

ASA-MRS-001

Model Recovery Signature - Blind Cross-Domain Attribution Test

STATUS: PROTOCOL PUBLISHED / EXECUTION MANIFEST NOT YET FROZEN

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?

Primary hypothesis

A model may leave a relatively persistent recovery signature in the way it detects error, changes direction, absorbs perturbation and returns toward a stable solution. If this geometry is partly model-specific, blinded trajectory records should support above-baseline attribution across tasks and domains.

Competing explanations

- H1 - model signature: recovery geometry transfers across domains.
- H2 - task signature: apparent identity disappears when tasks change.
- H3 - interaction signature: the stable unit is model-by-domain, not model alone.
- H0 - no transferable signature: attribution does not exceed registered baselines.

Experimental design

Layer	Design requirement
Models	At least three model families; identities hidden during analysis
Domains	At least three task domains with distinct surface content
Trials	Repeated perturbation and recovery sequences per model-domain pair
Blinding	Randomized opaque run IDs; ground truth sealed until attribution is frozen
Holdout	Unseen tasks and at least one held-out domain

Observable trajectory only

ASA receives time-ordered external records of attempts, perturbations, corrections and recovery states. It does not receive model names, provider labels, architecture details, hidden reasoning traces or the final ground-truth identity during analysis.

What will be tested

Trajectory representation

The registered measurement layer will describe externally observable change: correction timing, direction changes, recovery depth, persistence after perturbation, stability after correction and sensitivity to repeated disruption. Exact private ASA Core metrics and parameters remain undisclosed.

Baselines

At minimum: random attribution; majority-class attribution; final-answer-only attribution; task-feature attribution without trajectory order; and a sequence baseline that does not use ASA recovery structure.

Primary evaluation

Held-out attribution performance across unseen tasks and a held-out domain, compared with all registered baselines. Confidence intervals, class balance and confusion structure will be reported. Training-domain success alone will not support the primary claim.

Falsification conditions

The model-signature hypothesis is weakened or rejected if attribution falls to baseline on held-out domains; task identity explains the result better than model identity; minor prompt changes destroy the grouping; only final-answer content carries the signal; or performance depends on data leakage, provider-specific formatting or unblinded labels.

Hostile controls

Label permutation; formatting normalization; prompt paraphrase; matched tasks; deliberately misleading surface style; truncated trajectories; reordered non-causal metadata; and model-domain swaps.

Freeze sequence

1. Select models, domains, perturbation families and sample counts.
2. Register baselines, exclusions, primary measure and failure thresholds.
3. Generate and hash the execution manifest.
4. Seal ground truth and randomize run identifiers.
5. Run ASA and freeze attribution outputs.
6. Reveal identity, evaluate and publish pass or fail.

Claim boundary

Protocol publication is not a positive result. ASA-MRS-001 will support a model recovery signature only if the pre-frozen held-out test exceeds registered baselines and survives the stated hostile controls. The execution manifest, hash, model set, sample counts and reveal date will be published when frozen.

Public commitment

The protocol, aggregate outcome, baseline comparison, errors and limitations will remain public. The ASA Core source, protected scoring logic, parameters, prompts and operational configuration remain private.