THESISAGENT: REVISING PERSISTENT THESES FOR LONG-HORIZON DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon agents must revise decision-supporting assumptions while preserving unaffected beliefs. This requires controlling revision scope and separating belief changes from execution constraints. Behavioral responsiveness must also be distinguished from decision quality. We propose ThesisAgent, which maintains these assumptions as a persistent decision thesis governed by an explicit transition contract. To address Unscoped Thesis Revision, Evidence-Local Revision separates hypothesis selection from revision strength. Its numerical updater changes only a bounded subset and exactly preserves unselected confidences. To address Belief–Execution Entanglement, Decision–Execution Separation exposes desired actions before feasibility enforcement so that belief effects can be audited at fixed execution context. To address Sensitivity–Utility Ambiguity, Intervention-Based Utility Audit combines paired hypothesis interventions with oracle-referenced regret in the controlled setting to assess decision dependence and quality separately. ThesisWorld experiments show that local updates improve utility over matched dense updates and reveal a trade-off between rapid response and retained correction. In ALFWorld, persistent thesis state improves task success by 6.0 percentage points over a matched control using the same execution controller. In the 43-month return-based financial proxy, ThesisAgent achieves a cumulative net return 23.7 percentage points above the highest return among the evaluated baselines. Our code is available at `https://anonymous.4open.science/r/THESISAGENT-948D`.

1 INTRODUCTION

Long-horizon agents must maintain the assumptions behind ongoing decisions while responding to new evidence. An investment may remain justified until a disclosure contradicts one of its supporting assumptions. Such evidence need not invalidate the remaining assumptions or require an immediate change in execution. Reliable adaptation therefore requires preserving what remains supported while revising what has changed. We call the set of falsifiable assumptions supporting an ongoing decision a *decision thesis*. Figure 1 illustrates how maintaining this thesis differs from regenerating a rationale at each step.

Existing agent frameworks provide important foundations for this problem. ReAct interleaves reasoning with actions (Yao et al., 2022), while Reflexion retains verbal feedback to guide subsequent attempts (Shinn et al., 2023). More directly, ABBEL conditions decisions on recursively updated beliefs (Lidayan et al., 2025). Agent-BRACE represents uncertain state through atomic claims (Singh et al., 2026), and HEP makes hypothesis evolution explicit and auditable (Takahara & Mizoguchi, 2026). These advances motivate studying the transition rules governing persistent state. Our focus is how evidence should revise an existing thesis and how the consequences of that revision should be assessed. This raises three challenges.

Unscoped Thesis Revision. Recursive belief updates can omit or incorrectly revise information, as observed in ABBEL (Lidayan et al., 2025). A-MEM revises existing memories as new information arrives (Xu et al., 2026), while Memory-R1 learns explicit memory-editing operations (Yan et al., 2026). For a persistent thesis, an observation may concern only a small subset of hypotheses. Evidence relevance must therefore be distinguished from revision strength. Reducing update magnitudes alone does not determine which hypotheses the evidence is permitted to change. Repeated weak revisions

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.

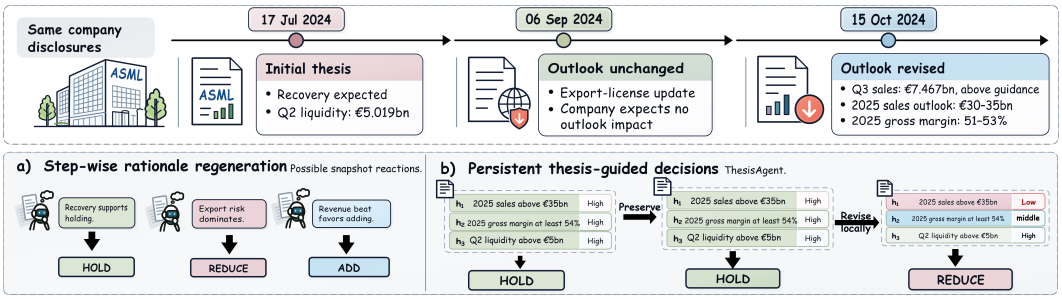

Figure 1: Investment decisions under shared ASML disclosures. Panel a illustrates snapshot reactions from step-wise rationale regeneration. Panel b illustrates selective revision of persistent hypotheses in ThesisAgent. The depicted states and actions are illustrative.

outside the intended scope can accumulate, making it difficult to preserve unaffected beliefs while accepting necessary corrections.

Belief–Execution Entanglement. Executed actions reflect both intended decisions and execution constraints. Recent frameworks enforce runtime rules on LLM agents or check action trajectories against explicit safety policies (Wang et al., 2026; Chen et al., 2025b; Dalal et al., 2018). An observed action change may therefore arise from revised beliefs or altered feasibility. Executed trajectories alone cannot distinguish these causes, making it difficult to determine whether the thesis actually influenced the decision. This motivates exposing desired actions before constraint enforcement and auditing thesis dependence at fixed execution context.

Sensitivity–Utility Ambiguity. Verbal explanations can omit factors that influence model outputs, limiting their value as evidence of decision dependence (Turpin et al., 2023; Chen et al., 2025a). Chain-of-thought monitoring can also miss undesired behavior (Korbak et al., 2025). Hypothesis interventions provide a more direct test of decision dependence, but responsiveness alone does not establish correctness. A sensitive policy may move toward a worse decision, while a stable policy may preserve a sound decision or ignore necessary evidence. Evaluating useful revision therefore requires a quality reference beyond the magnitude of behavioral change.

We introduce **ThesisAgent**, a framework that specifies an explicit transition contract for persistent decision theses. To address *Unscoped Thesis Revision*, its *Evidence-Local Revision* mechanism selects a bounded set of relevant hypotheses before updating their confidence. The numerical updater exactly preserves unselected confidences, with the quality of the selected scope depending on the evidence interpreter. To address *Belief–Execution Entanglement*, its *Decision–Execution Separation* interface exposes the desired action before feasibility enforcement. This permits auditing thesis dependence separately from execution adjustments while holding execution context fixed. To address *Sensitivity–Utility Ambiguity*, its *Intervention-Based Utility Audit* evaluates paired hypothesis interventions alongside oracle-referenced regret in the controlled setting. Task utility provides a separate outcome measure. Figure 2 summarizes the framework.

ThesisWorld instantiates this contract with a deterministic updater and a known numerical action map. Its consistency rule filters proposal residuals around the belief-based reference. We analyze when reference error removes the denoising benefit. This setting tests the transition mechanism under controlled assumptions. ALFWorld provides an interactive adaptation with a task-specific controller, while the financial proxy uses return-derived signals. The implementations differ across these settings, so their results support distinct claims.

Our contributions are as follows.

- We introduce ThesisAgent and its explicit transition contract for persistent decision theses. The framework connects scoped belief revision to a decision interface with a separate behavioral audit.
- We specify a bounded numerical update that exactly preserves confidences outside the selected scope. We also characterize the consistency rule as ordinary residual filtering and

derive a reference-error condition for reducing action-readout mean squared error under stated assumptions.

- We evaluate the contract on 800 paired ThesisWorld episodes, obtaining normalized return 0.841 and counterfactual decision regret 0.1536. Matched controls expose a trade-off between rapid response and retained correction. In ALFWorld, persistent thesis state raises success from 124 to 132 of 134 tasks with the same PhaseGuard controller. A separate financial proxy examines persistent signal accumulation over 43 monthly returns.

2 Related Work

Persistent memory and belief revision. Generative Agents retrieve and reflect on past experiences, while MemGPT manages information across memory tiers (Park et al., 2023; Packer et al., 2023). Recent systems make memory organization more adaptive. A-MEM evolves linked notes, and Zep represents changing information through a temporal knowledge graph (Xu et al., 2026; Rasmussen et al., 2025). Mem0 consolidates conversational information, while MemoryOS coordinates updates across hierarchical stores (Chhikara et al., 2025; Kang et al., 2025). Memory-R1 and Mem-α further learn memory management through reinforcement learning (Yan et al., 2026; Wang et al., 2025). These approaches address what information to retain and how to organize it. Explicit belief models connect persistent state more directly to decisions. ABBEL uses a recurrent belief bottleneck, and Agent-BRACE conditions actions on uncertain atomic claims (Lidayan et al., 2025; Singh et al., 2026). HEP also records hypothesis evolution and verification (Takahara & Mizoguchi, 2026). Building on these precedents, ThesisAgent specifies which existing hypotheses an observation is permitted to revise before determining revision strength. Its numerical updater preserves unselected confidences exactly. This transition contract addresses Unscoped Thesis Revision without assuming that memory retrieval alone determines a valid update.

Execution interfaces and behavioral audit. Action-level safeguards separate proposed behavior from permitted execution. Safety layers correct continuous actions to satisfy constraints (Dalal et al., 2018). AgentSpec enforces explicit runtime rules, while ShieldAgent checks agent behavior against safety policies (Wang et al., 2026; Chen et al., 2025b). ThesisAgent uses this separation to address Belief–Execution Entanglement by auditing desired decisions at fixed execution context. A separate issue is whether an agent's stated reasoning explains its behavior. Studies of language models and reasoning models find that verbal explanations can omit influences on their outputs (Turpin et al., 2023; Chen et al., 2025a). Work on chain-of-thought monitorability likewise emphasizes both its potential and its limitations (Korbak et al., 2025). LongMemEval tests updated knowledge in sustained interactions, whereas MemoryArena evaluates how experience retained across sessions supports later actions (Wu et al., 2024; He et al., 2026). These evaluation goals motivate examining both persistent state and behavior. To address Sensitivity–Utility Ambiguity, our paired interventions measure the desired action's dependence on thesis state. When an appropriate quality reference is available, the audit separately assesses whether the induced action change improves the decision.

Financial agents and evaluation. FinMem uses layered financial memory, while FinCon adapts investment beliefs through conceptual verbal reinforcement (Yu et al., 2025; 2024). FinRobot supports financial analysis and equity research (Zhou et al., 2024). TradingAgents coordinates specialist agents, and Trading-R1 trains structured investment reasoning with reinforcement learning (Xiao et al., 2024; 2025). FinPos already separates directional reasoning from risk-aware position adjustment, extending agent decisions to continuous position management (Liu & Dang, 2025). Evaluation also extends beyond trading: FinSearchComp examines financial search and reasoning, while FinGAIA tests multi-step financial tasks requiring tool use (Hu et al., 2026; Zeng et al., 2025). FinWorld provides a shared platform for financial modeling and deployment (Zhang et al., 2026). Our financial experiment examines how persistent signal accumulation maps to portfolio decisions under shared return-derived inputs. Its mechanism-inspired adapters support controlled comparisons under a common proxy protocol. Their results do not establish a ranking of the native systems or evaluate their financial text reasoning.

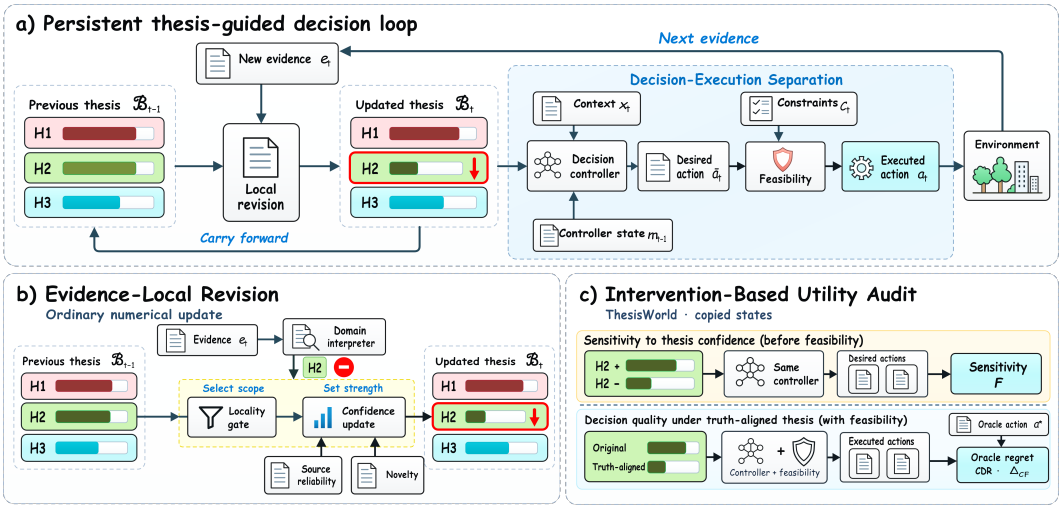

Figure 2: Overview of ThesisAgent. (a) Persistent thesis-guided decision loop with separate decision and execution stages. (b) Evidence-local numerical revision separates update scope from revision strength. (c) Paired thesis interventions assess policy sensitivity, with decision quality evaluated separately when a suitable reference is available. Hypotheses and confidences are illustrative.

3 Method

ThesisAgent organizes long-horizon decisions around a persistent thesis. Its transition contract specifies which assumptions evidence may revise and separates the resulting desired decision from its execution. A paired intervention audit assesses decision dependence and response quality separately. We define these interfaces below and give a concrete confidence update. Domain adapters supply evidence interpretation and action semantics. Figure 2 summarizes the framework. Table 5 specifies how the evaluated adapters realize its components.

3.1 Problem Setup and Persistent State

At step t, the agent receives evidence e_t and retains a bounded thesis $B_t = \{h_{t,k} : k \in \mathcal{I}_t\}$, with $|\mathcal{I}_t| \leq K_{\max}$. Each hypothesis has a stable identifier and proposition q_k. Confidence $c_{t,k}$ records its current support. A verification rule specifies what observation can resolve it within its horizon. Lifecycle status distinguishes active from terminal hypotheses, and evidence records retain provenance. Separating identity from confidence lets the agent revise an assumption's support without silently changing its meaning.

The evidence interpreter proposes hypothesis-specific records, and the updater commits permitted changes:

$$z_t = \mathcal{M}_{\text{dom}}(e_t, B_{t-1}), \qquad B_t = U_\theta(B_{t-1}, z_t). \tag{1}$$

For textual observations, interpretation can use a language model. Structured inputs can use a numerical map. These interfaces describe functional roles and need not require separate model calls. Execution context x_t records current operational conditions, with feasible actions in $\mathcal{C}_t$. Evidence about a route closure can revise a route-open hypothesis while leaving an energy assessment unchanged. Resource constraints can separately alter execution. Appendix E specifies the state schema and lifecycle.

3.2 Evidence-Local Revision

To address Unscoped Thesis Revision, we separate update eligibility from revision strength. Soft weighting alone can accumulate small changes on unrelated hypotheses. The gate first identifies the editable set $\mathcal{K}_t$. The updater then revises only that set during an ordinary evidence update.

For confidence-valued hypotheses, the interpreter supplies relevance $r_{t,k} \in [0,1]$ and a signed revision $\Delta_{t,k} \in [-\beta, \beta]$ for each active hypothesis. The gate selects at most M hypotheses with the largest relevance above γ_r, using deterministic tie-breaking. Source reliability $\rho_t \in [0,1]$ scales admitted evidence, while novelty $n_t \in [0,1]$ discounts repetition. Our numerical updater accumulates the admitted revision in log-odds space:

$$c_{t,k} = \begin{cases} \mathrm{clip}_{[\epsilon,1-\epsilon]}[\sigma(\mathrm{logit}(c_{t-1,k}) + r_{t,k}\rho_t n_t \Delta_{t,k})], & k \in \mathcal{K}_t, \\ c_{t-1,k}, & k \notin \mathcal{K}_t. \end{cases} \tag{2}$$

Here σ is the logistic function, and $\epsilon \in (0, 1/2)$ also bounds initialization. This rule retains prior support and assigns opposite effects to supporting and contradicting evidence. Reliability and novelty control the strength of an eligible revision without enlarging its scope.

Every unselected confidence is preserved exactly by this numerical update. A mistaken gate can still admit an unrelated hypothesis, so the guarantee concerns the selected scope rather than semantic correctness. Prompted textual revision expresses the locality objective without inheriting this hard guarantee. Verification and proposition replacement are explicit lifecycle operations outside ordinary revision. Appendix A specifies their implementations. Appendix F bounds drift from false-positive gate selections.

3.3 Decision–Execution Separation

To address Belief–Execution Entanglement, we expose the desired action before execution handling. An auxiliary state m_{t-1} records the adapter's declared controller history or execution ledger. We write

$$\begin{aligned} \tilde{a}_t &= \pi_\phi(B_t, x_t, m_{t-1}; \xi_t), \\ a_t &= \mathcal{E}_{\mathrm{dom}}(\tilde{a}_t, x_t, m_{t-1}) \in \mathcal{C}_t. \end{aligned} \tag{3}$$

Here ξ_t denotes controller randomness when present. Parameter subscripts describe implementation choices and do not prescribe gradient training. The execution adapter may project a continuous proposal or apply a task controller to a discrete action. It declares the auxiliary state retained after execution for the next step.

This interface exposes where an intended decision changes during execution. Auditing at fixed context and auxiliary state can then isolate the effect of a thesis intervention. Declaring m_{t-1} also avoids assuming that the current thesis alone determines every action. A consistency controller can be implemented within this interface. Appendix E.4 gives a scalar realization that filters proposal residuals around a specified thesis-supported transition. Its reference assumptions and denoising conditions are analyzed in Appendix F.

3.4 Intervention-Based Utility Audit

To address Sensitivity–Utility Ambiguity, we evaluate decision dependence separately from the quality of a correction. Construct paired copies $B_t^{k,+}$ and $B_t^{k,-}$ that perturb one hypothesis while preserving the others. For a numerical confidence and step $\eta > 0$, the copies use $\min(1, c_{t,k} + \eta)$ and $\max(0, c_{t,k} - \eta)$, respectively. Both branches fix x_t and $\mathcal{C}_t$ and reuse the same pre-intervention state m_{t-1} and randomness ξ_t. The updater is not rerun. Desired-action sensitivity is

$$\begin{aligned} \tilde{a}_t^{k,\pm} &= \pi_\phi(B_t^{k,\pm}, x_t, m_{t-1}; \xi_t), \\ F_{t,k} &= D_a(\tilde{a}_t^{k,+}, \tilde{a}_t^{k,-}), \end{aligned} \tag{4}$$

where D_a is a distance appropriate to the desired-action space. A nonzero difference establishes local dependence, not improvement.

Quality assessment requires an independently justified reference. Let B_t^{cf} contain a reference-supported correction to hypothesis k, and let $Q_t(a)$ be a task-defined quality score, with larger values preferred. Apply the same decision and execution interfaces to obtain a_t^{cf}, keeping the paired conditions fixed. The quality change is

$$\Delta_Q = \mathbb{E}_{t,k}\left[Q_t(a_t^{\mathrm{cf}}) - Q_t(a_t)\right]. \tag{5}$$

The reference and scoring rule must be held fixed across branches. When ground truth is available, a truth-aligned correction and negative oracle error instantiate this interface. Without a valid local

Table 1: Numerical adapters on 800 paired ThesisWorld episodes with shared action noise. Values are mean ± SEM over ten 80-episode groups. Lower is better for Brier, UBC, and CDR; higher otherwise.

Method	Brier ↓	Attr. F1 ↑	UBC ↓	RRR ↑	CDR ↓	ThesisWorld Return ↑
Full-History	0.137 ± 0.002	0.799 ± 0.002	0.01260 ± 0.00005	**0.807 ± 0.012**	0.1842 ± 0.0029	0.811 ± 0.003
Rolling-Summary	0.174 ± 0.001	0.726 ± 0.002	0.02196 ± 0.00006	0.699 ± 0.014	0.1869 ± 0.0015	0.806 ± 0.002
ReAct	0.248 ± 0.000	0.657 ± 0.001	0.14051 ± 0.00034	0.704 ± 0.017	0.2476 ± 0.0027	0.739 ± 0.002
Reflexion	0.164 ± 0.001	0.726 ± 0.002	0.02094 ± 0.00006	0.693 ± 0.013	0.1823 ± 0.0015	0.811 ± 0.002
ABBEL-style bottleneck	0.147 ± 0.002	0.799 ± 0.002	0.02298 ± 0.00009	0.706 ± 0.016	0.1828 ± 0.0029	0.813 ± 0.003
Agent-BRACE-style atomic	0.115 ± 0.001	0.922 ± 0.001	0.00351 ± 0.00002	0.806 ± 0.006	0.1661 ± 0.0026	0.830 ± 0.003
ThesisAgent	**0.104 ± 0.001**	**0.923 ± 0.002**	**0.00060 ± 0.00001**	0.806 ± 0.012	**0.1536 ± 0.0018**	**0.841 ± 0.002**

quality reference, task outcomes are reported separately and sensitivity is not interpreted as a quality gain. Appendix B defines the reported numerical specialization. All interventions operate on copies and leave the live decision trajectory unchanged.

4 Experiments

We evaluate state quality, action dependence, and task utility. ThesisWorld supplies controlled tests of the numerical core; ALFWorld and finance test domain adaptations with explicitly different implementations. External-agent comparisons use mechanism-inspired adapters rather than native systems. Appendix B defines the metrics, and Appendix C specifies the data, configurations, and audit history.

4.1 Evaluation design

We ask whether local revisions preserve accurate beliefs (RQ1), whether desired actions respond usefully to thesis changes (RQ2), and how the interface behaves in domain adaptations (RQ3).

ThesisWorld. The primary experiment uses 800 paired episodes of 48 steps with six latent propositions, noisy evidence embeddings, duplicate observations, and delayed factual reveals. We compare against Full-History, Rolling-Summary, ReAct (Yao et al., 2022), Reflexion (Shinn et al., 2023), an ABBEL-style bottleneck (Lidayan et al., 2025), and an Agent-BRACE-style atomic baseline (Singh et al., 2026). All six baselines are numerical mechanism adapters. All methods share exogenous noise draws, action-noise scale 0.045, and a common thesis-based action readout, with no LLM inference or gradient training. Separate component controls change only the named mechanism. We report ten-group SEM and episode-paired tests. Holm correction covers 30 primary comparisons, with seven post-hoc ablation contrasts forming a separate family. Sensitivity sweeps use disjoint validation seeds. Appendices B and C.1 define the metrics, adapters, and test families.

ALFWorld. We conduct a post-hoc evaluation on all 134 solvable official `valid_unseen` tasks. All methods use `gpt-4.1-mini` at temperature zero, a common admissible-action catalog, and at most 50 steps. Baselines include Full-History, Rolling-Summary, ReAct (Yao et al., 2022), Reflexion (Shinn et al., 2023), an ABBEL-style bottleneck (Lidayan et al., 2025), and an Agent-BRACE-style atomic baseline (Singh et al., 2026). These baselines are implemented as prompt adaptations. Matched ablations compare ThesisAgent with a no-persistent-thesis variant, both with and without PhaseGuard. PhaseGuard is a stateful task-specific controller that can replace valid proposals. Its ledger remains present in the no-persistent-thesis + PhaseGuard control.

Finance. We report 43 monthly returns, January 2023–July 2026, for 49 US industry portfolios. Baselines include Equal Weight (DeMiguel et al., 2009), Momentum (Efficiency, 1993), XGBoost (Chen & Guestrin, 2016), and Rolling Summary. We also compare numerical adaptations of FinMem (Yu et al., 2025), FinCon (Yu et al., 2024), TradingAgents (Xiao et al., 2024), Trading-R1 (Xiao et al., 2025), and FinPos (Liu & Dang, 2025). Momentum, lagged-return surprise, and distance from a trailing price peak supply three thesis slots. These are return-derived proxies, not company fundamentals or analyst expectations. All strategies use lagged inputs, monthly rebalancing, and 10 bp costs per unit of return-drifted half-ℓ_1 turnover. ThesisAgent's top-16% portfolio is additionally evaluated at the common top-20% projection. Matched monthly-reset and six-month EMA controls assess the effect

Table 2: Post-hoc ALFWorld evaluation on all 134 solvable valid-unseen tasks. ThesisAgent + PhaseGuard is the full system; no-persistent-thesis + PhaseGuard is the matched ablation. Success is primary. Steps and Steps (succ.) report mean environment steps over all tasks and successful tasks, respectively. Calls and tokens are totals over all tasks. Best: dark gray and bold; second-best: light gray and underlined; ties share formatting.

Method	Success ↑	Steps ↓	Steps (succ.) ↓	Invalid ↓	Calls ↓	Tokens ↓
Full-History	20/134 (14.9%)	45.37	18.95	0.31%	6079	6.878M
Rolling-Summary	25/134 (18.7%)	43.88	17.20	0.03%	5880	4.455M
ReAct	17/134 (12.7%)	45.53	**14.76**	0.33%	6101	4.649M
Reflexion	17/134 (12.7%)	46.01	18.59	1.33%	6166	4.769M
ABBEL-style bottleneck	20/134 (14.9%)	45.25	18.15	0.69%	6063	4.614M
Agent-BRACE-style atomic	18/134 (13.4%)	45.40	15.72	1.12%	6083	5.058M
No-persistent-thesis (revisable)	103/134 (76.9%)	25.46	18.08	**0.00%**	4003	3.742M
ThesisAgent (revisable)	119/134 (88.8%)	20.97	17.31	**0.00%**	3338	4.286M
No-persistent-thesis + PhaseGuard	124/134 (92.5%)	19.31	16.84	**0.00%**	2581	**2.423M**
ThesisAgent + PhaseGuard	**132/134 (98.5%)**	**17.03**	16.53	**0.00%**	**2282**	3.031M

Table 3: Recent-period proxy results, January 31, 2023–July 31, 2026 (43 months). Methods share lagged inputs, independently initialized portfolios, and a cost of 10 bp per unit of half-ℓ_1 turnover. Net return is cumulative; excess is the difference from Equal Weight in percentage points (pp). Best: dark gray and bold; second-best: light gray and underlined; ties share formatting.

Method	RankIC ↑	Cumulative Net Return ↑	Excess (pp) ↑	Sharpe ↑	Max Drawdown ↓	Turnover ↓
Equal Weight	–	75.7%	0.0	1.18	11.0%	**3.0%**
Momentum	0.007	52.7%	-23.0	0.75	12.3%	27.6%
XGBoost	0.010	87.4%	11.7	0.97	16.4%	35.6%
Rolling Summary	0.022	76.7%	1.0	0.96	13.6%	34.9%
FinMem	0.003	81.8%	6.1	1.14	10.7%	56.2%
FinCon	-0.019	63.5%	-12.2	0.96	12.5%	62.4%
TradingAgents	-0.013	86.5%	10.8	1.19	12.3%	58.9%
Trading-R1	-0.058	35.0%	-40.7	0.73	**9.8%**	37.6%
FinPos	-0.024	58.3%	-17.4	0.94	11.6%	24.9%
ThesisAgent	**0.039**	**111.1%**	**35.4**	**1.33**	12.6%	6.6%

of persistent signal accumulation (Appendix H). This single market path supports exploratory point estimates, without cross-seed significance claims. Appendix C.2 specifies chronological fitting and verification, while Appendix D provides the extended setup and leakage controls.

4.2 Main Comparison

On ThesisWorld, ThesisAgent achieves the lowest Brier score (0.104), UBC (0.00060), and CDR (0.1536), with the highest return (0.841) (Table 1). These gains align with local revision protecting unaffected beliefs and persistent state retaining corrections, while consistency limits unsupported action changes. RRR remains close to the strongest baselines, emphasizing gains in retained belief quality and decision utility.

In the post-hoc ALFWorld evaluation, thesis persistence raises success from 103/134 to 119/134 without PhaseGuard and from 124/134 to 132/134 with it (Table 2). These gains support complementary roles for persistent beliefs and execution control. The guarded comparison retains the same PhaseGuard ledger, showing a 6.0 pp gain from thesis persistence at higher token cost.

Across the 43-month financial proxy, ThesisAgent exceeds XGBoost by 23.7 pp in cumulative net return and achieves the highest Sharpe ratio (1.33) and RankIC (0.039) among the evaluated adapters (Table 3). Figure 4 complements these aggregate results with cumulative wealth trajectories and performance under varying transaction costs. Under the common top-20% projection, it retains the leading return and Sharpe, at 107.6% and 1.36, respectively (Table 7).

Matched financial controls further examine why persistence helps. With identical evidence, feature weights, costs, and portfolio projection, persistent accumulation yields 107.6% return, compared with 61.8% for monthly reset and 82.4% for six-month EMA, while reducing turnover relative to both controls (Appendix H). The reset comparison supports retaining signals across months, while

Table 4: Ablation and mechanism diagnostics under a shared evaluation protocol. Panel A removes or replaces one internal ThesisAgent component at a time. Panel B is a matched alternative-mechanism control rather than an internal component ablation. Full ThesisAgent is the shared reference for both panels. Delay is the capped score defined in Appendix B. Best: dark gray and bold; second-best: light gray and underlined; ties share formatting.

Variant	Brier ↓	UBC ↓	Sust. Rev. Delay ↓	UTR ↓	RRR ↑	CDR ↓	Task Utility ↑
Panel A: Component ablations							
w/o Persistent Thesis State	0.206 ± 0.002	0.01857 ± 0.00009	21.00 ± 0.00	0.002 ± 0.000	**0.886 ± 0.010**	0.2585 ± 0.0049	0.726 ± 0.006
w/o Locality Gate (Dense Revision)	0.118 ± 0.001	0.01025 ± 0.00005	1.55 ± 0.05	0.004 ± 0.001	0.722 ± 0.012	0.1620 ± 0.0019	0.835 ± 0.002
w/o Consistency Regularization	**0.104 ± 0.001**	**0.00060 ± 0.00001**	**1.30 ± 0.02**	0.020 ± 0.001	0.824 ± 0.010	0.1655 ± 0.0021	0.833 ± 0.002
w/o Factual Verification	0.112 ± 0.001	0.00145 ± 0.00003	2.10 ± 0.06	0.003 ± 0.000	0.800 ± 0.013	0.1605 ± 0.0018	0.836 ± 0.002
Panel B: Matched mechanism control							
Action Smoothing	**0.104 ± 0.001**	**0.00060 ± 0.00001**	**1.30 ± 0.02**	**0.000 ± 0.000**	0.863 ± 0.009	0.1662 ± 0.0022	0.832 ± 0.002
Shared reference							
Full ThesisAgent	**0.104 ± 0.001**	**0.00060 ± 0.00001**	**1.30 ± 0.02**	0.002 ± 0.000	0.806 ± 0.012	**0.1536 ± 0.0018**	**0.841 ± 0.002**

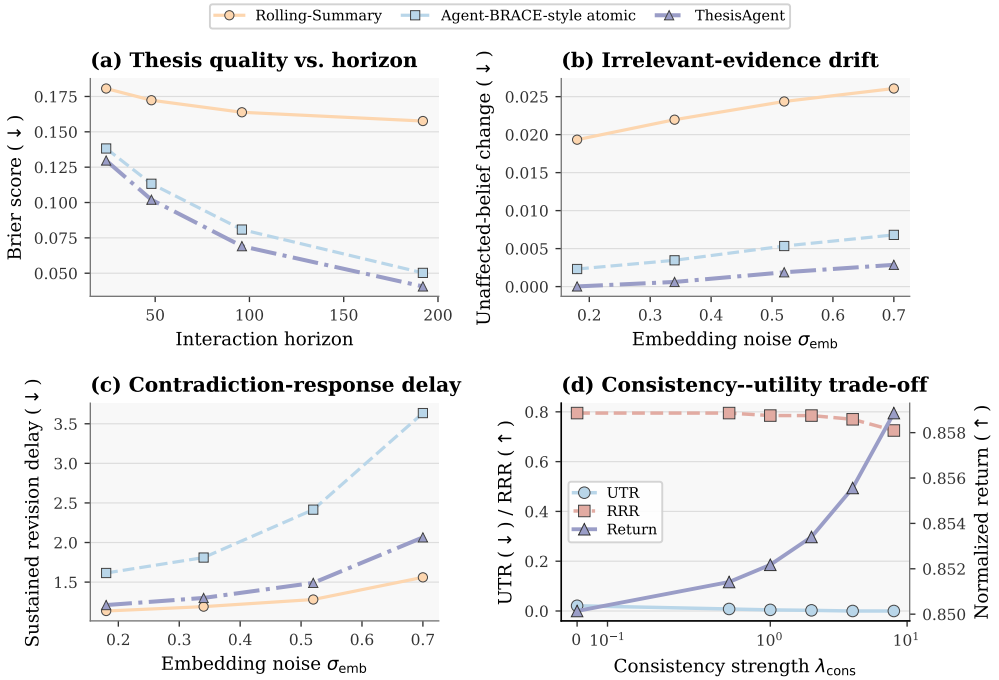

Figure 3: Mechanism analysis for RQ1–RQ2: long-horizon thesis quality, robustness to irrelevant evidence, response to reliable contradiction, and the trade-off between thesis–action consistency and task utility.

the EMA comparison shows an advantage over the tested fixed smoother. An ordinary numerical accumulator reproduces the portfolio path exactly, identifying retained signal accumulation as the operative numerical mechanism. These findings concern the evaluated adaptations within this market window.

4.3 Ablation and Mechanism Analysis

Table 4 evaluates component contributions and matched action smoothing under shared noise draws, noise scale, and action map. Dense revision changes only update eligibility. Appendix A specifies the switches.

Component contributions. Resetting thesis state lowers return from 0.841 to 0.726 and fails sustained retention, showing why corrections must persist across observations. Dense revision raises UBC from 0.00060 to 0.01025, supporting the locality gate's role in protecting unaffected beliefs. Removing verification increases Brier from 0.104 to 0.112, demonstrating the value of delayed factual

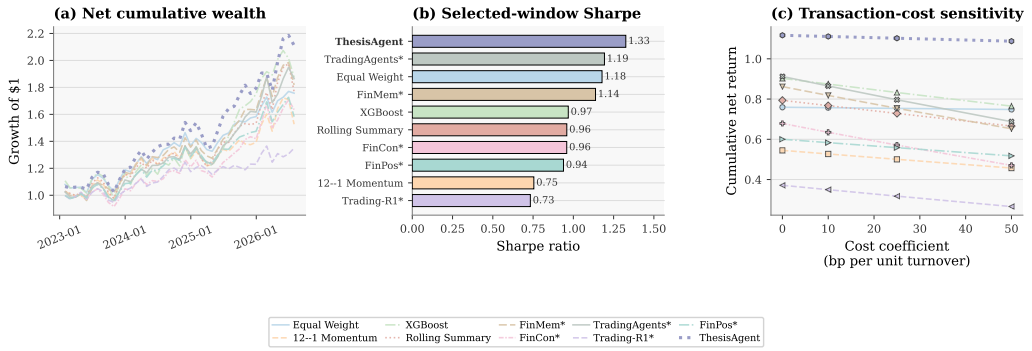

Figure 4: Exploratory January 2023–July 2026 financial reporting window: transaction-cost-adjusted cumulative wealth, method-level Sharpe ratios, and sensitivity to execution costs. Portfolios are independently initialized at the start of the window; asterisks denote external-agent mechanisms.

feedback. Removing consistency preserves belief metrics but raises CDR from 0.1536 to 0.1655, locating its effect at the action layer.

Counterfactual audit. At fixed context and action noise, truth-aligned interventions reduce sampled regret by 0.01004 (Table 6). The evidence-sign audit finds sensitivity to relevant evidence but zero sensitivity to irrelevant evidence in the audit sample, compared with 0.04220 for dense revision. These results support numerical locality and useful state dependence, without establishing semantic faithfulness of natural-language reasoning.

Consistency–utility trade-off. Action smoothing and state reset attain higher RRR but lower return, so responsiveness must be assessed alongside sustained retention and decision quality. On disjoint seeds, larger λ_{cons} reduces unsupported transitions and improves return over the tested range, but the largest value lowers RRR (Figure 3). The primary setting remains $\lambda = 2$.

Numerical mechanism and reference accuracy. Across 17 prespecified conditions and 800 additional paired episodes on disjoint seeds, an independent array filter reproduces all confidences and actions, while residual filtering matches the consistency controller to numerical precision (Appendix G). Under matched proposal noise, sparse revision outperforms dense revision in return, and exact-reference consistency reduces residual MSE approximately fivefold over steps 12–47. Its return gain is negligible without proposal noise, while reference-error SD 0.06 raises residual MSE above direct readout. These controls identify reference-dependent noise reduction as the numerical consistency mechanism.

5 Discussion and Conclusion

ThesisAgent specifies an explicit transition contract for local thesis revision and audits its effects on decisions. The results support evaluating persistence, evidence locality, and decision quality jointly. The contribution builds on established hypothesis representations and filtering mechanisms.

ThesisWorld evaluates numerical adapters using supplied evidence annotations and an exact action map. ALFWorld shows persistence gains with and without PhaseGuard, but controller state and recent-history access prevent interpreting these gains as evidence for a strict thesis bottleneck. Financial findings remain exploratory, covering return-derived proxies over 43 months with post-hoc refinements and unequal tuning effort. Broader validation requires learned semantic attribution, independently estimated action-transition references, and untouched interactive tasks.

AI use statement

The authors wrote the initial manuscript in their native language and used GPT to translate and refine it into academic English. The authors then reviewed and revised the text sentence by sentence to

preserve the intended meaning and ensure logical consistency and clarity. For the figures, the authors created hand-drawn sketches and used GPT to generate visual drafts, which served as references for redrawing the final figures in Microsoft PowerPoint. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript and figures.

Ethics statement

This work studies long-horizon agent decision making and includes a financial decision-making instantiation. The reported experiments are intended for research evaluation and should not be interpreted as personalized investment advice or as a guarantee of financial performance.

Reproducibility statement

Appendices A–C specify the implemented state transitions, domain adapters, metrics, and evaluation settings. The artifact contains frozen inputs, trajectories, paired metrics, source hashes, and scripts that regenerate tables directly inside this single manuscript source. Financial features are lagged relative to the evaluated returns; the public source snapshot and its limitations are documented in Appendix C.2.

References

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025a.

Zhaorun Chen, Mintong Kang, and Bo Li. Shieldagent: Shielding agents via verifiable safety policy reasoning. *arXiv preprint arXiv:2503.22738*, 2025b.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies*, 22(5):1915–1953, 2009.

Stock Market Efficiency. Returns to buying winners and selling losers: Implications for. *The Journal of Finance*, 48(1):65–91, 1993.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, et al. Memoryarena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313*, 2026.

Liang Hu, Jianpeng Jiao, Jiashuo Liu, Dongyuan Mutu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, et al. Finsearchcomp: Towards a realistic, expert-level evaluation of financial search and reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 46402–46428, 2026.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25972–25981, 2025.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.

Aly Lidayan, Jakob Bjorner, Satvik Golechha, Kartik Goyal, and Alane Suhr. Abbel: Learning natural-language belief states for memory-efficient interaction. *arXiv preprint arXiv:2512.20111*, 2025.

Bijia Liu and Ronghao Dang. Finpos: A position-aware trading agent system for real financial markets. *arXiv preprint arXiv:2510.27251*, 2025.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Joykirat Singh, Zaid Khan, Archiki Prasad, Justin Chih-Yao Chen, Akshay Nambi, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. Agent-brace: Decoupling beliefs from actions in long-horizon tasks via verbalized state uncertainty. *arXiv preprint arXiv:2605.11436*, 2026.

Izumi Takahara and Teruyasu Mizoguchi. Toward auditable ai scientists: A hypothesis evolution protocol for llm agents. *arXiv preprint arXiv:2607.09195*, 2026.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Haoyu Wang, Christopher M Poskitt, and Jun Sun. Agentspec: Customizable runtime enforcement for safe and reliable llm agents. In *Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering*, pp. 2938–2950, 2026.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-{\alpha}: Learning memory construction via reinforcement learning. *arXiv preprint arXiv:2509.25911*, 2025.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

Yijia Xiao, Edward Sun, Tong Chen, Fang Wu, Di Luo, and Wei Wang. Trading-r1: Financial trading with llm reasoning via reinforcement learning. *arXiv preprint arXiv:2509.11420*, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z Pan, et al. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan W Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2024.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Jordan W Suchow, Denghui Zhang, and Khaldoun Khashanah. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *IEEE Transactions on Big Data*, 2025.

Lingfeng Zeng, Fangqi Lou, Zixuan Wang, Jiajie Xu, Jinyi Niu, Mengping Li, Yifan Dong, Qi Qi, Wei Zhang, Ziwei Yang, et al. Fingaia: A chinese benchmark for ai agents in real-world financial domain. *arXiv preprint arXiv:2507.17186*, 2025.

Wentao Zhang, Yilei Zhao, Chuqiao Zong, Xinrun Wang, and Bo An. Finworld: An all-in-one open-source platform for end-to-end financial ai research and deployment. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2890–2900, 2026.

Tianyu Zhou, Pinqiao Wang, Yilin Wu, and Hongyang Yang. Finrobot: Ai agent for equity research and valuation with large language models. *arXiv preprint arXiv:2411.08804*, 2024.

A Additional Method Details

Table 5: Implemented scope by domain. The framework is realized by separate numerical and prompted adapters with the learning procedures listed below.

Component	ThesisWorld	Finance proxy	ALFWorld adaptation
Evidence interpreter	Cosine similarity to numerical hypothesis prototypes	Lagged, cross-sectionally standardized return features	Frozen `gpt-4.1-mini` prompt
State update	Hard top-1 local log-odds update and terminal verification	Three numerical slots with monthly verification and rollover	Prompted textual revision; no numerical hard gate
Policy	Fixed weighted confidence map	Frozen confidence-weighted score and long-only projection	Single admissible action ID, with recent interaction context
Consistency	Closed-form scalar quadratic rule, $\lambda = 2$	No separate learned consistency predictor	No distribution-level consistency loss
Execution	Clip scalar action to $[-1, 1]$	Monthly weights and transaction-cost accounting	Admissibility plus optional stateful PhaseGuard
Learning	No gradient training; fixed simulation settings	Thesis parameters selected on development data; baseline fitting disclosed	No model fine-tuning; prompt/controller development
Claim tested	Local revision, persistence, scalar decision audit	Historical proxy allocation under the stated window	Adapted system behavior and memory/controller comparisons

State operations and scope guarantee. Numerical hypotheses retain an identifier, proposition, confidence, verification horizon and rule, lifecycle state, and evidence provenance. Admission checks required fields, unique IDs, normalized exact-text duplicates, and capacity; these are syntactic rather than entailment checks. Retirement preserves confidence and provenance while disabling ordinary revision. Terminal slots may receive active successors. Finance distinguishes these with a storage-slot ID and issuance month; identity and proposition text remain fixed within each lifetime.

Candidates are ordered by descending relevance, then hypothesis ID and evidence ID. The gate admits at most one record per hypothesis and at most M hypotheses. Unselected hypotheses retain the same

immutable objects, giving exactly zero ordinary-update drift outside the selected scope. False-positive selections are covered by Appendix F. Reveals set finite terminal confidence and block subsequent ordinary revision. Log-odds clipping is equivalent to $\epsilon = \sigma(-L)$ in Equation 25.

Numerical evidence and reliability. ThesisWorld interprets an observation embedding by its cosine similarity to each hypothesis prototype. Relevance is the positive part of this similarity, eligibility requires $r > 0.45$, and $M = 1$. The signed revision is 1.34 times the observation sign. The simulator provides source reliability ρ; duplicate evidence receives novelty $n = 0.18$, otherwise $n = 1$. A delayed reveal promotes the corresponding record to relevance one and supplies its binary outcome. Confidence log-odds are capped at $L = 6$. These simulator-supplied reliability and duplicate labels define the controlled evidence interface.

Decision and consistency rule. For normalized positive weights w_k, the fixed readout is $f(c) = \tanh(2\sum_k w_k(2c_k - 1))$. All numerical comparators use $u_t = f(c_t) + \xi_t$, with $\xi_t \sim \mathcal{N}(0, 0.045^2)$ and identical action-noise draws from an independent stream on each paired episode. ThesisAgent uses the known transition $G_t = f(c_t) - f(c_{t-1})$ and $\lambda = 2$ in Equation 27. The action-smoothing control instead applies $0.72\tilde{a}_{t-1} + 0.28u_t$. Appendix G separately tests imperfect references.

Strict component ablations. The no-persistent-state variant rebuilds the thesis from the episode-initial prior at each observation, retaining the evidence interpreter, verification, action map, action noise, and consistency rule. Dense revision makes all positive-relevance active hypotheses eligible with update budget K; all other update terms, noise, and forgetting settings are unchanged. The artifact key `Global Rewrite` refers to this numerical *w/o Locality Gate (Dense Revision)* control for backward compatibility. The consistency ablation sets $\lambda = 0$, and the verification ablation suppresses delayed reveals. Every component control retains the thesis-based action readout.

A.1 PhaseGuard and the ALFWorld Memory Interface

The ALFWorld adaptation jointly proposes a revised textual thesis and an admissible action ID in one language-model response. Revisable variants receive up to eight recent action–observation records. PhaseGuard variants receive four. Each record includes the executed action and subsequent observation, together with a no-progress indicator. The thesis schema distinguishes verified facts from uncertainties and retains invalidated assumptions. It also records the task phase and next subgoal. Locality is a prompt instruction in this adaptation. It does not implement the numerical top-M gate.

Controller rules and state. PhaseGuard parses the target object, required transformation, destination, and object count. It first selects an admissible target pickup or needed receptacle opening. When carrying the target, it prioritizes the required cleaning, heating, cooling, or lighting step and navigation to the relevant facility, followed by delivery. Otherwise, it retains an eligible proposal or selects an unvisited location, with least-used navigation as fallback. Ties follow admissible-action ordering or deterministic action keys. Its persistent ledger tracks visited locations, action counts, carried objects, transformation status, and delivered instances, updating from executed actions. The controller reads no textual thesis and can replace a valid proposal, so it provides task-specific control beyond admissibility.

Matched thesis removal. The no-persistent-thesis controls reset thesis text while preserving the matched recent interaction records. Guarded variants also share the same PhaseGuard rules and persistent ledger. The runner flag `--preserve-ablation-history` selects this setting; intervention counts from earlier joint-memory-removal runs are kept separate. Table 2 thus measures the added thesis state within the same history/controller interface, rather than a strict thesis-only history bottleneck.

A.2 Learned adapters as future work

The interface can accommodate learned bounded evidence scores and a task-trained policy. A consistency objective could use detached thesis inputs to block gradients into the updater, alongside a task objective. Its transition target would need independent justification: fitting it only to reproduce the current policy can yield trivial agreement. Learning this reference, including for distribution-valued

discrete actions, remains unevaluated future work. The reported experiments use the implementations in Table 5.

B Evaluation Metric Definitions

This section defines the metrics used in the main tables and figures. Unless stated otherwise, episode-level ThesisWorld quantities are first computed within an episode and then averaged over the paired test episodes; arrows in the tables give the preferred direction.

B.1 Thesis quality and evidence revision

Let $c_{t,k} \in [0,1]$ be the confidence assigned to hypothesis k after evidence step t, and let $z_k \in \{0,1\}$ be its simulator-provided truth value. The **Brier score** and **negative log-likelihood (NLL)** are

$$\text{Brier} = \frac{1}{TK} \sum_{t=1}^{T} \sum_{k=1}^{K} (c_{t,k} - z_k)^2, \tag{6}$$

$$\text{NLL} = -\frac{1}{TK} \sum_{t=1}^{T} \sum_{k=1}^{K} \left[z_k \log c_{t,k} + (1 - z_k) \log(1 - c_{t,k}) \right], \tag{7}$$

with probabilities clipped away from zero for numerical stability. Lower values are better. **Expected calibration error (ECE)** partitions final confidences into equal-width bins and sums each bin's mass times the absolute gap between mean confidence and empirical frequency. **Evidence Attribution F1** compares the predicted observation–hypothesis relevance mask with the simulator's relevance mask and uses $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$.

Unaffected-Belief Change (UBC) measures unintended movement among hypotheses known to be irrelevant to the current observation:

$$\text{UBC} = \mathbb{E}\left[|c_{t,k} - c_{t-1,k}| \mid y^{\text{rel}}_{t,k} = 0 \right]. \tag{8}$$

Lower UBC indicates better isolation of local revisions. Let t_0 be the start of the simulator's designated decisive-evidence burst for the deliberately incorrect prior, and define truth-aligned confidence as $q_{t,k} = z_k c_{t,k} + (1 - z_k)(1 - c_{t,k})$. **Sustained Revision Delay** is the smallest $d \geq 1$ such that $q_{t_0+d-1+j,k} \geq 0.60$ for every $j \in \{0,1,2,3\}$ within the observed trajectory. The triggering observation counts as step one. If no such run occurs, the implementation assigns the number of remaining post-update states as a cap. Tables report the mean of this *capped score*, not a survival-analysis estimate or a mean conditional on successful revision. A maximum value can represent failure to revise, not successful revision at the last step. The four-state rule extends beyond the simulator's three-observation decisive-evidence burst. UBC, delay, and task return are interpreted together so that neither indiscriminate updating nor never updating appears optimal.

B.2 Thesis–action grounding

Let $\tilde{a}_t$ denote the desired action before feasibility projection and let $\overline{\Delta B}_t = K^{-1} \sum_k |c_{t,k} - c_{t-1,k}|$. The **Unsupported Transition Rate (UTR)** is

$$\text{UTR} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left\{ |\tilde{a}_t - \tilde{a}_{t-1}| > 0.18 \ \wedge \ \overline{\Delta B}_t < 0.025 \right\}, \tag{9}$$

evaluated where execution context and feasibility constraints are controlled. Lower values indicate fewer action changes unsupported by a material thesis change.

Required-Revision Response (RRR) tests whether the designated decisive-evidence burst moves the desired action toward the independently known oracle action $a^\star(z)$. For its start at t_0, define

$$d_{\text{pre}} = |\tilde{a}_{\max(t_0-1,0)} - a^\star(z)|, \tag{10}$$

$$d_{\text{post}} = \min_{0 \leq j \leq 3} |\tilde{a}_{t_0+j} - a^\star(z)|. \tag{11}$$

An episode receives RRR equal to one when $d_{\text{post}} \leq \max(d_{\text{pre}} - 0.025, 0.04)$ and zero otherwise; reported RRR is the episode mean. Higher values indicate more frequent task-relevant responses.

For the ThesisWorld specialization of Equation 5, use $Q_t(a) = -|a - a^\star(z)|/2$, where the oracle depends on simulator truth z. The truth-aligned state intervention is

$$B_t^{k,\text{align}} = B_t[c_{t,k} \leftarrow \text{clip}(c_{t,k} + \eta(2z_k - 1), 0, 1)]\,. \tag{12}$$

Each branch follows the method's complete decision pipeline, including its consistency filter or action smoother when present, followed by execution projection. The branches reuse the same pre-intervention controller state and exogenous action-noise draw, with observation context fixed. The state updater is not rerun. If a_t^{cf} is the resulting executed action, **Counterfactual Decision Regret (CDR)** is

$$\text{CDR} = \mathbb{E}_{t,k}\left[\frac{|a_t^{\text{cf}} - a^\star(z)|}{2}\right]. \tag{13}$$

The expectation is taken over eight equally spaced episode states and all hypotheses, and the divisor is the width of the normalized action interval $[-1, 1]$. The simulator oracle is independent of the deterministic consistency reference. CDR measures remaining post-intervention error; intervention benefit is assessed against unperturbed regret on the same sampled states using the paired reduction

$$\Delta_{\text{CF}} = \mathbb{E}_{t,k}\left[\frac{|a_t - a^\star| - |a_t^{\text{cf}} - a^\star|}{2}\right], \tag{14}$$

with $a^\star = a^\star(z)$. The divisor is the width of the normalized action interval. This is Δ_Q under the stated negative-error score. Positive values indicate reduced error after the truth-aligned intervention. The bidirectional sensitivity $F_{t,k}$ in Equation 29 is measured on desired actions before clipping; executed-action sensitivity is retained separately in the artifact. These diagnostics assess local numerical state use.

B.3 Task and interaction utility

For ThesisWorld, **normalized episode return** is $T^{-1}\sum_t[1 - |a_t - a^\star(z)|/2]$, where a_t is the executed action; **success** equals one when the final action is within 0.25 of the oracle action. In ALFWorld, **task success** is the fraction of the 134 tasks completed within the step budget. **Steps** averages environment steps over all tasks, whereas **Steps (succ.)** conditions on successful tasks. **Invalid-action rate** is the number of rejected actions divided by total environment steps. **Calls** and **tokens** are total logical model calls and total model tokens, respectively, over the complete split.

B.4 Financial signal and portfolio metrics

For asset i and signal month t, **RankIC** is the cross-sectional Spearman correlation between the pre-projection model score and the next realized return,

$$\text{RankIC}_t = \text{Spearman}_i\left(s_{i,t}, r_{i,t+1}\right), \qquad \text{RankIC} = \frac{1}{M}\sum_{t=1}^{M}\text{RankIC}_t. \tag{15}$$

Equal weight has no cross-sectional score and therefore no RankIC. A positive RankIC indicates useful ranking information but does not by itself imply profitable net execution.

Let w_t be the new target weights and $\widetilde{w}_{t-1}$ the previous target after drifting with the returns realized since the preceding rebalance. Monthly turnover and net portfolio return are

$$\text{TO}_t = \frac{1}{2}\|w_t - \widetilde{w}_{t-1}\|_1, \tag{16}$$

$$R_t^{\text{net}} = w_t^\top r_{t+1} - \kappa\,\text{TO}_t, \tag{17}$$

where $\kappa = 0.001$, or 10 bp per unit of reported half-ℓ_1 turnover, in the main table. This is the coefficient on the stated turnover measure, not on total traded notional. Portfolios begin with zero risky-asset exposure; establishing a fully invested portfolio therefore contributes $\text{TO} = 0.5$ under this convention. With cumulative wealth $W_t = \prod_{u=1}^{t}(1 + R_u^{\text{net}})$, the reported **cumulative net return** is $W_M - 1$. **Excess cumulative return** is its difference from Equal Weight in percentage points, not

a compounded active-return series. The annualized **Sharpe ratio** is $\sqrt{12}\,\overline{R}/s_R$, using the sample standard deviation of monthly net returns and no risk-free-rate subtraction. **Maximum drawdown** is $\max_{1\le t\le M}[1 - W_t/\max_{1\le u\le t} W_u]$. The implementation uses observed month-end wealth peaks, without prepending $W_0 = 1$ to the drawdown series. Reported **turnover** is $M^{-1}\sum_t \mathrm{TO}_t$.

C Additional Experimental Details

ALFWorld development and provenance. An audit of the earlier frozen subset identified target-parsing errors, non-target pickup, prerequisite-order violations, and 50-step search exhaustion. PhaseGuard candidates were screened only on `valid_seen`: persistent and no-persistent-thesis variants both solved 8/8 screening tasks and 20/20 tasks on a separate confirmation subset. After inspecting the 20-task result, we kept the runner and benchmark source hashes fixed and ran the post-hoc census of all 134 solvable `valid_unseen` tasks, including tasks touched during development.

All ten methods share the indexed action-ID interface and 6,700-call ceiling; retained runs contain zero request-failure records and use 2,282–6,166 logical calls. Earlier cached responses are reused only when method, episode, step, model, system-prompt, and user-prompt hashes match. A non-table ThesisAgent-Legacy development row was stopped before aggregation and excluded. The artifact retains this exclusion note, complete manifests for the six baseline and four ThesisAgent rows, 1,340 method–task trajectories, paired outcomes, intervention flags, model-call caches, budget ledgers, and source hashes.

Post-hoc metric corrections. CDR was changed from distance to a thesis-conditioned reference to oracle-referenced executed-action error (Equation 13), with $\eta = 0.20$, baseline regret, and paired change. Revision delay was changed from first threshold crossing to confidence at least 0.60 for four consecutive states, counting the trigger as step one and capping failures at the remaining window (Appendix B). These corrections distinguish decision quality from reference agreement and retained correction from transient response. Current tables use the corrected metrics; earlier intervention-size sweeps remain archived separately.

Sweep and selection audit. A post-hoc audit corrected overlap between the original consistency sweep and formal-test seeds. The current sweep uses only disjoint seeds 10,000–10,009. A subsequent search over supported-transition scales found no candidate satisfying the frozen joint RRR, UTR, CDR, and utility promotion gates, leaving the primary configuration unchanged. The artifact retains the null-selection ledger and continuous before/after-distance diagnostic. Controller definitions and common action-noise settings are specified in Appendix A.

C.1 ThesisWorld generator and comparison protocol

An episode has six independent Bernoulli-0.5 latent truths and six positive decision weights drawn uniformly from $[0.55, 1.45]$ and normalized to sum to one. Initial confidence is drawn from $[0.54, 0.72]$ for true propositions and $[0.28, 0.46]$ for false ones. The highest-weight proposition is deliberately initialized incorrectly at 0.22 or 0.78, respectively, for every method. The six hypothesis prototypes are the first six basis vectors in ten dimensions.

The base episode has 48 steps. Outside the decisive burst and duplicate branch, evidence is relevant with probability 0.76. Relevant embeddings add Gaussian noise with coordinate standard deviation 0.34 to the target prototype. Source reliability is sampled from $\{0.58, 0.72, 0.88, 0.95\}$ with probabilities $\{0.15, 0.30, 0.35, 0.20\}$; the evidence sign agrees with truth with that reliability. The duplicate branch has probability 0.16 when a previous relevant event is available and perturbs its embedding with standard deviation 0.025 while retaining its sign and source. Irrelevant evidence has a random sign and a Gaussian embedding whose first six coordinates are multiplied by 0.22. Three consecutive decisive observations begin at zero-based step $\mathrm{round}(0.56T)$, target the deliberately incorrect prior, have reliability 0.98, and use 0.35 times the base embedding noise. After the midpoint, eligible newly generated relevant events reveal their target truth with probability 0.16. Reveals are observed evidence shared by methods, not unrestricted access to latent truth.

Main and ablation experiments use seed groups $s = 0, \ldots, 9$ and 80 episodes per group, with episode seed $100000s + e$ for $e = 0, \ldots, 79$. Each horizon/noise setting uses 20 episodes per group.

Table 6: State-intervention diagnostics on the same 800 ThesisWorld episodes, with $\eta = 0.20$. Entries are means $\pm$ SEM over ten seed groups. Baseline and aligned regret use identical sampled states; Δ_{CF} is their paired reduction. Sensitivity F measures dependence without a universal preferred direction; interpret regret reductions relative to each method's starting error.

Method	Desired-action F	Baseline regret	Aligned regret (CDR)	Δ_{CF}
Full-History	0.17826 ± 0.00099	0.19421 ± 0.00300	0.18420 ± 0.00285	0.01002 ± 0.00030
Rolling-Summary	0.22374 ± 0.00062	0.19932 ± 0.00168	0.18694 ± 0.00148	0.01238 ± 0.00031
ReAct	0.21701 ± 0.00028	0.26184 ± 0.00301	0.24760 ± 0.00268	0.01424 ± 0.00048
Reflexion	0.21456 ± 0.00074	0.19378 ± 0.00167	0.18235 ± 0.00153	0.01143 ± 0.00029
ABBEL-style bottleneck	0.17070 ± 0.00104	0.19167 ± 0.00299	0.18281 ± 0.00291	0.00886 ± 0.00022
Agent-BRACE-style atomic	0.17629 ± 0.00076	0.17585 ± 0.00280	0.16612 ± 0.00262	0.00973 ± 0.00028
ThesisAgent	0.18002 ± 0.00075	0.16365 ± 0.00203	0.15361 ± 0.00184	0.01004 ± 0.00033

Horizons are $24, 48, 96, 192$; in that sweep the burst starts at $\lfloor 0.56T \rfloor$. Noise standard deviations are $0.18, 0.34, 0.52, 0.70$. The consistency sweep uses $\lambda \in \{0, 0.5, 1, 2, 4, 8\}$ and separate seed groups $10000, \ldots, 10009$. These sweeps evaluate the fixed numerical update and controller at each horizon and noise level.

Numerical comparator definitions. All methods use the same action map and noise scale, but differ in belief updates. Full-History recomputes prior log-odds plus accumulated attributed evidence with coefficient 0.78. Rolling-Summary retains 0.91 of previous log-odds and adds evidence with coefficient 0.78. The ReAct-style adapter uses only the current attributed evidence with coefficient 1.65. The Reflexion-style adapter uses retention 0.93, evidence coefficient 0.82, and feedback every eight steps only for outcomes already revealed. The ABBEL-style adapter uses retention 0.97, evidence coefficient 0.95, and a shared mean-evidence term of 0.30. The Agent-BRACE-style adapter combines local evidence coefficient 0.93 with cross-claim coupling 0.018. Non-core comparators reset revealed log-odds to ± 4 and retain their subsequent update rules; the thesis core uses terminal lifecycle states and ± 6. The main table therefore compares complete numerical adapters, while internal component controls isolate the named switches.

Protocol corrections and output provenance. Post-hoc control corrections introduced common action noise, removed unrevealed-truth access from Reflexion-style feedback, and replaced multi-factor ablations with the single-component controls in Appendix A. Current tables and Figure 3 use the corrected source hashes and outputs frozen in `paper_revision_20260926/world`; historical metric and parameter-search records are archived separately.

Evidence-level intervention protocol. Twenty separate episodes with seeds $2000000, \ldots, 2000019$ test the complete updater. At each pre-update state, two branches replace the current evidence sign by $+1$ and -1, holding its embedding, reliability, novelty, policy context, and action noise fixed. Verification is disabled in both probe branches; the ordinary trajectory continues with the original event. Each method has 762 relevant and 198 irrelevant paired probes. ThesisAgent's mean desired-action effect is 0.10673 on relevant events and zero on irrelevant events. The dense-revision control gives 0.15221 and 0.04220, respectively. The zero effect is specific to these synthetic probes.

C.2 Financial data and execution protocol

The dataset is the Kenneth French library's 49 value-weighted US industry portfolios, downloaded as the monthly `49_Industry_Portfolios` CSV archive. The retained raw file identifies the July 2026 CRSP database snapshot. Features use history beginning in January 2000, and the reported evaluation contains 43 monthly returns from January 2023 through July 2026 for all 49 portfolios. The artifact records the public source URL and SHA-256 hashes of the raw file and processed dataset. Lagged use of this snapshot does not establish historical publication-vintage availability or absence of later revisions.

Three cross-sectionally standardized features supply momentum, return-surprise, and price-to-peak evidence. Confidences start at 0.5. Each feature updates its corresponding slot with relevance, reliability, and novelty equal to one, using signed revision $0.30 \tanh(e/1.5)$ and log-odds cap $L = 2.5$. The component score weights are $(0.55, 0.10, 0.35)$. After the next return is observable, its comparison

Table 7: Common top-quintile selection audit on the same January 2023–July 2026 interval. All score-based rows select ten of 49 portfolios, with FinPos retaining its quantity and persistence rules. Scores are frozen; only ThesisAgent's selection fraction changes relative to Table 3. Net return is cumulative and excess is in percentage points. Best: dark gray and bold; second-best: light gray and underlined; ties share formatting.

Method	RankIC ↑	Cumulative Net Return ↑	Excess (pp) ↑	Sharpe ↑	Max Drawdown ↓	Turnover ↓
Equal Weight (DeMiguel et al., 2009)	–	75.7%	0.0	1.18	11.0%	**3.0%**
Momentum (Efficiency, 1993)	0.007	52.7%	-23.0	0.75	12.3%	27.6%
XGBoost (Chen & Guestrin, 2016)	0.010	87.4%	11.7	0.97	16.4%	35.6%
Rolling Summary	0.022	76.7%	1.0	0.96	13.6%	34.9%
FinMem (Yu et al., 2025)	0.003	81.8%	6.1	1.14	10.7%	56.2%
FinCon (Yu et al., 2024)	-0.019	63.5%	-12.2	0.96	12.5%	62.4%
TradingAgents (Xiao et al., 2024)	-0.013	86.5%	10.8	1.19	12.3%	58.9%
Trading-R1 (Xiao et al., 2025)	-0.058	35.0%	-40.7	0.73	**9.8%**	37.6%
FinPos (Liu & Dang, 2025)	-0.024	58.3%	-17.4	0.94	11.6%	24.9%
ThesisAgent	**0.039**	**107.6%**	**31.9**	**1.36**	13.1%	7.5%

with the contemporaneous cross-sectional median verifies the expired monthly proposition. The successor inherits the pre-verification confidence, so the terminal label does not directly become the next prediction. The development record contains 720 first-stage and 168 second-stage candidates over 2005–2016 windows across four markets. Baseline development effort was not matched to this search budget.

The first reported portfolio is formed using December 2022 signals. All methods begin the reporting interval with zero risky holdings, but their causal signal histories are retained. Main-table ThesisAgent selects the top 16%, or eight portfolios; other score-based rows select the top 20%, or ten. Equal Weight holds all 49. FinPos additionally retains inverse-volatility quantities, a 0.20 position cap, and 0.60 position persistence. Table 7 changes only ThesisAgent's selection fraction to 20%, using the same frozen scores and accounting. This controls selection concentration, not all position-sizing rules or development effort.

XGBoost uses 260 estimators, depth three, learning rate 0.025, minimum child weight 12, subsampling 0.80, column subsampling 0.85, and L2 regularization four, with seed 42 and chronological fitting. The external-agent rows use the return-proxy mechanisms described below, not news, filings, or pretrained language-model calls. The frozen-score audit reconstructs the original reporting-window gross returns, net returns, turnover, and RankIC with zero maximum absolute discrepancy. Inputs, scores, positions, and protocol hashes are retained in the financial audit artifact.

Financial-agent adaptation audit. FinMem is represented by shallow, intermediate, and deep causal memory states with the decay horizons specified by the paper. FinCon uses specialized proxy analysts, a manager aggregation rule, a downside-risk penalty, and strictly lagged performance feedback. TradingAgents uses separate technical, event, sentiment, and valuation-proxy reports followed by bull–bear synthesis and risk review. Trading-R1 uses structured thesis, evidence, and volatility-adjusted decision stages; its decision surrogate is trained only on development targets realized by December 2016. FinPos separates direction from inverse-volatility quantity choice and blends desired weights with the carried position. Every adapter is generated by the released runner, receives the same return-derived information set, and is marked as an adaptation because the public portfolio data do not contain the original systems' point-in-time textual inputs or model checkpoints.

Financial reporting scope. All financial results use the same 43-month interval and independent portfolio initialization. The artifact records development and reporting choices; results are exploratory point estimates on this inspected path.

D Extended Evaluation Setup and Statistical Protocol

D.1 Research Questions

RQ1: Belief quality and revision. We measure calibration, evidence attribution, drift on unaffected hypotheses, and sustained response to reliable contradiction to distinguish retained corrections from inertia.

RQ2: Decision dependence and quality. We measure unsupported transitions, required-revision response, and paired hypothesis interventions with oracle-referenced regret to separate policy sensitivity from useful correction.

RQ3: Behavior across domain adaptations. Task success and utility are evaluated in ThesisWorld, the complete solvable ALFWorld split, and the transaction-cost-adjusted financial proxy. Table 5 identifies the mechanisms tested in each adapter.

D.2 Experimental Environments

Controlled long-horizon environment. Appendix C.1 specifies ThesisWorld's generator, numerical adapters, and horizon/noise sweeps. Component ablations match the update and decision pipeline except for the named switch. Label-exposure calibration uses disjoint validation seeds and selects the reveal rate using exposure counts rather than method scores.

Interactive cross-domain benchmark. ALFWorld (Shridhar et al., 2020) 0.4.2's text-only `AlfredTWEnv` supplies object-location and prerequisite-completion tasks. SHA-256 ordering with seed 3701 fixes the order of all 134 solvable `valid_unseen` tasks. Each method uses `gpt-4.1-mini` at temperature zero, a zero-based admissible-action catalog, a 50-step budget, and a 6,700-call ceiling. The manifest SHA-256 is `d9efc5248aab2d51a716b0db3de5ac909146016cd5df2091b7570ee5b9409be4`. The census followed subset inspection and includes development-touched tasks, as documented in Appendix C.

Long-horizon equity decision making. The 49 industry-portfolio return series supply three proxy thesis slots:

$$B_{i,t} = \left\{ B_{i,t}^{\text{fact}}, B_{i,t}^{\text{surprise}}, B_{i,t}^{\text{value}} \right\}, \tag{18}$$

Six-month momentum, lagged-return surprise, and distance from the trailing 12-month price peak instantiate these slots. The shared code maintains monthly identities, delayed verification, and successors initialized from pre-verification confidence. Appendix C.2 specifies the January 2023–July 2026 window, development parameters, input cutoffs, and top-16%/top-20% projection comparison.

D.3 Baselines and Controlled Comparisons

The main tables compare complete mechanism-inspired adapters under common inputs, action interfaces, and execution accounting, rather than native system reproductions. Numerical update rules and financial adapters are specified in Appendices C.1 and C.2. Component ablations isolate the switches in Appendix A; ALFWorld thesis removal preserves matched history and controller state (Appendix A.1). The financial state controls additionally fix features, weights, and portfolio projection (Appendix H).

D.4 Metrics

Appendix B defines all thresholds and aggregation rules. Belief metrics assess retained state, action diagnostics assess dependence and response quality, and task outcomes assess full-trajectory utility. Financial RankIC measures signal ranking, while return, Sharpe, drawdown, and turnover characterize portfolio execution.

D.5 Statistical Protocol and Leakage Controls

ThesisWorld uses the same 800 independent episodes across methods and reports SEM over ten seed-group means. Two-sided episode-paired tests compare each of six adapters with ThesisAgent: Wilcoxon tests for Brier, UBC, CDR, and return, and exact discordant-pair tests for binary RRR. Holm correction covers all 30 comparisons jointly; attribution F1 is descriptive. Seven directional post-hoc ablation contrasts form a separate Holm family.

ALFWorld reports success counts and rates over the 134-task post-hoc census. Without PhaseGuard, the 95% Wilson intervals are $[69.0\%, 83.2\%]$ and $[82.4\%, 93.1\%]$ for the no-persistent-thesis and ThesisAgent revisable variants, respectively. With PhaseGuard, the corresponding intervals are

$[86.8\%, 95.9\%]$ and $[94.7\%, 99.6\%]$. These intervals summarize task-level success rather than repeated-run variability. Matched success-rate differences are interpreted descriptively within the post-hoc census, without claims of paired statistical significance. Finance reports point estimates on one 43-month path, without cross-seed significance claims.

Financial scores for month $t+1$ use observations through month t. XGBoost fitting uses only observations preceding each evaluation fold; FinCon feedback waits for its associated realized return; Trading-R1 fitting uses only development-period targets. ThesisAgent verifies the proposition issued at $t-1$ once its return is observed at t, then rolls over to the next active horizon before processing month-t evidence. A realized return never forms its own forecast. Turnover uses return-drifted previous holdings (Appendix B). These numerical adapters make no pretrained language-model calls; the historical snapshot-vintage limitation remains as stated in Appendix C.2.

E Expanded State Interface

This appendix specifies the state schema and numerical realization of the general interfaces in Section 3. Table 5 separates the numerical adapters from ALFWorld's prompted thesis and execution controller.

E.1 Problem Formulation and ThesisAgent Overview

At each step, evidence e_t revises the previous thesis, while execution context x_t determines operational conditions and feasible actions $\mathcal{C}_t$:

$$\begin{aligned} z_t &= \mathcal{M}_{\mathrm{dom}}\big(e_t, \mathrm{Serialize}(B_{t-1})\big), \\ B_t &= U_\theta(B_{t-1}, z_t). \end{aligned} \tag{19}$$

The interpreter produces hypothesis-specific evidence records using a language model or numerical map; several functional roles may share one model call. The numerical experiments use fixed hypothesis slots. The updater commits persistent changes. The numerical controller retains $m_{t-1} = (B_{t-1}, \tilde{a}_{t-1})$ and uses

$$\begin{aligned} \tilde{a}_t &= \pi_\phi(B_t, x_t, m_{t-1}; \xi_t), \\ a_t &= \Pi_{\mathcal{C}_t}(\tilde{a}_t), \end{aligned} \tag{20}$$

This specializes Equation 3 to projection $\Pi_{\mathcal{C}_t}$, with proposal noise ξ_t. The base readout uses B_t; recurrent filtering additionally uses the declared preceding thesis and desired action, without a separate raw-history input. ALFWorld's additional recent-history and controller channels are specified in Appendix A.1.

Decision thesis versus a general belief state. A thesis retains decision-supporting assumptions with stable identities and explicit resolution conditions. In an illustrative delivery task, a verified closure revises the route-open hypothesis while preserving the charge-sufficient hypothesis. The policy can change routes without reconstructing the energy assessment. This illustrates the transition contract; numerical tests use fixed propositions and action maps.

E.2 Persistent State Representation

A bounded collection retains hypothesis identities across revisions:

$$B_t = \{h_{t,k} \mid k \in \mathcal{I}_t\}, \qquad |\mathcal{I}_t| \le K_{\max}, \tag{21}$$

Here $\mathcal{I}_t$ indexes retained hypotheses. Each record is

$$h_{t,k} = \big(q_k, c_{t,k}, \tau_k, v_k, s_{t,k}, \mathcal{E}^+_{t,k}, \mathcal{E}^-_{t,k}\big). \tag{22}$$

The proposition q_k remains fixed as confidence $c_{t,k}$ changes. Horizon τ_k and rule v_k specify when and how an observation can resolve it. Lifecycle state $s_{t,k}$ records active, verified, invalidated, or retired status, while $\mathcal{E}^+_{t,k}$ and $\mathcal{E}^-_{t,k}$ retain supporting and contradicting provenance. Verification requires the designated observation; expiry or retirement alone does not establish falsity. Candidate admission checks required fields, identity, exact-text duplicates, and capacity as specified in Appendix A. These storage checks do not certify semantic atomicity or real-world verifiability; open-ended hypothesis discovery is outside the fixed-slot evaluation.

E.3 Evidence-Local Revision

The update separates eligibility from revision strength so that weak relevance alone cannot repeatedly alter every hypothesis.

Evidence attribution. For an active hypothesis, record $z_{t,k}$ links evidence to q_k and its source, with relevance $r_{t,k} \in [0,1]$ and signed revision $\Delta_{t,k} \in [-\beta, \beta]$. Positive and negative revisions express support and contradiction; relevance may be high even when the directional revision is small. Appendix A specifies numerical scores.

Locality gate. For active identifiers $\mathcal{A}_{t-1}$, select

$$\begin{aligned} \mathcal{R}_t &= \{k \in \mathcal{A}_{t-1} \mid r_{t,k} > \gamma_r\}, \\ \mathcal{K}_t &= \mathrm{TopM}(\mathcal{R}_t, r_{t,\cdot}). \end{aligned} \tag{23}$$

The threshold γ_r controls eligibility; M bounds the number of editable hypotheses. The gate ranks by relevance with deterministic tie-breaking and retains all eligible entries when fewer than M qualify. Relevance also scales each admitted revision.

Bounded confidence revision. Source reliability $\rho_t \in [0,1]$ and novelty $n_t \in [0,1]$ scale admitted evidence, discounting repetition. With $\epsilon \in (0, 1/2)$ bounding stored confidence, compute

$$\widehat{c}_{t,k} = \sigma(\mathrm{logit}(c_{t-1,k}) + r_{t,k}\rho_t n_t \Delta_{t,k})\,. \tag{24}$$

and commit

$$c_{t,k} = \begin{cases} \mathrm{clip}(\widehat{c}_{t,k}, \epsilon, 1-\epsilon), & k \in \mathcal{K}_t, \\ c_{t-1,k}, & k \in \mathcal{I}_{t-1} \setminus \mathcal{K}_t. \end{cases} \tag{25}$$

Initialization uses the same bounds. Verification overrides the proposal, sets confidence to $\sigma(L)$ or $\sigma(-L)$ for a true or false outcome, and marks the hypothesis terminal. Terminal entries are excluded from ordinary revision. The bounded log-odds rule preserves unselected confidences exactly; attribution mistakes are quantified in Appendix F. Accepted updates retain provenance, and admission and lifecycle changes remain explicit operations.

E.4 Decision–Execution Separation

The numerical readout and recurrent state in Equation 20 determine the desired decision. Feasibility handling is a separate operation, allowing thesis effects to be audited at fixed execution context.

Desired decisions and feasible actions. The numerical execution adapter minimizes deviation from the proposal:

$$a_t = \Pi_{\mathcal{C}_t}(\tilde{a}_t) = \arg\min_{a \in \mathcal{C}_t} D_{\mathrm{exec}}(a, \tilde{a}_t). \tag{26}$$

ThesisWorld clips its scalar action to $[-1, 1]$. ALFWorld instead proposes one admissible action ID, which PhaseGuard may replace to enforce task prerequisites (Appendix A.1); it exposes no action distribution for a numerical consistency loss.

Thesis–action consistency. At fixed execution context, the numerical controller uses a known readout $f(B_t)$, proposal $u_t = f(B_t) + \xi_t$, and supported transition $\widehat{\Delta\tilde{a}}_t = f(B_t) - f(B_{t-1})$. For $\lambda_{\mathrm{cons}} \geq 0$, it solves

$$\begin{aligned} \tilde{a}_t &= \arg\min_a \big[(a - u_t)^2 + \lambda_{\mathrm{cons}}(a - \tilde{a}_{t-1} - \widehat{\Delta\tilde{a}}_t)^2\big] \\ &= \frac{u_t + \lambda_{\mathrm{cons}}(\tilde{a}_{t-1} + \widehat{\Delta\tilde{a}}_t)}{1 + \lambda_{\mathrm{cons}}}. \end{aligned} \tag{27}$$

With $\varepsilon_t = \tilde{a}_t - f(B_t)$, this is ordinary residual filtering: $\varepsilon_t = (\xi_t + \lambda_{\mathrm{cons}}\varepsilon_{t-1})/(1 + \lambda_{\mathrm{cons}})$. Unlike direct action smoothing, it filters deviations around the changing belief readout. Appendix F characterizes reference error, and Appendix G tests denoising and utility separately. Learning an independent transition reference remains future work (Appendix A.2).

Execution procedure.

1. Initialize the bounded thesis with stable IDs and verification rules. Set $\tilde{a}_{-1} = 0$ and retain the pre-update thesis.
2. Interpret evidence into relevance, sign, reliability, novelty, and available verification outcomes.
3. Select at most M eligible active hypotheses and apply Equation 25. Then commit verification and provenance.
4. Form the desired decision using Equation 27 in ThesisWorld or the stated portfolio rule in finance, and execute the feasible action.
5. Retain $m_t = (B_t, \tilde{a}_t)$ for the next numerical decision. Audits use state copies and matched noise, leaving the live trajectory unchanged.

E.5 Intervention-Based Utility Audit

Intervene on one stored hypothesis, including terminal slots, while holding the other hypotheses, execution context, preceding controller state m_{t-1}, and noise ξ_t fixed. For $\eta > 0$,

$$\begin{aligned} B_t^{k,+} &= B_t[c_{t,k} \leftarrow \min(1, c_{t,k} + \eta)]\,, \\ B_t^{k,-} &= B_t[c_{t,k} \leftarrow \max(0, c_{t,k} - \eta)]\,. \end{aligned} \tag{28}$$

Audit inputs use $[0, 1]$ bounds rather than the persistent update's $[\epsilon, 1 - \epsilon]$ bounds. Without rerunning the updater, compute pre-execution sensitivity:

$$\begin{aligned} \tilde{a}_t^{k,\pm} &= \pi_\phi(B_t^{k,\pm}, x_t, m_{t-1}; \xi_t), \\ F_{t,k} &= D_a(\tilde{a}_t^{k,+}, \tilde{a}_t^{k,-}). \end{aligned} \tag{29}$$

The reported numerical policy uses a desired-action distance D_a; these probes are not computed on sampled ALFWorld action IDs. Sensitivity measures local dependence and can be small under saturation. Response quality is assessed separately using truth-aligned interventions followed by the same controller and projection. CDR measures remaining oracle error (Equation 13), and Δ_{CF} measures its paired reduction (Equation 14). The oracle uses latent truth, whereas the consistency reference uses estimated beliefs. Appendix B specifies sampling and aggregation. These tests concern internal state use rather than external environmental causal effects.

F Locality and Reference-Error Analysis

We characterize ordinary-update locality and reference-dependent residual error for Equation 27, using a specified numerical action map at fixed execution context. The results apply to this scalar realization under the assumptions below.

Ordinary-update locality bound. Let $\mathcal{U}_t$ be the truly unrelated hypothesis indices, with $|\mathcal{U}_t| > 0$. Consider an ordinary confidence update with no verification overrides, an already bounded prior confidence, and $|\Delta_{t,k}| \le \beta$. The sigmoid is $1/4$-Lipschitz and clipping to an interval containing the prior confidence cannot increase the change. Equation 2 therefore implies

$$\frac{1}{|\mathcal{U}_t|} \sum_{k \in \mathcal{U}_t} |c_{t,k} - c_{t-1,k}| \le \frac{\beta \rho_t n_t}{4|\mathcal{U}_t|} \sum_{k \in \mathcal{K}_t \cap \mathcal{U}_t} r_{t,k}. \tag{30}$$

Unselected entries contribute zero. Applying the Lipschitz inequality to selected log-odds increments and summing over false-positive selections proves the bound. Drift thus depends on the relevance mass admitted to unrelated hypotheses. Verification and proposition replacement are outside ordinary revision. Tests use the implemented cap $L = 6$.

Residual identity with an imperfect transition. Let $\lambda = \lambda_{\text{cons}} \geq 0$. Write $f_t = f(B_t)$, $u_t = f_t + \xi_t$, and $\widehat{G}_t = f_t - f_{t-1} + \delta_t$. Use $\widehat{G}_t$ as the supported transition in Equation 27, and define $\alpha = \lambda/(1+\lambda)$ and $\varepsilon_t = \tilde{a}_t - f_t$. Direct substitution yields

$$\varepsilon_t = \alpha\varepsilon_{t-1} + (1-\alpha)\xi_t + \alpha\delta_t. \tag{31}$$

This holds for arbitrary reference errors, including state-dependent ones. The simulator initializes $\tilde{a}_{-1} = 0$, so $\varepsilon_{-1} = -f(B_{-1})$; setting this residual to zero instead would change the evaluated controller.

For the moment calculation, condition on a fixed belief trajectory. Assume the proposal errors are independent and identically distributed over time with mean zero and variance σ^2, and the transition errors are independent and identically distributed with mean μ and variance τ^2. The two sequences are mutually independent and independent of the initial residual. Iterating Equation 31 for $t+1$ steps gives

$$\mathbb{E}[\varepsilon_t] = \alpha^{t+1}\varepsilon_{-1} + \lambda\mu(1-\alpha^{t+1}), \tag{32}$$

$$\text{Var}(\varepsilon_t) = \frac{\sigma^2 + \lambda^2\tau^2}{1+2\lambda}(1-\alpha^{2(t+1)}). \tag{33}$$

The variance follows from summing the geometric series of independent innovation variances. Taking the limit and adding squared bias yields

$$\text{MSE}_\infty = \lambda^2\mu^2 + \frac{\sigma^2 + \lambda^2\tau^2}{1+2\lambda}. \tag{34}$$

For $\lambda > 0$, comparison with the noisy direct readout's residual MSE σ^2 gives the exact stationary improvement criterion

$$\lambda\tau^2 + \lambda(1+2\lambda)\mu^2 < 2\sigma^2. \tag{35}$$

At the evaluated $\lambda = 2$, this becomes $\tau^2 + 5\mu^2 < \sigma^2$ for stationary residual-MSE improvement. With an exact reference, the stationary variance is $\sigma^2/(1+2\lambda)$. A constant reference bias is amplified to $\lambda\mu$. Multiplicative errors in $\widehat{G}_t$ are generally state-dependent and do not satisfy the i.i.d. assumptions; they are evaluated empirically rather than assigned this moment formula.

Bounded errors and utility. If $|\xi_t| \leq s$ and $|\delta_t| \leq b$, iterating the absolute-value inequality in Equation 31 instead yields

$$|\varepsilon_t| \leq \alpha^{t+1}|\varepsilon_{-1}| + (s+\lambda b)(1-\alpha^{t+1}). \tag{36}$$

Clipping to $[-1,1]$ is non-expansive. Because the belief-driven reference f_t can itself be incorrect, lower residual error alone does not guarantee lower oracle regret. We report full-episode utility, including initialization, alongside residual MSE. Algebra and bound checks are in `tests/test_residual_analysis.py`.

G Matched Local-Revision and Reference Controls

These controls separate update locality, action-readout denoising, and reference quality. After inspecting the main results, we froze all 17 post-hoc conditions, seed namespaces, source hashes, and analysis settings before execution with `scripts/run_mechanism_controls.py`. Table 8 retains every condition; the primary configuration is unchanged.

Paired replay. Seed groups 20,000–20,009 contain 80 episodes each, with seed $100{,}000g + i$ for $i = 0, \ldots, 79$, disjoint from main and validation runs. The 48-step generator is unchanged. Readout controls reuse cached sparse beliefs and identical standardized action-noise draws; transition noise uses a separate stream. Proposal-noise SD is zero or 0.045. Controllers share the known action map, clipping to $[-1,1]$, and initial desired action zero. We compare direct readout, exact-reference consistency at $\lambda = 2$, its equivalent residual filter, and action EMA with matched persistence $2/3$ (the main smoothing control uses 0.72). A dense-update arm changes only the locality gate under the same exact-reference controller and noise scale.

Table 8: All 17 frozen controller conditions on 800 additional paired episodes. Return covers the full episode; residual MSE is relative to the belief readout $f(B_t)$ rather than the oracle. Matched groups separate proposal noise, reference quality, and update scope.

Condition	Return	Residual MSE	MSE, steps 12–47
No proposal noise; identical sparse beliefs			
Direct readout	0.841160	0.000000	0.00000000
Exact-reference consistency	0.841280	0.001514	0.00000012
Ordinary residual filter	0.841280	0.001514	0.00000012
Matched action EMA	0.836540	0.005760	0.00428724
Proposal-noise SD 0.045; identical sparse beliefs			
Direct readout	0.839830	0.002015	0.00201332
Exact-reference consistency	0.843120	0.001902	0.00039686
Ordinary residual filter	0.843120	0.001902	0.00039686
Matched action EMA	0.836380	0.006121	0.00465190
$\widehat{G} = 0.5\Delta f$	0.840620	0.002915	0.00145222
$\widehat{G} = 1.5\Delta f$	0.840410	0.003082	0.00148581
Additive reference-error SD 0.01	0.842930	0.001979	0.00047329
Additive reference-error SD 0.03	0.842240	0.002594	0.00109072
Additive reference-error SD 0.045	0.841510	0.003459	0.00196030
Additive reference-error SD 0.06	0.840520	0.004670	0.00317832
Reference bias -0.01	0.842520	0.002277	0.00080899
Reference bias $+0.01$	0.842470	0.002275	0.00078405
State-scope control; proposal-noise SD 0.045			
Dense update, exact-reference consistency	0.837480	0.001902	0.00039686

Reference perturbations and statistics. With λ fixed, we scale the supported transition by 0.5 or 1.5, add independent Gaussian error with SD 0.01, 0.03, 0.045, or 0.06, or add bias ± 0.01. Return includes all 48 steps and initialization. Residual MSE is relative to $f(B_t)$, with a prespecified post-transient window at zero-based steps 12–47. The artifact retains all 13,600 episode-condition records, state/action/terminal trajectories, group means, and paired contrasts. Intervals use a Student-t multiplier with nine degrees of freedom over ten paired group means; they are pointwise descriptive intervals without multiplicity adjustment.

Incremental effects under matched conditions. At proposal-noise SD 0.045, sparse revision yields return 0.843120 versus 0.837480 for dense revision, a 0.005640 gain under the same controller. On identical sparse beliefs, exact-reference consistency yields 0.843120 versus 0.839830 for direct readout and 0.836380 for matched action EMA. Its post-transient residual MSE is 0.00039686 versus 0.00201332 for direct readout. These comparisons separate the effect of restricting revision scope from filtering action-proposal noise.

Reference quality and initialization. Without proposal noise, the exact controller retains only the decaying initialization residual $-\alpha^{t+1} f(B_{-1})$; return is 0.841280 versus 0.841160 for direct readout. Reference-error SD 0.06 instead raises post-transient MSE to 0.00317832, exceeding direct readout. Scaling the transition by 1.5 gives return 0.840410 and MSE 0.00148581: return remains above direct readout while MSE is worse than with the exact reference. Thus reference quality determines the denoising benefit, while task utility requires its own evaluation.

Implementation-equivalence checks. An independent array implementation uses the same threshold, active-only top-one selection, source/novelty scaling, clipping, reveal handling, terminal exclusion, and tie-breaking, without calling the thesis updater. All 800 confidence/action trajectories and terminal statuses match exactly; the equivalent residual controller differs by at most 2.22×10^{-16}. Separate evaluator smoke checks are retained in the artifact. These are implementation checks of the same rules, distinct from alternative-mechanism comparisons. Proposition and provenance metadata provide audit records; an additional predictive contribution from those fields is not established by these numerical controls.

Table 9: Same-evidence financial controls for January 2023–July 2026, with frozen features, weights, costs, and a common top-20% signal projection. The independent accumulator shares the persistent update's parameters and is an implementation-equivalence check. Best: dark gray and bold; second-best: light gray and underlined; ties share formatting.

Control	RankIC ↑	Cumulative net return ↑	Sharpe ↑	Max drawdown ↓	Turnover ↓
Equal Weight	–	75.70%	1.178	**11.04%**	**3.05%**
Frozen persistent state	**0.0393**	**107.62%**	**1.362**	13.10%	7.52%
Memoryless, same evidence	-0.0048	61.80%	0.872	15.20%	39.80%
Fixed six-month EMA	0.0186	82.40%	1.214	12.05%	15.30%
Independent numerical accumulator	**0.0393**	**107.62%**	**1.362**	13.10%	7.52%

H Same-Evidence Financial State Controls

This post-hoc diagnostic compares state retention with frozen features, weights, costs, and a common top-20% projection for all signal controls. The January 2023–July 2026 window and reference method had already been inspected; control settings and source hashes were frozen before execution. Results are exploratory comparisons on that same market path.

Matched inputs and decoder. For each industry and signal month, let x_t contain the frozen six-month momentum, lagged-return surprise, and distance from the trailing 12-month price peak features. With the existing weights $w = (0.55, 0.10, 0.35)$, define

$$u_{t,k} = 0.30 \tanh(x_{t,k}/1.50), \qquad s_t = \sum_k w_k[2\sigma(z_{t,k}) - 1]. \tag{37}$$

All controls share this increment and score decoder. They differ only in how they retain the numerical state:

$$z_t^{\text{reset}} = u_t, \tag{38}$$

$$z_t^{\text{EMA}} = d z_{t-1}^{\text{EMA}} + (1-d)u_t, \qquad d = 2^{-1/6}, \tag{39}$$

$$z_t^{\text{acc}} = \text{clip}_{[-2.5,2.5]}(z_{t-1}^{\text{acc}} + u_t). \tag{40}$$

Each coordinate starts at zero at the first complete historical observation for that industry. Ineligible observations do not update state. Historical features supply causal warm-up; risky holdings are reset to zero at the reporting boundary. The EMA half-life is fixed, not selected by performance. It changes both forgetting and state amplitude relative to the accumulator, so it is an alternative filter rather than an isolated intervention on forgetting alone.

All signal strategies hold the ten highest-scoring industries with equal weights out of the same universe of 49. Equal Weight holds all 49 as the common benchmark. Costs are 0.001 times half-ℓ_1 turnover from return-drifted holdings. Each path has exactly 43 monthly returns and initial turnover 0.5.

State-retention effects. Under matched inputs and projection, persistent accumulation has higher return and Sharpe and lower turnover than both monthly reset and the fixed six-month EMA (Table 9). Its drawdown is lower than reset but higher than EMA. These results support retained signal accumulation relative to the two specified alternatives on the inspected window.

Implementation-equivalence check. The script `scripts/run_finance_state_controls.py` exactly reproduces the frozen top-20% gross/net return, turnover, and RankIC paths. The persistent path has the same 13.10% drawdown, displayed as 13.1% in Table 7. The independent accumulator matches historical scores within 3.25×10^{-15} and the selected net-return path exactly; saved hypothesis confidences also decode to the frozen scores. Verification carry is one and verification strength is zero, so verification updates audit records without changing the score trajectory. The equivalence row validates the numerical implementation rather than an additional predictive effect of symbolic provenance. All controls and monthly paths are retained, with tests for causal input perturbations, per-industry isolation, and missing-input handling. No parameters were retuned and no new LLM calls were used for these controls.