

ASA-MRS-001

Design and Evaluation Manifest v1.0

STATUS: DESIGN FROZEN / NO OUTCOME OBSERVED / MODEL ROSTER SEAL PENDING

Cryptographic commitment

SHA-256 of the canonical public JSON manifest:

cf97f36a89553e53fde501d50dfa0c2413ca8a6ccae39ea6a51abdedc1932c0

Research question

Can a model be identified from the geometry of its recovery trajectory across unseen tasks, without access to its name, architecture or final answer?

Fixed experimental design

Fixed element	Registered value
Model families	3 (exact builds sealed in a separate pre-run roster)
Domains	4; leave-one-domain-out evaluation
Tasks	6 templates per domain
Perturbations	2 families
Repetitions	2 per model-domain-task-perturbation cell
Total	288 planned recovery trajectories

Domains and perturbations

Domains: constraint satisfaction; structured planning; causal diagnosis; evidence-bounded revision.

Perturbations: contradictory but resolvable evidence; constraint or goal revision.

Primary endpoint

Macro balanced accuracy for blinded model-family attribution, pooled across four leave-one-domain-out folds. Uncertainty: 95% stratified bootstrap interval with 10,000 resamples.

How the claim can pass - or fail

Registered baselines

Random uniform; majority class; final-answer-only; unordered task features; and an ordered sequence baseline without ASA recovery structure.

Primary success rule - every condition is required

1. Macro balanced accuracy is at least 0.45.
2. Absolute advantage over the strongest registered baseline is at least 0.10.
3. The 95% bootstrap lower bound exceeds the strongest baseline point estimate.
4. Advantage is positive in at least three of four held-out-domain folds.
5. No critical leakage or blinding violation occurs.

Failure and falsification

Failure is recorded if the complete success rule is not met. The model-signature explanation is further weakened if task identity explains attribution better; formatting normalization or paraphrase destroys the effect; only final-answer content carries it; or any identity leakage, post-reveal tuning or unregistered exclusion occurs.

Hostile controls

Label permutation; formatting normalization; prompt paraphrase; matched task surfaces; trajectory truncation; non-causal metadata reordering; and model-domain swap audit.

Exclusions and stop rules

Only transport failure before output, an irrecoverably corrupt record, or a pre-reveal protocol violation may be excluded. Weak performance, unusual behavior and outlier status are not exclusions. Stop only for confirmed leakage, more than 10% non-analytical data loss, or a systematic runner defect. Never stop because the result looks negative.

Freeze and reveal order

Publish manifest and hash; seal exact model builds separately; generate opaque IDs; execute blind; freeze and hash attribution; reveal identity; evaluate once; publish PASS, FAIL or INCONCLUSIVE with all exclusions and limitations.

Claim boundary

Passing would support a transferable recovery signature under this bounded design. It would not establish universal model identification, production readiness, mechanistic explanation or access to hidden reasoning.

Immutable record

This file is immutable. Any change creates version 1.1 or a successor study, with a new hash and an explicit change log. No amendment may replace this frozen record.