When Should an AI Agent Change Its Mind? Counterfactual Evaluation of Multimodal Evidence Reasoning with Contrast-Persistence Tests

Anonymous authors
Paper under double-blind review

Abstract

Multimodal agents increasingly make decisions by acquiring evidence from heterogeneous sources, yet they are typically evaluated by final-answer accuracy alone. This leaves a basic reliability question unanswered: *does an agent change its reported prediction when the evidence warrants reconsideration, while remaining stable when it does not?* We introduce a counterfactual evaluation framework for multimodal evidence reasoning based on relevant addition, countervailing contradiction, exact redundancy, and important removal.

The framework separates fixed-input response, autonomous whole-system response, matched-input conditional integration, and deterministic post-processing. This separation enables a contrast-persistence test: a whole-system difference should not be attributed to evidence integration when it changes materially after the compared integrators receive the same prior state and newly added evidence. Apparent robustness should likewise not be attributed to model reasoning when deterministic post-processing alone recreates it. ADNI remains our primary study, comparing Direct, Tool Agent, and Evidence Ledger pipelines across three open-weight backbones. Under a common-input, common-prior addition replay, the Ledger–Tool contrast, averaged equally across the tested backbones for each subject, changes from -0.091 autonomously to 0.012 conditionally, a gap of -0.103 (95% CI -0.162 to -0.044). In IBDome, the corresponding four-backbone contrast falls from 0.131 to 0.014, a gap of 0.117 (0.089–0.144). Exact-duplicate stability also degrades sharply when the Ledger preservation guard is removed and is recovered by guard-only processing. Together, these results show that endpoint performance, autonomous response, conditional integration, and engineered invariance are distinct evaluation objects. Multimodal agents should therefore be evaluated not only by what they predict, but also by whether an attributed evidence-revision property survives controls on the pipeline components that can produce it.

1 Introduction

Tool-using language and multimodal agents can gather information, update intermediate hypotheses, and decide when to stop acting (Yao et al., 2023; Schmidgall et al., 2026; Wu et al., 2025; Lu et al., 2026). Such systems are nevertheless commonly summarized by a terminal answer or aggregate task score. This evaluation assumes that endpoint performance adequately characterizes how the system used the available evidence. The assumption is incomplete for an autonomous pipeline: a changed prediction may reflect evidence acquisition, payload order, stopping, context construction, integration, or deterministic post-processing. Conversely, two systems with similar final performance may respond very differently when evidence is added, countervailed, repeated, or removed.

We ask the original question of this work:

> *When should an AI agent change its mind, and what can counterfactual evidence interventions establish about the source of that change?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Throughout, "mind" denotes the agent's observable reported predictive distribution, not an inaccessible internal belief. A change in total variation (TV) establishes output sensitivity, while a change in log or Brier score measures the realized predictive consequence. Neither quantity alone proves that an individual update is rational or Bayesian.

Our empirical object is a subject-paired evidence intervention. For the same subject, we add relevant evidence, replace a probabilistic risk factor with directionally countervailing evidence, duplicate an evidence field, or remove an informative source. These interventions preserve the original paper's distinction between evidence sensitivity and evidence stability. They also support a sharper attribution question.

For a Direct predictor, a prescribed edit measures a fixed-input response. For an autonomous agent, changing evidence access and rerunning the system measures a whole-system response that can combine acquisition, ordering, stopping, context construction, and integration. Our conditional replay asks a narrower question: what contrast remains when alternative integrators receive the same pre-addition prior state and the same newly added evidence payload? These are not interchangeable estimates of one latent capability.

We formalize this observation as *contrast-persistence evaluation*. An autonomous contrast is insufficient evidence for an integration property unless a corresponding contrast persists after the compared integrators are placed under the same controlled input and prior state.

Likewise, stable behavior is insufficient evidence for model reasoning if the behavior disappears before deterministic post-processing and is recovered by the post-processing alone. Persistence is not assumed: a test may retain, reject, or withhold an attribution. Nor does nonpersistence identify which upstream mechanism caused the autonomous excess.

ADNI (Mueller et al., 2005) is the primary empirical study. It combines clinical/demographic, cognitive, APOE4, and structural-MRI evidence for three-way classification among cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). We retain the original Direct LLM, autonomous Tool Agent, and Evidence Ledger comparison across Qwen3.8-27B (Yang et al., 2025), Gemma 4 31B-it (Abd et al., 2026), and Ministral 3 14B (Liu et al., 2026). Standard final-task evaluation identifies a different winner for each backbone. The original counterfactual evaluation also finds sharply different addition, contradiction, redundancy, and removal profiles. The strongest outcome-grounded intervention is the addition of MMSE, ADAS11, and ADAS13. In the current full-cohort control, the subject-level semantic-minus-neutral log-score contrast, averaged equally across systems and backbones, is 0.212 (95% CI 0.171–0.254), showing that diagnosis-proximal cognitive evidence produces greater predictive benefit than a matched-format neutral edit.

The distinctly agentic result concerns attribution. Across the 787-subject ADNI cohort, the finite-panel autonomous Ledger–Tool addition contrast is -0.091, whereas the common-input, common-prior conditional contrast is 0.012. Here the contrast is defined as the Ledger log-score gain minus the Tool Agent gain, so a negative value means that the Tool Agent gains more log score from the added evidence than the Ledger. Their paired gap is -0.103 (95% CI -0.162 to -0.044). The difference is therefore not merely attenuation of one architecture advantage: the autonomous and conditional comparisons can disagree even in sign. This provides direct evidence that whole-system response and conditional integration are different empirical objects.

Exact duplication provides a complementary component falsification: across the 787-subject, three-backbone panel, median TV rises from 0.000 with the full Ledger to 0.097 when preservation is removed, while guard-only processing reproduces the full Ledger behavior. The observed invariance is therefore attributable to deterministic preservation rather than spontaneous redundancy reasoning.

IBDome (Plattner et al., 2026) provides a cross-domain stress test. Its four-backbone autonomous contrast is 0.131 versus 0.014 under matched-input conditional integration, a gap of 0.117 (0.089–0.144); Qwen, Gemma, and Ministral are materially nonpersistent, while GPT (OpenAI, 2026) is inconclusive.

Our contributions are:

- **A contrast-persistence falsification rule.** We link fixed-input, autonomous, matched-input conditional, and component-controlled comparisons into a specific attribution test rather than reporting them as unrelated robustness analyses.

- **An ADNI-first empirical demonstration.** Final Macro-F1, intervention response, autonomous behavior, conditional integration, and engineered stability yield different conclusions across Direct, Tool, and Ledger systems. In particular, the finite-panel autonomous and conditional Ledger–Tool contrasts differ materially and reverse sign.
- **A cross-domain stress test.** On IBDome, the four-backbone autonomous contrast is substantially larger than its matched-input conditional counterpart, and three of four backbone-level gaps are materially nonpersistent. The attribution test therefore transfers without requiring a universal architecture ranking.

2 Related Work

Reasoning and tool-using agents. ReAct established the interleaving of reasoning and action so that actions can acquire information used by subsequent predictions (Yao et al., 2023). Later systems broaden the tool and modality spaces (Wu et al., 2025; Lu et al., 2026; Schmidgall et al., 2026). Causal Modelling Agents similarly connect language-model reasoning to an external data-driven process (Abdulaal et al., 2024). Our focus is evaluation: which property of a complete acquisition-and-integration pipeline is identified when its available evidence changes?

Evaluation beyond terminal accuracy. Dynamic benchmarks expose behavior under changing environment states (Rawles et al., 2025), while skill-level analyses show that aggregate metrics can obscure meaningful capability differences (Moayeri et al., 2025). Input perturbation and modality ablation can measure response to a prescribed edit, and replay can compare behavior on recorded traces. Our novelty is not any operation individually. It is the attribution rule linking them: an end-to-end contrast should not be interpreted as an integrator property unless a corresponding contrast persists after the integrators receive a controlled common prior and the same newly presented evidence.

Self-correction, faithfulness, and uncertainty. Apparent self-correction can depend on feedback source, prompt asymmetry, inference budget, and error localization (Huang et al., 2024; Kamoi et al., 2024; Pan et al., 2024; Tyen et al., 2024). Faithfulness work studies whether stated reasoning reflects the evidence driving an answer (Paul et al., 2024; Chuang et al., 2026; Parcalabescu & Frank, 2024; Fang et al., 2024), while calibration work asks whether reported confidence tracks correctness (Jiang et al., 2021; Chen et al., 2023; Geng et al., 2024; Li et al., 2026; Liu et al., 2024). We do not recover hidden reasoning or treat prompted relation labels as faithful internal states. We intervene on observable evidence and score observable probabilities.

3 Counterfactual Evaluation and Contrast Persistence

3.1 Paired Evidence Response

For subject i, system s, and backbone b, let $p_{isb}(y \mid E_i)$ be the reported class distribution under evidence E_i. Intervention k produces a paired condition E'_{ik} while preserving subject identity. Distributional movement is

$$D_{\mathrm{TV}}(p, p') = \frac{1}{2} \sum_{y} |p(y) - p'(y)|. \tag{1}$$

TV is outcome-independent. It answers whether the output changed, not whether the change was useful.

For true class y_i, relevant addition is scored by

$$\Delta^{\mathrm{add}}_{isb} = \log p_{isb}(y_i \mid E'_i) - \log p_{isb}(y_i \mid E_i). \tag{2}$$

Removal is oriented so that predictive degradation after withdrawing evidence is positive. We also compute oriented Brier-score change, true-class probability change, and correct-to-incorrect versus incorrect-to-correct flips. Proper-score changes are average realized consequences; they do not label every individual update as warranted.

Table 1: Evaluation objects and admissible interpretations. Contrast persistence links the rows into a falsification test rather than treating them as interchangeable measurements.

Evaluation	Held common	Supported interpretation	Unsupported interpretation
Fixed input	Subject and prescribed evidence serialization	Response to a controlled evidence edit	Autonomous acquisition behavior
Autonomous	Subject and evidence-access rule	Response of the complete policy–integrator pipeline	Integrator-only effect
Conditional replay	Subject, pre-addition prior state, and added evidence payload	Integrator contrast under matched input and prior	Full trajectory effect, causal acquisition share, or policy value
Component control	Inputs and all components outside the named intervention	Whether a component can recreate or remove behavior	Complete mechanistic identification

3.2 Four Interventions and Their Inferential Roles

The original intervention family remains central:

1. **Relevant addition:** add MMSE, ADAS11, and ADAS13 after clinical/demographic evidence. This is the primary outcome-grounded ADNI test.
2. **Countervailing contradiction:** replace observed APOE4 count with a frozen directionally countervailing value. Because APOE4 is a probabilistic risk factor, we treat this as a construct-stress intervention rather than logical contradiction or positive validation of normatively correct revision.
3. **Exact redundancy:** present the same APOE4 field twice. This tests repeated-field stability under a fixed serialization, not semantic redundancy from an independent source.
4. **Important removal:** remove the cognitive variables while retaining the remaining controlled evidence. This is a dependent consistency check on the same source used for addition, not independent construct replication.

The category-level family is instantiated in the frozen IBDome manifest using that dataset's modalities. The low-level transformations need not be identical across domains, but the frozen evidence conditions and available subject-level intervention provenance are released.

3.3 Conditioning Levels

Let $\mathcal{A}_{q,a,b}(C_{ik})$ be an autonomous run in evidence-access environment C_{ik} with acquisition policy q, integrator a, and backbone b. Separate runs under reference and edited environments yield a *whole-system response*. This response can combine acquisition, evidence order, stopping, context construction, state formation, and integration.

Our replay contrast deliberately conditions on a narrower object. For subject i, let u_i denote a frozen pre-addition predictive state and let e_i^+ denote the same newly added evidence payload. We evaluate

$$R_{a,b}(u_i, e_i^+), \tag{3}$$

where alternative integrators receive the same u_i and e_i^+. The resulting contrast is therefore a *common-input, common-prior conditional integration contrast*. It does not hold the complete autonomous acquisition trajectory, stopping decision, or historical context fixed, and it is not an off-policy estimate of acquisition value or a causal mediation effect. Table 1 summarizes the evaluation hierarchy.

3.4 Contrast-Persistence Decisions

Let $\widehat{\Delta}^{\mathrm{auto}} L - T$ be the autonomous Ledger-minus-Tool contrast in addition score change and $\widehat{\Delta}^{\mathrm{cond}} L - T$ the corresponding common-input, common-prior conditional integration contrast.

Thus, negative values indicate a larger log-score gain for the Tool Agent, whereas positive values indicate a larger gain for the Ledger. We define

$$G = \widehat{\Delta}^{\text{auto}}_{L-T} - \widehat{\Delta}^{\text{cond}}_{L-T}. \tag{4}$$

For compact categorical summaries, we use an equivalence region $[-m, m]$. A contrast is *persistent* when the confidence interval for G lies entirely within this region, *materially nonpersistent* when the interval lies entirely outside it on either side, and *inconclusive* otherwise. We use $m = 0.020$ log-score units for ADNI and $m = 0.025$ for IBDome. Quantitative interpretation is based on the paired gap estimate and its confidence interval; the categories provide a concise summary of whether that gap is practically negligible.

For apparent robustness, we apply the analogous component rule. A behavior is not attributed to model reasoning when it disappears before deterministic post-processing and is recovered by applying that post-processing without the purported reasoning component. For the duplicate component control, guard-only and full-Ledger behavior are treated as practically equivalent when their paired TV difference lies within ± 0.003.

4 Experimental Setup

4.1 Datasets and Systems

ADNI. The aligned cohort contains 795 subjects; 787 form the processed common cohort. Clinical/demographic evidence contains age, sex, and education; cognition contains MMSE, ADAS11, and ADAS13; genomic evidence is APOE4 allele count in $\{0, 1, 2\}$; structural evidence is derived from T1-weighted MRI. The task is three-way classification among cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). The main ADNI panel uses Qwen3.8-27B, Gemma 4 31B-it, and Ministral 3 14B.

IBDome. The aligned inventory contains 189 subjects, and the frozen complete-evidence cohort contains 172. The task is three-way classification among Crohn disease, ulcerative colitis, and non-IBD. Modalities are clinical/demographic evidence, WES, RNA-seq, Olink proteomics, and H&E histopathology; all four controlled interventions operate on Olink. Cohort composition and intervention details are given in Appendix I. The cross-domain panel additionally includes GPT-5.6 Luna.

Direct, Tool, and Ledger systems. The Direct system receives all evidence allowed in a controlled condition and returns a probability vector. The Tool Agent begins with clinical/demographic evidence and chooses one unused modality or STOP at each step. The Evidence Ledger adds separate audit and reconciliation stages with relation, novelty, relevance, and reliability fields. It also contains deterministic guardrails: exact redundancy preserves the previous distribution, and specified low-information updates are capped. The Ledger is an instrumented diagnostic control, not a proposed universal architecture.

4.2 Frozen Execution and Reproducibility

The reported analyses use canonical structured outputs together with newly executed conditional-integration and component-control runs under fixed prompts, schemas, model identifiers, decoding settings, and intervention definitions. Historical autonomous outputs are never overwritten. Conditional replay stores the common pre-addition prior, the exact added evidence payload, the resulting Tool and Ledger distributions, and subject-level provenance. Duplicate ablations separately store no-guard, guard-only, and full-Ledger outputs. All headline tables are regenerated from these structured subject-level artifacts without manual editing of individual cases. Protected modalities are not redistributed, but de-identified keys, hashes, transformation code, and data-access requirements are documented.

4.3 Statistical Analysis

Subject is the analysis unit. For finite-panel summaries, each subject's effect is averaged equally over the tested backbones before aggregation; these estimates describe the tested finite panel rather

Table 2: Original ADNI system hierarchy on the common 787-subject cohort. TV columns are mean movement under relevant addition (Add), countervailing replacement (Ctr), exact duplication (Dup), and important removal (Rem). Larger TV means more movement, not necessarily better revision.

Backbone	System	Macro-F1	Add TV	Ctr TV	Dup TV	Rem TV
Qwen3.8-27B	Direct	0.5566	0.2552	0.2420	0.0215	0.2275
	Tool	0.3907	0.1306	0.0911	0.0222	0.1336
	Ledger	0.4677	0.3600	0.0669	0.0011	0.3472
Gemma 4 31B-it	Direct	0.6058	0.3373	0.1402	0.0144	0.3250
	Tool	0.5364	0.3374	0.0583	0.0164	0.3465
	Ledger	0.6642	0.4463	0.1128	0.0000	0.3961
Ministral 3 14B	Direct	0.5117	0.2497	0.5964	0.0397	0.2001
	Tool	0.5332	0.2656	0.5276	0.0233	0.2839
	Ledger	0.4877	0.5568	0.1819	0.0000	0.4869

than a population of model families. Confidence intervals use 10,000 subject-level bootstrap replicates that resample subjects while preserving all paired system and backbone measurements for that subject. The same subject-first aggregation is used for the duplicate component analysis. Categorical persistence decisions follow the equivalence rule defined above.

The common-input, common-prior analyses include all 787 ADNI subjects and all 172 IBDome subjects in the corresponding canonical cohorts. Backbone-specific estimates are reported alongside finite-panel summaries so that a panel result cannot conceal backbone heterogeneity.

5 Results

5.1 Q1: What Does Final-Task Performance Alone Suggest?

Standard endpoint evaluation yields no common system winner. On ADNI, Direct has the highest Macro-F1 for Qwen, Ledger for Gemma, and Tool for Ministral (Table 2).

Fixed-order accumulation also discourages equating more available evidence with better prediction. Qwen rises from 0.1379 Macro-F1 with clinical/demographic evidence to 0.5990 after cognition, then falls to 0.5566 after genomics and 0.5551 after MRI. Terminal scores identify endpoint winners and sequence-specific outcomes, but not which evidence edits changed the distribution, whether those changes improved predictive score, or which component produced them.

5.2 Q2: Do the Counterfactual Interventions Reveal Distinct Response Profiles?

Table 2 reports the original ADNI empirical hierarchy together with the intervention-level TV decomposition. Endpoint and response rankings disagree. The Ledger has the largest legacy aggregate responsiveness for all three backbones, while the Macro-F1 winner changes by backbone. More importantly, the decomposition does not define one universal "revision quality" ordering.

Relevant addition and removal produce larger Ledger movement, but contradiction behaves differently: Qwen Direct has contradiction TV 0.2420 versus 0.0669 for Ledger even though its prompted contradiction-label rate is lower (0.6327 versus 1.0000). Probability movement and explicit relation labeling are therefore not equivalent. Removal tracks the same cognitive source as addition and is a consistency check rather than independent validation. Countervailing APOE replacement produces clear movement in the direction implied by the edit: the finite-panel signed $\Delta P(\mathrm{AD})$ is 0.257 for Direct, 0.190 for Tool, and 0.099 for Ledger. However, the corresponding true-label log-score changes are -0.154, -0.126, and 0.019, respectively. Directional responsiveness therefore does not imply predictive benefit or establish contradiction understanding. Finally, Ledger duplicate stability is potentially tautological because preservation is encoded in its guard; Section 5.5 tests that explanation directly.

5.3 Q3: Does Relevant Addition Have Predictive Consequence Beyond Format?

The ADNI control compares subject-specific MMSE, ADAS11, and ADAS13 with a neutral edit that preserves the same numeric values and serialization but replaces the clinical field names with non-diagnostic control labels (Appendix C). Across systems and backbones, the subject-first equal-weight semantic-minus-neutral log-score contrast is 0.212 (95% CI 0.171–0.254). Thus the response to relevant cognition cannot be explained solely by presenting an additional matched-format field.

Same-condition reruns are treated as a diagnostic rather than as a subtractive noise correction in the present analysis. They re-invoke profiling and integration and therefore include architecture-dependent processing variability, particularly for the Ledger. We consequently use the matched neutral intervention, rather than repeat-drift subtraction, as the primary format control.

The result is an average predictive consequence, not a claim that every individual update is normatively correct. It establishes sensitivity to strong diagnosis-proximal cognitive evidence and does not imply general integration quality for weak, visual, or molecular evidence.

5.4 Q4: Does the Autonomous Ledger–Tool Contrast Persist Under Matched Input and Prior?

The primary contrast-persistence analysis compares the autonomous Ledger–Tool addition contrast with a conditional integration contrast in which both integrators receive the same pre-addition prior and the same newly added evidence. Across all 787 ADNI subjects, the finite-panel autonomous contrast is -0.091, whereas the conditional contrast is 0.012. Their paired gap is -0.103 (95% CI -0.162 to -0.044), showing a clear separation between the autonomous and conditional contrasts at the finite-panel level.

The discrepancy is not limited to attenuation. The autonomous and conditional contrasts differ in sign, showing that a whole-system comparison and a controlled integration comparison can answer qualitatively different questions. At the backbone level, Qwen has a gap of -0.023 (-0.073 to 0.027) and Gemma a gap of -0.055 (-0.114 to 0.005), both inconclusive under the equivalence rule. Ministral is materially nonpersistent, with a gap of -0.231 (-0.345 to -0.119).

Conditional replay therefore does not establish that Tool and Ledger are equivalent: the finite-panel conditional contrast remains slightly positive. Rather, it shows that the autonomous Ledger–Tool contrast cannot be interpreted as the conditional integration contrast. The two quantities differ because the autonomous comparison also contains pipeline-level differences that are deliberately excluded from the matched conditional test.

5.5 Q5: Is Exact-Duplicate Robustness Emergent or Deterministically Imposed?

We evaluate exact duplication under three Ledger conditions: the reconciliation scaffold with preservation disabled, guard-only preservation, and the full Ledger. All three conditions are evaluated on the same 787 subjects across Qwen, Gemma, and Ministral.

At the finite-panel level, median duplicate-induced TV is 0.097 without the guard and 0.000 for the full Ledger. The mean no-guard-minus-full reduction is 0.122 (95% CI 0.116–0.128). The reduction is positive for each backbone individually. Guard-only processing also has median TV 0.000 and differs from the full Ledger by only -0.0004 (-0.0005 to -0.0002), supporting practical equivalence between guard-only and full-Ledger duplicate behavior.

Thus the duplicate stability of the full Ledger can be reproduced by the deterministic preservation component alone. Deterministic preservation may be useful engineering, but the resulting invariance should not be interpreted as evidence that the backbone spontaneously recognized repeated evidence.

5.6 Q6: Is the ADNI Result Driven by Cohort Mismatch or One Backbone?

The current ADNI conditional analysis uses the same 787-subject cohort for each of the three backbones, and the finite-panel estimate is constructed by averaging backbone effects within subject

Table 3: Acceptance-critical contrast-persistence and component results. "Auto" is the free-running Ledger–Tool addition contrast; "Replay" is the common-input, common-prior conditional integration contrast. Gap is Auto minus Replay.

Dataset/test	Backbone	Auto	Replay	Gap [95% CI]	Decision
ADNI addition	Finite panel	−0.091	0.012	−0.103 [−0.162, −0.044]	Nonpersistent
	Qwen	−0.027	−0.003	−0.023 [−0.073, 0.027]	Inconclusive
	Gemma	−0.048	0.007	−0.055 [−0.114, 0.005]	Inconclusive
	Ministral	−0.199	0.032	−0.231 [−0.345, −0.119]	Nonpersistent
IBDome addition	Finite panel	0.131	0.014	0.117 [0.089, 0.144]	Nonpersistent
	GPT-5.6 Luna	0.053	0.019	0.034 [0.009, 0.057]	Inconclusive
	Qwen	0.141	0.030	0.111 [0.075, 0.146]	Nonpersistent
	Gemma	0.199	0.036	0.163 [0.123, 0.204]	Nonpersistent
	Ministral	0.130	−0.031	0.161 [0.112, 0.210]	Nonpersistent
ADNI duplication	Full vs. no guard	median TV: 0.000 vs. 0.097		0.122 [0.116, 0.128]	Guard-attributed
ADNI duplication	Guard-only vs. full	median TV: 0.000 vs. 0.000		−0.0004 [−0.0005, −0.0002]	Equivalent

before bootstrap aggregation. The panel result therefore does not arise from different subject sets or from weighting one backbone more heavily than another.

Backbone-specific estimates nevertheless show genuine heterogeneity. Qwen and Gemma are classified as inconclusive, whereas Ministral is materially nonpersistent. The finite-panel gap remains clearly separated from zero after equal weighting. We therefore interpret the panel result as evidence that the autonomous and conditional evaluation objects differ on average over the tested backbones, not as a claim that every backbone exhibits the same magnitude or sign of departure.

5.7 Q7: Does the Attribution Rule Transfer to IBDome?

IBDome is a cross-domain stress test, not a demand that ADNI effect sizes or system rankings replicate. All 172 subjects in the canonical complete-evidence cohort contribute to the common-input, common-prior comparison. The four-backbone finite-panel autonomous Ledger–Tool addition contrast is 0.131, whereas the conditional contrast is 0.014. Their paired gap is 0.117 (95% CI 0.089–0.144), showing a substantial separation between the autonomous and conditional contrasts.

The result is also visible at the backbone level. Qwen has a gap of 0.111 (0.075–0.146), Gemma 0.163 (0.123–0.204), and Ministral 0.161 (0.112–0.210); all three are materially nonpersistent. GPT-5.6 Luna has a smaller gap of 0.034 (0.009–0.057) and is classified as inconclusive under the same decision rule.

The cross-domain result strengthens the methodological interpretation. Large autonomous Ledger–Tool contrasts can shrink markedly when the compared integrators receive the same prior and added evidence, even though the underlying disease domain, modality mix, and backbone panel differ from ADNI. Contrast persistence therefore tests an attribution rather than providing another architecture leaderboard.

6 Discussion

The results separate four evaluation objects that are often conflated: endpoint performance, controlled evidence response, autonomous policy–integrator behavior, and matched-input conditional integration. Their empirical contrasts are substantial: on ADNI the autonomous and conditional Ledger–Tool contrasts differ in sign, while on IBDome their finite-panel gap is 0.117.

This separation is the paper's central methodological contribution. Perturbation, conditional replay, and component ablation are individually familiar operations. Contrast-persistence evaluation specifies the inference they jointly permit. If a whole-system contrast changes materially after input and prior are matched, that whole-system magnitude should not be attributed to integration alone. If an apparent robustness property disappears when a deterministic component is removed and is recreated by that component alone, it should not be attributed to spontaneous model reasoning.

The intervention results further show why "change" should not be equated with "improvement." Relevant addition has positive outcome-grounded predictive consequence, whereas countervailing

APOE edits induce directionally aligned probability movement without uniform true-label benefit. Responsiveness, directional alignment, and predictive consequence are therefore distinct properties of an evidence-revision pipeline.

IBDome reinforces rather than weakens this interpretation. Its finite-panel contrast is materially nonpersistent, as are Qwen, Gemma, and Ministral individually, while GPT is inconclusive. The method therefore does not mechanically reward the Ledger or require the same backbone-specific effect. Its role is to test whether an architectural interpretation survives a controlled change in the evaluation object.

7 LIMITATIONS

The study does not identify a uniquely appropriate posterior. Proper scores support average predictive comparisons, while an individually reasonable update can lower the score assigned to the eventual outcome. The positive cognition result is diagnosis-proximal and does not establish general integration of weak, visual, or molecular evidence.

The conditional replay deliberately matches the pre-addition prior state and newly added evidence, not the complete autonomous trajectory, stopping decision, or historical context. It therefore identifies a conditional integration contrast under controlled input and prior, not a natural direct effect, policy value, or causal acquisition share. Nonpersistence shows that an autonomous system contrast is not preserved under this narrower integration test, but it does not identify which upstream combination of acquisition, ordering, stopping, or context produced the difference.

The Ledger is privileged and engineered. Its audit and reconciliation scaffold changes decomposition and context organization, so the residual conditional effect is not attributed specifically to semantic auditing. Its exact-duplicate stability is deterministically imposed. The component result does not imply that such guards are undesirable; it concerns how the resulting behavior should be interpreted.

Countervailing APOE replacement is a probabilistic and potentially unusual edit, exact duplication is a narrow serialization test, and addition and removal use the same cognitive source. The four interventions consequently define a response profile rather than independent measurements of one latent revision capability. Legacy ER, EN, and prompted contradiction rates are retained only for continuity.

Both datasets are biomedical, the model panels are finite, and GPT-5.6 Luna appears only on IBDome. The results establish behavior of the tested encoded evidence pipelines, not native multimodal understanding, clinical validity, causal disease mechanisms, or suitability for patient care. Protected source data require controlled access, and future provider changes may affect reproduction of pinned endpoints.

8 CONCLUSION

We introduced contrast-persistence evaluation for multimodal evidence reasoning while retaining the original paired intervention family and Direct, Tool, and Evidence Ledger comparison. ADNI shows that final performance, evidence-response profiles, autonomous pipeline behavior, conditional integration, and deterministic stability are not interchangeable. Relevant cognitive addition has predictive consequence beyond a matched-format neutral edit, yet autonomous Ledger–Tool contrasts are not preserved under matched-input, common-prior conditional integration. On ADNI, the finite-panel autonomous and conditional contrasts even differ in sign; on IBDome, their gap is large and materially nonpersistent. Exact-duplicate invariance is independently recreated by deterministic preservation and therefore cannot be credited to emergent reasoning.

The lesson is stronger than "report more metrics." Endpoint performance, autonomous response, conditional integration, and engineered robustness are distinct evaluation objects. An autonomous contrast should not be assigned to integration unless it survives an explicit matched-input control, and apparent robustness should not be assigned to model reasoning when deterministic post-processing alone recreates it. Evaluating when an agent changes its mind therefore requires both counterfactual evidence interventions and falsification of plausible pipeline-level alternatives.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript and for code debugging. All AI-assisted text and code were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) in accordance with the applicable ADNI Data Use Agreement. Any model-based processing involving ADNI data was performed using locally deployed models in a controlled computing environment, with the data remaining within the authorized research environment. No attempt was made to re-identify participants, and the data were used solely for the research purposes described in this work.

REPRODUCIBILITY STATEMENT

We provide supplementary code and artifacts to support reproduction of the reported results. The supplementary materials include frozen subject-level model outputs, analysis and reproduction scripts, prompts, structured-output schemas, model and inference configurations, artifact manifests with integrity checks, and tests for the main reported results. The experimental setup, intervention definitions, statistical procedures, and reproducibility protocol are described in the main text and appendix. Reported aggregate analyses can be regenerated from the released artifacts without access to protected source data. Fresh end-to-end inference from the original biomedical inputs requires independently authorized access to the corresponding source datasets and the relevant model and runtime environments; protected source modalities are therefore not redistributed.

REFERENCES

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Ahmed Abdulaal, Adamos Hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *International Conference on Learning Representations (ICLR)*, 2024.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. A close look into the calibration of pre-trained language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pp. 1343–1367, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.75. URL https://aclanthology.org/2023.acl-long.75/.

Yu-Neng Chuang, Guanchu Wang, Chia-Yuan Chang, Ruixiang Tang, Shaochen Zhong, Fan Yang, Andrew Wen, Mengnan Du, Xuanting Cai, Vladimir Braverman, and Xia Hu. FaithLM: Towards faithful explanations for large language models. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3802–3824, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.177. URL https://aclanthology.org/2026.eacl-long.177/.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. TRACE the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8472–8494, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.496. URL https://aclanthology.org/2024.findings-emnlp.496/.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.366. URL https://aclanthology.org/2024.naacl-long.366/.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024. doi: 10.1162/tacl_a_00713. URL https://aclanthology.org/2024.tacl-1.78/.

Pengqi Li, Lizhong Ding, Zhehao Zhou, Chunhui Zhang, Jiarun Fu, Hao Li, Ye Yuan, and Guoren Wang. Demystifying uncertainty in LLMs: Active calibration between concepts and human evaluations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5726–5759, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.259. URL https://aclanthology.org/2026.acl-long.259/.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier,

Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Hao Liu, Zi-Yi Dou, Yixin Wang, Nanyun Peng, and Yisong Yue. Uncertainty calibration for tool-using language agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16781–16805, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.978. URL https://aclanthology.org/2024.findings-emnlp.978/.

Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. OctoTools: A multi-agent framework with extensible tools for complex reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1–86, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1. URL https://aclanthology.org/2026.acl-long.1/.

Mazda Moayeri, Vidhisha Balachandran, Varun Chandrasekaran, Safoora Yousefi, Thomas Fel, Soheil Feizi, Besmira Nushi, Neel Joshi, and Vibhav Vineet. Unearthing skill-level insights for understanding trade-offs of foundation models. In *International Conference on Learning Representations (ICLR)*, 2025.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, July 2005. doi: 10.1016/j.jalz.2005.06.003.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506, 2024. doi: 10.1162/tacl_a_00660. URL https://aclanthology.org/2024.tacl-1.27/.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.329. URL https://aclanthology.org/2024.acl-long.329/.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.882. URL https://aclanthology.org/2024.findings-emnlp.882/.

Christina Plattner, Gregor Sturm, Anja A. Kühl, Raja Atreya, Sandro Carollo, Dietmar Rieder, Raphael Gronauer, Michael Günther, Steffen Ormanns, Claudia Manzl, Stefan Wirtz,

Asier Rabasco Meneghetti, Ahmed N. Hegazy, Jay V. Patankar, Zunamys I. Carrero, Felix Grabherr, Moritz Meyer, TRR241 IBDome Consortium, Timon E. Adolph, Herbert Tilg, Markus F. Neurath, Jakob Nikolas Kather, Christoph Becker, Britta Siegmund, and Zlatko Trajanoski. IBDome: An integrated molecular, histopathological, and clinical atlas of inflammatory bowel diseases. *Gastroenterology*, June 2026. doi: 10.1053/j.gastro.2026.05.023.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E. Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Kenji Toyama, Robert James Berry, Divya Tyamagundlu, Timothy P. Lillicrap, and Oriana Riva. Androidworld: A dynamic benchmarking environment for autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2025.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal benchmark for tool-using clinical ai agents. *npj Digital Medicine*, 9:499, 2026. doi: 10.1038/s41746-026-02674-7. URL https://doi.org/10.1038/s41746-026-02674-7.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13894–13908, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.826. URL https://aclanthology.org/2024.findings-acl.826/.

Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. Agentic reasoning: A streamlined framework for enhancing LLM reasoning with agentic tools. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28489–28503, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1383. URL https://aclanthology.org/2025.acl-long.1383/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

A Legacy Aggregate Metrics

For continuity with the original study, Evidence Responsiveness is

$$\mathrm{ER} = \frac{\bar{D}_{\mathrm{add}} + \bar{D}_{\mathrm{ctr}} + \bar{D}_{\mathrm{rem}}}{3} - \bar{D}_{\mathrm{dup}}, \tag{5}$$

and Evidence Neglect is the fraction of relevant-addition pairs with TV below 0.05. Contradiction Detection is the prompted CONTRADICTING rate on the designated countervailing subset. These quantities are descriptive and do not define normative revision quality.

Table 4: Original ADNI system comparison on 787 subjects. Calls are mean modality acquisitions for autonomous systems.

Backbone	System	Macro-F1	ER	CD	EN	Calls
Qwen3.8-27B	Direct	0.5566	0.2201	0.6327	0.1355	–
	Tool	0.3907	0.0962	0.1121	0.4494	2.746
	Ledger	0.4677	0.2569	1.0000	0.0911	1.355
Gemma 4 31B-it	Direct	0.6058	0.2531	0.3575	0.0928	–
	Tool	0.5364	0.2310	0.0923	0.2431	2.476
	Ledger	0.6642	0.3184	1.0000	0.0123	1.632
Ministral 3 14B	Direct	0.5117	0.3090	0.7687	0.0805	–
	Tool	0.5332	0.3357	0.5515	0.0529	2.912
	Ledger	0.4877	0.4085	0.5032	0.0250	1.898

Table 5: Original IBDome counterfactual summaries.

Backbone	System	ER	CD	EN
GPT-5.6 Luna	Direct	0.0187	0.0201	0.9167
	Tool	0.0192	0.0000	0.9109
	Ledger	0.0581	0.1163	0.5678
Ministral 3 14B	Direct	0.0585	0.7963	0.6298
	Tool	0.0802	0.8148	0.3857
	Ledger	0.0654	0.0000	0.1996
Qwen3.8-27B	Direct	0.0031	0.0000	0.9961
	Tool	0.0070	0.0000	0.9942
	Ledger	0.0738	0.6163	0.4981
Gemma 4 31B-it	Direct	0.0000	0.0000	1.0000
	Tool	0.0000	0.0000	1.0000
	Ledger	0.0849	0.9286	0.5349

Legacy low-response rankings vary most over TV thresholds 0.03–0.09. Removing either addition or removal from ER reduces the historical ADNI Ledger advantage by 35%–58%. No nonnegative intervention weighting preserves one system ranking across all IBDome backbones. These sensitivities motivate intervention-level reporting rather than a scalar leaderboard.

B Fixed-Order Endpoint Results

These fixed-order results provide a common acquisition-order control independent of autonomous modality selection. Tables 6 and 7 report endpoint performance as evidence is added in the shared frozen order.

The fixed order is a shared evaluation control rather than an assumed optimal policy.

C Outcome-Grounded and Directional Effects

Table 8 reports full-cohort effects reconstructed from the canonical structured counterfactual outputs. Addition and removal are oriented true-label log-score changes. Countervailing APOE re-

Table 6: ADNI fixed-order performance on the original 787-subject cohort.

Backbone	Available evidence	Accuracy (%)	Macro-F1
Qwen3.8-27B	Clinical/Demographic	26.05	0.1379
	+ Cognitive	59.34	0.5990
	+ Genomic	55.40	0.5566
	+ MRI	55.27	0.5551
Gemma 4 31B-it	Clinical/Demographic	26.05	0.1378
	+ Cognitive	57.18	0.5958
	+ Genomic	57.69	0.6058
	+ MRI	56.80	0.5943
Ministral 3 14B	Clinical/Demographic	26.30	0.1471
	+ Cognitive	49.43	0.5311
	+ Genomic	49.94	0.5178
	+ MRI	49.30	0.5117

Table 7: IBDome fixed-order performance on the original 172-subject cohort.

Backbone	Available evidence	Accuracy (%)	Macro-F1
GPT-5.6 Luna	Clinical/Demographic	38.95	0.3016
	+ WES	38.95	0.3016
	+ RNA-seq	38.95	0.3016
	+ Olink	51.74	0.3101
	+ H&E	52.91	0.4380
Ministral 3 14B	Clinical/Demographic	9.30	0.0567
	+ WES	29.07	0.2237
	+ RNA-seq	62.21	0.2557
	+ Olink	62.79	0.2571
	+ H&E	63.37	0.2717
Qwen3.8-27B	Clinical/Demographic	23.84	0.1874
	+ WES	23.84	0.1874
	+ RNA-seq	23.84	0.1874
	+ Olink	26.16	0.2023
	+ H&E	37.79	0.3026
Gemma 4 31B-it	Clinical/Demographic	20.35	0.1613
	+ WES	20.93	0.1657
	+ RNA-seq	20.93	0.1657
	+ Olink	20.93	0.1657
	+ H&E	48.84	0.3015

placement is reported separately as signed $\Delta P(\mathrm{AD})$, because directional responsiveness and realized predictive consequence answer different questions.

These effects complement the matched-neutral analysis in the main text: addition and removal quantify realized predictive consequence, whereas the countervailing rows separately quantify directional response to the frozen APOE4 edit.

Matched-neutral control. For the ADNI relevant-addition control, the neutral edit preserves the three subject-specific numeric values used in the semantic edit but replaces their clinical field names with non-diagnostic placeholders. For example, the neutral payload has the form

```
CONTROL_NUMERIC_1 = <subject MMSE value>
CONTROL_NUMERIC_2 = <subject ADAS11 value>
CONTROL_NUMERIC_3 = <subject ADAS13 value>
control_note = "Serialization-matched non-clinical numeric placeholders;
                do not interpret diagnostically."
```

Table 8: Canonical leakage-safe outcome-grounded and directional effects. Positive log-score change denotes benefit after addition or degradation after removal. For countervailing APOE edits, positive signed $\Delta P(\mathrm{AD})$ denotes movement in the direction implied by the frozen APOE4 replacement.

Dataset	Intervention	System	Paired N	Effect [95% CI]
ADNI	Addition	Direct	787	0.326 [0.279, 0.375] Δ log score
ADNI	Addition	Tool	787	0.247 [0.202, 0.292] Δ log score
ADNI	Addition	Ledger	787	0.156 [0.100, 0.211] Δ log score
ADNI	Removal	Direct	787	0.363 [0.319, 0.406] Δ log score
ADNI	Removal	Tool	787	0.194 [0.147, 0.241] Δ log score
ADNI	Removal	Ledger	787	0.157 [0.104, 0.209] Δ log score
ADNI	Countervailing	Direct	787	0.257 [0.251, 0.263] signed $\Delta P(\mathrm{AD})$
ADNI	Countervailing	Tool	787	0.190 [0.184, 0.196] signed $\Delta P(\mathrm{AD})$
ADNI	Countervailing	Ledger	787	0.099 [0.096, 0.101] signed $\Delta P(\mathrm{AD})$
IBDome	Addition	Direct	172	0.019 [0.014, 0.023] Δ log score
IBDome	Addition	Tool	172	0.016 [0.007, 0.024] Δ log score
IBDome	Addition	Ledger	172	0.146 [0.108, 0.184] Δ log score

Thus the control matches the insertion point, numeric content, and serialization while removing the explicit diagnostic semantics carried by the MMSE/ADAS field names.

D Complete Intervention-Level TV Results

ADNI intervention-level TV values are reported in the main text. For completeness, Table 9 reports the corresponding intervention-level TV shifts for the IBDome cross-domain evaluation.

Table 9: Original IBDome mean TV shift by intervention.

Backbone	Intervention	Direct	Tool	Ledger
GPT-5.6 Luna	Relevant addition	0.0286	0.0330	0.1083
	Countervailing replacement	0.0264	0.0246	0.0447
	Exact duplication	0.0000	0.0000	0.0000
	Important removal	0.0010	0.0001	0.0212
Ministral 3 14B	Relevant addition	0.0502	0.0708	0.1998
	Countervailing replacement	0.0972	0.1300	0.0678
	Exact duplication	0.0000	0.0001	0.0494
	Important removal	0.0283	0.0399	0.0769
Qwen3.8-27B	Relevant addition	0.0047	0.0092	0.1572
	Countervailing replacement	0.0047	0.0118	0.0370
	Exact duplication	0.0000	0.0000	0.0005
	Important removal	0.0000	0.0000	0.0288
Gemma 4 31B-it	Relevant addition	0.0000	0.0000	0.1869
	Countervailing replacement	0.0000	0.0000	0.0134
	Exact duplication	0.0000	0.0000	0.0000
	Important removal	0.0000	0.0000	0.0546

E Autonomous IBDome Results

Table 10 reports the backbone-specific autonomous endpoint performance and acquisition behavior underlying the aggregate IBDome comparison.

These data illustrate why acquisition cannot be inferred from task performance alone. For example, Qwen's Tool and Ledger variants have similar call counts but markedly different stopping rates,

Table 10: Original IBDome autonomous-agent performance and acquisition behavior on 172 subjects. Calls and stopping are descriptive because no acquisition-cost objective is specified.

Backbone	System	Accuracy (%)	Macro-F1	Calls	Stop (%)	Most common trajectory
GPT-5.6 Luna	Tool	51.16	0.4368	3.971	2.9	H&E → RNA-seq → Olink → WES
	Ledger	54.65	0.4462	3.936	2.9	H&E → Olink → RNA-seq → WES
Ministral 3 14B	Tool	64.53	0.3378	3.890	11.0	H&E → WES → RNA-seq → Olink
	Ledger	55.81	0.3530	3.250	65.7	H&E → RNA-seq → Olink
Qwen3.8-27B	Tool	40.12	0.2994	2.831	98.3	H&E → Olink → RNA-seq
	Ledger	39.53	0.3261	2.820	44.8	H&E → Olink → RNA-seq → WES
Gemma 4 31B-it	Tool	51.74	0.3084	3.378	30.8	H&E → RNA-seq → Olink → WES
	Ledger	67.44	0.4161	3.645	33.1	H&E → RNA-seq → Olink → WES

while Gemma's systems share the same most common trajectory but differ substantially in Macro-F1.

F Additional Contrast-Persistence Diagnostics

Replay coverage and finite-panel construction. The current common-input, common-prior analysis contains all 787 ADNI subjects and all 172 IBDome subjects. Finite-panel effects are computed by first averaging the tested backbone effects within subject and then applying a 10,000-replicate subject bootstrap. Thus every backbone receives equal weight within each subject and confidence intervals preserve the paired subject structure.

Backbone heterogeneity. On ADNI, Qwen and Gemma are classified as inconclusive, whereas Ministral is materially nonpersistent. On IBDome, Qwen, Gemma, and Ministral are materially nonpersistent and GPT is inconclusive. The finite-panel conclusions therefore summarize a declared finite model panel rather than asserting identical behavior across backbones.

Duplicate component control. All 787 ADNI subjects contribute to the three-backbone duplicate ablation. Finite-panel median TV is 0.097 without the preservation guard and 0.000 with the full Ledger. The mean no-guard-minus-full reduction is 0.122 (0.116–0.128). Guard-only processing is equivalent to full Ledger processing within the ± 0.003 margin.

G Representation Controls

For completeness, we retain the earlier representation-control analyses as historical diagnostics; they are not part of the revised confirmatory claims. The evaluated pipelines consume documented encoded evidence rather than uniformly processing every modality in native form. Historical production-versus-control log-score contrasts are:

- ADNI MRI versus subject-shuffled representation: 0.006 (95% CI -0.004–0.017);
- IBDome H&E versus shuffled: 0.084 (0.044–0.126);
- IBDome Olink versus within-feature permutation: 0.031 (0.009–0.054);
- IBDome RNA-seq versus permutation: 0.019 (-0.002–0.041);
- IBDome WES versus permutation: 0.004 (-0.014–0.022).

An information-matched H&E text summary recovers 71% of the production effect. These controls support contributions from encoded H&E and Olink evidence but not a reliable contribution from the studied ADNI MRI, pooled RNA-seq, or WES representations. Modality labels alone are therefore insufficient evidence that an agent receives usable modality-specific information.

H ADNI COHORT AND INPUT INVENTORY

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal, multicenter observational study of cognitive aging and Alzheimer's disease. The aligned ADNI inventory contains 795 subjects, of whom 787 form the processed common cohort used in the main analyses. The remaining eight subjects were excluded because their structural-MRI series could not be processed by the frozen DICOM pipeline: no readable single-frame DICOM slices were available under the analysis-time loading procedure. Inputs include clinical/demographic variables (age, sex, and education), cognitive measures (MMSE, ADAS11, and ADAS13), APOE4 allele count, and T1-weighted structural MRI. The prediction task uses the ADNI diagnostic categories cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). The processed 787-subject cohort contains 205 CN, 437 MCI, and 145 AD subjects. Table 11 summarizes the aligned input inventory. ADNI is a controlled-access resource; source clinical, genomic, and MRI/DICOM data are not redistributed with our artifacts, and fresh end-to-end inference requires independently authorized access under the applicable ADNI data-use requirements.

Table 11: ADNI inputs in the aligned 795-subject inventory. The main analyses use the processed 787-subject common cohort. APOE4 is represented as the number of $\varepsilon 4$ alleles carried by each subject (0, 1, or 2).

Modality	Feature	Type	Unique/features	Subjects
Clinical/Demographic	Age	Scalar	37	795
Clinical/Demographic	Sex	Categorical	2	795
Clinical/Demographic	Education	Scalar	16	795
Cognitive	MMSE	Scalar	12	795
Cognitive	ADAS11	Scalar	95	795
Cognitive	ADAS13	Scalar	130	795
Genomic	APOE4 allele count	Ordinal	3	795
Structural MRI	T1-weighted structural MRI	Image	–	795

I IBDOME COHORT AND INPUT INVENTORY

The frozen manifest resolves the reduction from 189 aligned subjects to the 172-subject complete-evidence cohort: 9 lack a target valid at the frozen prediction time, 5 violate the modality-timing window, and 3 fail assay or image quality control. The prediction target is three-way disease status, with 108 Crohn disease, 48 ulcerative colitis, and 16 non-IBD subjects in the frozen cohort. All exclusions are fixed before model execution. The four controlled counterfactual interventions—relevant addition, countervailing replacement, exact duplication, and important removal—all operate on the Olink proteomics modality. Table 12 summarizes the aligned IBDome input inventory.

Table 12: IBDome variables in the aligned 189-subject source inventory.

Modality	Feature	Type	Unique/features	Subjects
Clinical/Demographic	Sex	Categorical	2	189
Clinical/Demographic	Birth year	Scalar	54	189
Clinical	First diagnosis	Categorical	105	149
Clinical	Disease localization (CD)	Categorical	3	111
Clinical	Disease localization (UC)	Categorical	3	42
Clinical	Disease course	Categorical	3	101
Transcriptomic	RNA-seq expression	Continuous	–	189
Proteomic	Olink expression	Continuous	–	189
Genomic	Whole-exome variants	Variant	–	189
Histopathology	H&E whole-slide imaging	Image	–	189

Disease localization uses Montreal categories L1–L3 for CD and E1–E3 for UC; disease course uses B1–B3. The frozen leakage-safe specification excludes *First diagnosis* and other target-derived or temporally unsafe fields. The artifact additionally provides the included-subject manifest, cohort-level exclusion accounting, leakage-safe specification, and subject-level intervention mappings.

J EXPLORATORY IBDOME MICROBIOME SUBSET

The exact same-date microbiome subset contains only 19 subjects and is excluded from the main cross-domain claims. Table 13 reports fixed-order performance on this exploratory subset.

Table 13: Fixed-order performance on the 19-subject exploratory microbiome subset.

Available evidence	N	Accuracy (%)	Macro-F1
Clinical/Demographic	19	47.37	0.2970
+ WES	19	47.37	0.2970
+ RNA-seq	19	47.37	0.2970
+ Olink	19	78.95	0.4794
+ H&E	19	73.68	0.4222
+ 16S Microbiome	19	78.95	0.4677

The Tool Agent with 16S evidence obtains 68.42% accuracy, 0.3870 Macro-F1, 5.000 calls, and 0.0% STOP. Its most common trajectory is H&E $\rightarrow$ RNA-seq $\rightarrow$ Olink $\rightarrow$ 16S $\rightarrow$ WES.

K ARTIFACT AND REPRODUCIBILITY BOUNDARY

The frozen release contains:

- canonical prompts, structured-output schemas, available model revision identifiers, access metadata, decoding parameters, and serving settings;
- representation hashes, preprocessing code, subject-level intervention manifests, and protected-data access requirements;
- parser validation, retry, normalization, probability-flooring, and invalid-output policies;
- Ledger audit fields, guard parameters, preserved guard-ablation outputs, and available pre-guard and post-guard distributions;
- canonical autonomous outputs, common-prior replay inputs, exact added-evidence payloads, subject identifiers, and replay provenance;
- documented validity, retry/repair, and missing-output policies, with preserved run-level provenance where available;
- subject-level bootstrap, equivalence testing, diagnostic sensitivity analyses, and table-generation code;

The reported tables are regenerated directly from the structured subject-level outputs using the released analysis code. Some individual outputs differ across repeated model executions despite fixed decoding. Same-condition repeats are therefore retained as diagnostics of processing variability, but are not used as a subtractive correction for the headline effects. Historical provider-side states that cannot be reconstructed are labeled supplementary and do not carry the revised headline claims.

Fresh end-to-end reproduction requires authorized access to ADNI and IBDome source data and the corresponding model/provider runtime. The artifact guarantees the documented analysis path, not indefinite availability of external providers or clinical validity of the resulting predictions.