task × Run × Condition. Only authorized observable events, states, tools, transformations, errors, recovery events, and outcomes are required; private chain-of-thought is not required.

DEV, PILOT, and HELDOUT partitions will be used to reduce overfitting and contamination.

MEASUREMENT AND VALIDATION

Evaluation will combine deterministic verification where objectively testable, expert-authored or expert-calibrated rubrics, and model-based judges treated as measurement instruments rather than independent truth.

Exploratory dimensions include final outcome, trajectory integrity, transformation fidelity, failure propagation, recovery, cross-run consistency, and transfer.

These dimensions will not be assumed to be universal metrics. Stability, sensitivity, redundancy, reproducibility, and relation to expert evaluation will be tested first. Disagreements among evaluation mechanisms will be retained as research data.

GOVERNANCE AND REPRODUCIBILITY

Data governance will be fixed before confirmatory execution. Mercor confidential materials, expert/reference materials, MENS-SCCA infrastructure, and authorized public derivatives will remain separated.

Primary Mercor data will remain in authorized environments. Provenance, transformations, training exposure, publication status, and usage restrictions will be logged.

Hypotheses, success/non-success criteria, partitions, and the confirmatory statistical plan will be preregistered before confirmatory analysis.

SIX-MONTH PLAN — 30 H/WEEK

Month 1 — Specification and preparation: tasks, variables, protocol, governance, preregistration, rubrics, instrumentation specification, synthetic fixtures, and initial validation. If APIs are available, run a micro-pilot; otherwise proceed with synthetic fixtures and instrument validation.

Month 2 — Pilot and calibration.

Month 3 — Authorized baseline executions.

Month 4 — Controlled perturbation, recovery, and transfer.

Month 5 — Confirmatory and robustness analysis.

Month 6 — Synthesis, technical report/paper, and authorized reproducibility package.

DELIVERABLES

1. Experimental protocol and preregistration.
2. Task, variable, rubric, and instrumentation specifications.
3. Data-boundary and provenance contract.
4. Longitudinal execution instrumentation/harness.
5. Empirical failure, recovery, and transfer taxonomy.
6. Baseline, perturbation, recovery, and transfer results where authorized.
7. Agreement/disagreement analysis across deterministic checks, experts, and model judges.
8. Technical report or paper including positive, negative, and limitation findings.
9. Code, fixtures, derived data, or public documentation only to the extent authorized.

EXPECTED CONTRIBUTION

The scientific contribution is evidence about whether longitudinal, traceable measurement adds incremental, reproducible, and useful information about frontier agents beyond final-outcome evaluation.

If supported, the study yields a complementary protocol and instrumentation approach. If not supported, it identifies which forms of trajectory instrumentation do not justify their complexity and under what conditions.