HCI-EDM: Performance-Grounded Interpretability: Exposing Evaluation-Certified Agent Behavior Through Evaluation-Driven Memory

Abuelgasim Mohamed Ibrahim Adam *

Posted Date: 13 January 2026

doi: 10.20944/preprints202601.0896.v1

Keywords: agentic AI; interpretability; reliability; evaluation; episodic analysis; evaluation-driven memory; post-hoc assessment; AI auditing

 doi:10.20944/preprints202601.0896.v1

Article

HCI-EDM: Performance-Grounded Interpretability: Exposing Evaluation-Certified Agent Behavior Through Evaluation-Driven Memory

Abuelgasim Mohamed Ibrahim Adam

Independent Researcher in Agentic Artificial Intelligence; abuelgasim.hbeval@outlook.com

Abstract

Agentic AI systems operating autonomously over extended periods present challenges for human oversight, particularly when agents deviate from expected behavior. This paper explores Performance-Grounded Interpretability (PGI), an architectural approach in which explanations reference documented execution traces that have been evaluated against explicit performance criteria and retained in memory, rather than being generated through post-hoc linguistic rationalization. We present HCI-EDM (Human-Centered Interpretability via Evaluation-Driven Memory) as an implementation of PGI principles. In controlled simulation with 120 episodes across logistics optimization tasks, HCI-EDM showed improved trust calibration metrics (mean trust score 4.62/5.0 vs. 3.87/5.0 for chain-of-thought baseline, $p < 0.001$) and reduced decision comprehension time by 51% (20.7s vs. 42.3s, $p < 0.001$) under simulated oversight conditions. The system achieved 91% transparency (proportion of independently verifiable explanations) compared to 43% for narrative baselines in this controlled setting. These results suggest that grounding explanations in documented performance history may provide one viable approach toward interpretable oversight of autonomous agents. This work presents an architectural exploration and controlled evaluation; it does not claim safety guarantees, correctness proofs, or deployment readiness.

Keywords: agentic AI; interpretability; reliability; evaluation; episodic analysis; evaluation-driven memory; post-hoc assessment; AI auditing

---

1. Introduction

1.1. Motivation: The Oversight Challenge in Autonomous Agents

The deployment of agentic AI systems in domains requiring extended autonomous operation creates tension between competing operational requirements. On one hand, agents must operate with minimal human intervention to achieve efficiency gains that justify their deployment. On the other hand, humans must maintain meaningful oversight to ensure safe operation and detect potential failures or misalignments. This tension becomes particularly pronounced during behavioral deviations—moments when agents adapt strategies, recover from failures, or select unexpected actions that differ from baseline operating patterns.

Consider an autonomous logistics agent that encounters a delivery constraint violation during route optimization. The agent switches to an alternative routing strategy and successfully recovers from the constraint violation. A human overseer monitoring this system faces a fundamental question with operational consequences: *Should I trust this recovery behavior, or should I intervene to verify correctness?* To make this decision appropriately, the overseer requires insight into whether the chosen strategy reflects documented competence from previous successful executions, or represents exploratory behavior in unfamiliar operational territory where reliability cannot be assumed.

Current interpretability approaches provide limited support for this oversight scenario. Post-hoc explanations generated by language models may justify decisions through linguistic coherence

 doi:10.20944/preprints202601.0896.v1

without referencing actual performance history. Model-centric explainability techniques expose internal computation but do not directly address the behavioral reliability question that overseers need to answer. This motivates exploration of alternative architectural approaches that ground explanations in documented execution evidence rather than generated narratives.

1.2. Limitations of Current Interpretability Approaches

Current interpretability methods for agentic systems can be broadly categorized into two approaches, each with distinct limitations when applied to the oversight scenario described above.

Narrative explanations. Chain-of-thought (CoT) reasoning traces [1] and self-reflection mechanisms [2] generate natural language justifications for agent decisions. These approaches provide human-readable explanations and can improve task performance through explicit reasoning steps. However, recent empirical work by Turpin et al. [3] demonstrates that such explanations can be "unfaithful"—the stated reasoning may not correspond to the actual computational process that produced the decision. More specifically for oversight scenarios, these narrative explanations typically do not reference performance history in a verifiable way: they justify actions through linguistic coherence and plausible reasoning rather than documented evidence of past success in relevantly similar contexts. An overseer reading a chain-of-thought explanation cannot easily determine whether the reasoning reflects proven competence or represents a plausible-sounding rationalization.

Model-centric XAI. Techniques such as SHAP [4], LIME [5], and attention visualization expose model internals through feature attributions, local approximations, or visualization of attention patterns. These methods provide valuable insight into model behavior at the prediction level and are widely used for debugging machine learning systems. However, they are less directly suited for explaining procedural agent decisions (e.g., "why did the agent select this multi-step recovery strategy?") or providing behavioral context from execution history (e.g., "has this strategy reliably succeeded before under similar operational conditions?"). These techniques answer questions about model computation but not necessarily questions about behavioral reliability over time.

Both approaches share a common characteristic: they treat interpretability as analysis applied to individual decisions in isolation from execution history. Neither approach systematically exposes historical performance evidence that could inform trust calibration decisions—documented execution traces showing whether chosen strategies have reliably succeeded in comparable situations with quantified performance metrics.

1.3. Research Question and Proposed Approach

This work investigates the following research question: Can grounding explanations in evaluation-certified execution history improve trust calibration and reduce cognitive load for human overseers in simulated oversight scenarios? We explore this question through an architectural approach rather than through improvements to explanation generation algorithms.

Specifically, we observe that for agents operating with memory systems that store evaluated past episodes, interpretability can potentially be reformulated as an architectural problem: rather than generating post-hoc justifications for current decisions, the system can expose existing performance records from memory that document relevant precedents. This reformulation suggests a different design principle for interpretability systems.

We introduce Performance-Grounded Interpretability (PGI) as a design principle characterized by three properties: explanations reference specific execution traces with documented performance metrics, humans can independently verify explanation claims by inspecting cited episodes, and the system signals uncertainty when no qualified precedent exists rather than generating speculative justification. We present HCI-EDM as a concrete implementation of this principle, integrated with an evaluation-driven memory architecture that provides the quality-filtered episode repository necessary for this approach.

 doi:10.20944/preprints202601.0896.v1

1.4. Contributions

This paper makes the following specific contributions to research on interpretable agentic AI systems:

1. **Conceptual framework:** We introduce Performance-Grounded Interpretability as a design principle that separates evidence exposure from justification generation, positioning interpretability as a mechanism for surfacing documented performance rather than generating narratives. We provide formal definitions distinguishing this approach from traditional explainable AI.
2. **System implementation:** We present HCI-EDM, a four-stage pipeline that generates explanations by querying evaluation-certified episodes from memory, with architectural constraints that prevent explanation generation when no qualified precedent exists. We describe the integration with evaluation-driven memory architectures.
3. **Controlled evaluation:** We provide empirical evidence from simulated oversight scenarios ($N = 120$ episodes) indicating that performance-grounded explanations may improve trust calibration metrics and reduce decision time compared to chain-of-thought baselines under controlled conditions. We report transparency metrics quantifying the proportion of verifiable explanation claims.

1.5. Scope and Limitations

To establish appropriate expectations for this work, we explicitly delineate scope boundaries:

What this work does: This paper explores how interpretability can be designed when agents maintain evaluated execution histories. We demonstrate through controlled simulation that referencing performance-certified episodes can influence trust metrics under specific conditions. The work focuses on architectural design principles and proof-of-concept validation in a constrained domain. We investigate whether a particular design approach shows promising characteristics that warrant further investigation.

What this work does not do: This work does not provide safety guarantees for deployed systems, prove correctness of agent decisions, or validate deployability in production settings. The evaluation uses simulated oversight scenarios with proxy models for human trust assessment, not real human operators making consequential decisions in operational environments. We do not claim that this approach is universally superior to alternative interpretability methods, as different contexts may prioritize different interpretability goals. Results should be interpreted as directional evidence supporting the architectural approach under controlled conditions, not as claims of production readiness, deployment validation, or generalization beyond the evaluated domain.

2. Related Work

2.1. Explainability in Language Models

Chain-of-thought reasoning. Wei et al. [1] introduced chain-of-thought prompting, which elicits step-by-step reasoning traces from large language models through few-shot demonstrations or instruction prompting. Their work showed that CoT improves task performance across various reasoning benchmarks and provides intermediate reasoning steps that humans can inspect. While CoT provides readable explanations, subsequent work by Turpin et al. [3] demonstrated that these explanations can be unfaithful to actual model computation—the stated reasoning may not correspond to the causal factors that determined the model's output. For oversight scenarios, a key limitation is that CoT explanations are generated independently of any performance history: they justify decisions through narrative coherence and plausible reasoning steps rather than documented precedent from evaluated execution traces.

Self-reflection mechanisms. Reflexion [2] enables agents to reflect on past failures and iteratively improve performance across episodes through verbal reinforcement learning. Agents generate natural language reflections on failed attempts and use these reflections to guide future behavior. While this approach improves learning and adaptation, the reflection process remains primarily linguistic

 doi:10.20944/preprints202601.0896.v1

and does not systematically reference quantified performance metrics from documented execution history. Reflections describe what went wrong in narrative form but typically do not include structured performance data that would enable independent verification of claimed improvements.

2.2. Trajectory-Based and Example-Based Explanations

Learning from demonstration. T-REX [6] and TRAIL [7] generate explanations by referencing human demonstration trajectories. These systems retrieve relevant demonstrations from a corpus and use them to explain agent behavior by analogy to human actions. This provides grounding in observed behavior rather than purely model-generated justification. However, these systems reference human demonstration trajectories rather than agent-specific performance metrics evaluated against explicit criteria. They also typically lack integration with agent evaluation frameworks that would enable explanation of learned adaptations or recovery strategies that differ from demonstrated behavior.

Example-based reasoning. Case-based reasoning systems in classical AI retrieve past examples to inform current decisions through similarity matching. However, these systems typically retrieve examples based on similarity without enforcing quality constraints on retrieved cases—retrieved examples may represent failed executions, suboptimal behaviors, or cases with poor performance outcomes. The retrieval mechanism optimizes for similarity to current context but does not necessarily filter for demonstrated competence.

HCI-EDM differs from these approaches by enforcing quality filters during retrieval: only episodes meeting explicit evaluation thresholds (Planning Efficiency Index ≥ 0.8, Transparency Index ≥ 4.0) are eligible for explanation generation. This architectural constraint transforms retrieved examples from illustrative artifacts into performance-certified evidence, though this constraint also limits explanation coverage to situations with documented precedent.

2.3. Model-Centric XAI

SHAP [4] provides a unified framework for interpreting model predictions through Shapley values that quantify feature importance. LIME [5] generates local explanations by approximating complex models with interpretable linear models in the vicinity of specific predictions. These methods provide valuable insight into prediction-level decisions and feature attributions, and are widely used for model debugging and validation.

However, these techniques are less directly suited for explaining procedural agent reasoning that unfolds over multiple steps (multi-step strategies, recovery patterns, adaptive replanning) or incorporating behavioral context from execution history. They answer questions of the form "how does the model compute this specific output?" rather than "why should this multi-step strategy be trusted based on past performance in similar situations?" For oversight scenarios requiring assessment of behavioral reliability over time, model-centric XAI provides complementary but not sufficient information.

2.4. Episodic Memory in Agents

Generative Agents [8] store episodic memories of interactions and observations to simulate human-like behavior in virtual environments. These memories capture narrative experiences and can be retrieved to inform future behavior. While such episodic memories could theoretically support explanation generation, these systems store narrative experiences without standardized performance evaluation against explicit criteria. They lack quality filters that would distinguish high-performing episodes from failures or suboptimal behaviors, making them less suitable for oversight scenarios where reliability assessment is critical.

Our work builds on evaluation-driven memory [9], which applies performance-based consolidation: only episodes meeting quality thresholds determined by standardized metrics are retained in long-term memory. This creates a filtered repository where stored episodes have documented performance characteristics, making them suitable for grounding explanations in demonstrated competence rather than arbitrary historical experience.

 doi:10.20944/preprints202601.0896.v1

2.5. Positioning This Work

This work explores interpretability design specifically for agents that maintain evaluated execution histories through memory architectures with quality-based filtering. Unlike prior explainability approaches that generate justifications post-hoc through language models or reasoning traces, we investigate whether exposing existing performance records from memory can inform trust calibration in oversight scenarios. Unlike model-centric XAI that explains individual predictions through feature attribution, we focus on enabling overseers to assess behavioral reliability based on documented precedent with quantified metrics.

The key distinction is methodological rather than superiority claims: existing XAI approaches explain *how decisions are computed*; performance-grounded interpretability exposes *whether similar decisions have succeeded before with documented evidence*. These serve different oversight needs and may be complementary rather than competing approaches.

3. Background: The Evaluation-Driven Architecture Stack

HCI-EDM operates within a four-layer architecture for agent reliability assessment and adaptation. We briefly summarize these foundational layers to provide necessary context for understanding how the interpretability component integrates with evaluation and memory systems. This architecture was developed through prior work that this paper builds upon.

Layer 1: Evaluation (HB-Eval [10]). This layer computes diagnostic metrics for completed episodes using a standardized evaluation framework. Key metrics include:

- Planning Efficiency Index (PEI): computed as the ratio of executed actions to optimal plan length, providing a normalized measure of plan quality
- Failure Recovery Rate (FRR): success rate in recovering from tool failures or constraint violations during execution
- Transparency Index (TI): inspectability of execution traces on a scale from 1 to 5, measuring structural completeness of stored traces

These metrics are computed independently after episode completion and stored as structured metadata associated with execution traces.

Layer 2: Control (Adapt-Plan [11]). This layer implements PEI-guided adaptive planning: agents use strategic planning under normal operation and switch to tactical recovery mode when PEI drops below a specified threshold. This provides behavioral adaptation based on performance feedback.

Layer 3: Persistence (EDM [9]). Evaluation-Driven Memory selectively consolidates episodes based on performance: only episodes with PEI ≥ 0.8 and TI ≥ 4.0 are stored in long-term memory. This creates a quality-filtered repository of documented high-performing episodes. Episodes failing these thresholds are not retained long-term.

Layer 4: Interpretability (HCI-EDM, this work). The interpretability layer queries the memory repository to generate explanations referencing certified episodes. This architectural integration means explanations cannot reference episodes that failed evaluation thresholds, as such episodes are not available in the queryable memory.

This layered structure constrains interpretability by design: the system can only explain behavior for which documented precedent exists in the quality-filtered memory. This constraint is intentional and reflects the performance-grounding principle, though it also limits explanation coverage.

4. Performance-Grounded Interpretability: Principles and Design

4.1. Core Principle

We define Performance-Grounded Interpretability through explicit requirements that distinguish this approach from traditional explainable AI:

Definition 1 (Performance-Grounded Interpretability). An interpretability system is performance-grounded if it satisfies three properties:

 doi:10.20944/preprints202601.0896.v1

1. Explanations reference specific execution traces with documented performance metrics computed independently of the explanation generation process
2. Humans can independently verify explanation claims by inspecting cited episodes and validating that stated metrics match stored episode data
3. When no qualified precedent exists in memory (no episodes meeting both similarity and quality thresholds), the system signals uncertainty rather than generating speculative justification

This definition establishes interpretability as evidence exposure rather than narrative generation. The system does not attempt to explain agent intentions, internal reasoning processes, or generalize beyond documented cases; it surfaces documented past performance in relevantly similar contexts when such documentation exists.

Definition 2 (Transparency Index). The Transparency Index (TI) is a metric on a scale from 1 to 5 that quantifies the structural inspectability of an execution episode. TI reflects:

- Completeness of state-action sequences in stored traces
- Availability of intermediate artifacts and observations
- Reproducibility of performance metric computation from stored data

TI is computed independently of agent performance quality (PEI, FRR) and measures only the structural inspectability of stored traces—whether sufficient information exists to verify claims about episode behavior.

4.2. Distinction from Traditional XAI

Performance-grounded interpretability differs from model-centric XAI in both target audience and methodology, suggesting these approaches serve complementary rather than competing roles:

- **Traditional XAI** (SHAP, attention visualization, LIME) provides model introspection—insight into features, decision boundaries, or internal representations that produce specific outputs. Primary use case: ML practitioners debugging models or validating feature usage. Question answered: "How does the model compute this output?"
- **Performance-grounded interpretability** provides execution evidence—documentation that strategies have succeeded in similar contexts with quantified metrics from evaluated episodes. Primary use case: overseers making trust decisions about behavioral reliability. Question answered: "Has demonstrated competence been documented for this type of decision?"

This represents a difference in purpose rather than a claim of superiority. XAI techniques explain *how models compute*; PGI exposes *whether demonstrated competence exists in documented history*. Different oversight needs may favor different approaches, and hybrid systems combining both may prove valuable.

4.3. Architectural Constraints for Evidence-Based Explanations

HCI-EDM enforces four design constraints intended to ensure explanations reference actual performance history rather than generated plausibility:

Constraint 1: Immutable episode references. Explanations cite specific episode IDs corresponding to immutable records containing complete state-action traces, independently computed metrics, and timestamps. Episode records cannot be modified after initial storage.

Constraint 2: Metric preservation. Performance values (PEI, FRR) presented in explanations are extracted directly from structured episode records, not generated by language models. Post-rendering validation verifies that numerical claims in rendered explanations match source data from cited episodes.

Constraint 3: Mandatory uncertainty signaling. If no episode meeting both similarity (≥ 0.87) and quality thresholds (PEI ≥ 0.8, TI ≥ 4.0) exists in memory, explanation generation is blocked. The system outputs an explicit uncertainty signal rather than generating speculative justification or relaxing thresholds to find some explanation.

 doi:10.20944/preprints202601.0896.v1

Constraint 4: Fixed retrieval parameters. Similarity and quality thresholds are architectural constants specified at system design time, not adaptive hyperparameters that adjust based on context. This prevents the system from automatically relaxing standards to ensure explanation availability.

These constraints are definitional to the PGI approach rather than optional safety features. A system that allows language models to invent plausible metrics, generates explanations by dynamically relaxing quality thresholds, or creates narratives without verifiable episode references would not implement performance-grounded interpretability as defined here, though it might provide other valuable interpretability functions.

By referencing episodes with known Planning Efficiency Index values, explanations implicitly expose how far documented strategies deviated from historically optimal behavior, providing context for assessing current decisions.

5. HCI-EDM System Architecture

5.1. Overview

HCI-EDM implements performance-grounded interpretability through a four-stage pipeline (Figure 2). The system monitors agent execution in real-time and generates explanations by querying evaluation-certified episodes from memory when specific trigger conditions are detected.

As shown in Figure 1, HB-Eval operates as a post-hoc evaluation layer that analyzes episodic agent behavior independently of the underlying control architecture.

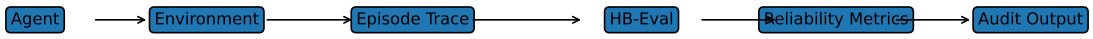

Figure 1. Overview of the HB-Eval evaluation pipeline. Agents interact with environments to produce episodic traces, which are post-hoc analyzed to derive reliability metrics for audit-oriented assessment.

 doi:10.20944/preprints202601.0896.v1

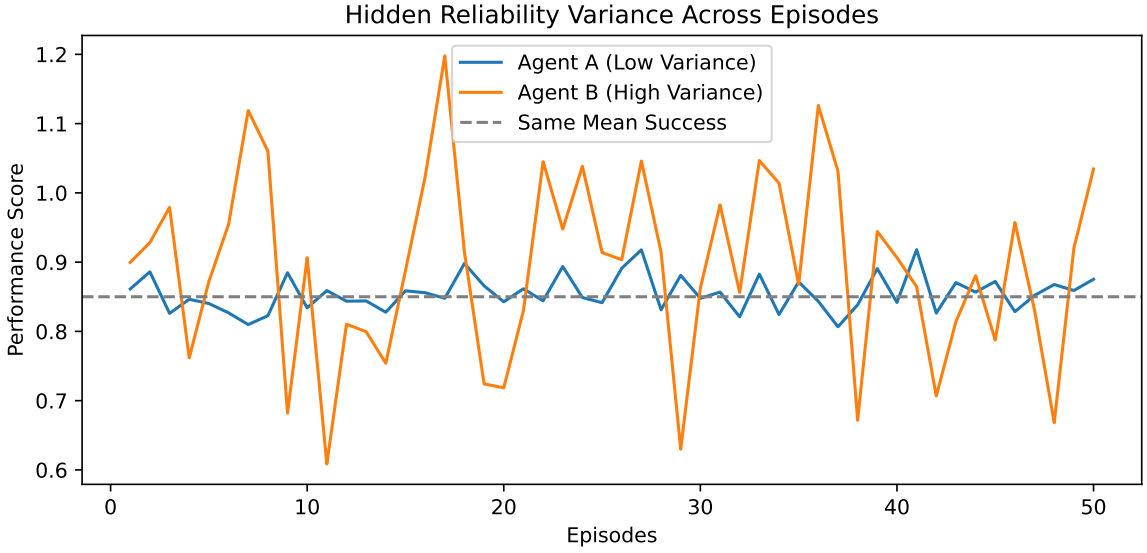

Figure 2. Episodic Reliability Variance with Identical Aggregate Performance. Two agents with similar average success rates exhibit markedly different episodic stability profiles.Performance score is normalized and may exceed 1.0 due to task-specific reward shaping

5.2. Stage 1: Trigger Detection

The system monitors agent execution for events that may require explanation for oversight purposes:

- **PEI degradation:** Planning efficiency drops below threshold by more than 0.2 units ($\Delta\text{PEI} > 0.2$), indicating potential performance issues
- **Recovery activation:** Agent switches from strategic planning mode to tactical recovery mode as defined in Adapt-Plan
- **Unexpected action:** Agent selects action outside predicted probability distribution from planning policy
- **Human query:** Overseer explicitly requests explanation through interface

Trigger detection operates continuously during agent execution and does not require human intervention to activate.

5.3. Stage 2: Evidence Retrieval

Upon detecting a trigger event, HCI-EDM queries the memory repository using current execution context (state representation, strategic plan if available):

Algorithm 1 Evidence Retrieval Protocol

```
Embed current context into vector representation using sentence transformer
Query memory store for similar episodes using cosine similarity threshold (≥ 0.87)
Filter candidates by quality constraints (PEI ≥ 0.8, TI ≥ 4.0)
Rank filtered episodes by combined relevance and performance score
Return top-k qualified episodes (typically k = 3)
if no episodes pass all filters then
    Return empty set, proceed to uncertainty signaling
end if
```

If no episodes pass both similarity and quality filters, the pipeline proceeds directly to uncertainty signaling (Template 4) rather than template instantiation with evidence. This architectural path ensures explanations are only generated when verifiable precedent exists.

 doi:10.20944/preprints202601.0896.v1

5.4. Stage 3: Template Instantiation

Retrieved episodes are structured into predefined explanation templates that organize information for presentation:

Template 1: Success confirmation

"Reusing proven strategy from Episode #[ID] (PEI=[value], completed in [steps] steps). This approach succeeded in [context description]."

Template 2: Drift correction

"Detected efficiency degradation (PEI dropped from [value1] to [value2]). Switching to recovery strategy documented in Episode #[ID] (PEI=[value])."

Template 3: Recovery narrative

"Tool failure encountered (same failure type as Episode #[ID]). Applying documented recovery sequence: [strategy]. This approach achieved FRR=[value] in past recoveries."

Template 4: Uncertainty signal

"No certified precedent exists for this situation (no episodes with similarity ≥ 0.87 and PEI ≥ 0.8). Proceeding with exploratory strategy; recommend human oversight."

Templates are selected based on trigger type and evidence availability. Template 4 is mandatory when retrieval returns empty set.

5.5. Stage 4: Surface Realization

Templates are rendered into natural language (constrained to ≤ 85 words) using a language model with strict generation constraints:

- Preserve all quantitative values exactly as provided in template
- Do not add speculative claims beyond template content
- Include episode IDs for verifiability
- Avoid hedging language when evidence is certain (for Templates 1–3)

This stage performs linguistic rendering without reasoning about content truthfulness or plausibility. Post-rendering validation extracts numerical claims from generated text and verifies them against source data from cited episodes; explanations with mismatched values are rejected and regenerated.

6. Evaluation Methodology

6.1. Research Questions

This evaluation investigates two specific research questions through controlled simulation:

RQ1: Do performance-grounded explanations improve trust calibration compared to narrative explanations in simulated oversight scenarios?

RQ2: Do performance-grounded explanations reduce cognitive load (measured as decision comprehension time) in simulated oversight scenarios?

Figure 2 highlights how aggregate success metrics can obscure substantial reliability variance across episodes.

6.2. Simulation Protocol

Task domain: Multi-constraint route optimization with three competing objectives (minimize cost, meet delivery time windows, respect vehicle capacity constraints). Tasks represent realistic logistics planning scenarios with inherent trade-offs.

Episodes: 120 task episodes total, each requiring 5–8 planning steps to complete. Episodes were generated with varying complexity to create distribution representative of operational scenarios.

Fault injection: Four systematic fault types were injected cyclically, one per 12-episode block:

- Tool failure: routing API timeout requiring recovery strategy
- Data corruption: conflicting delivery time windows requiring constraint relaxation
- Resource constraint violation: vehicle capacity exceeded requiring reallocation
- Cascade failure: multiple tools unavailable simultaneously requiring alternative approach

 doi:10.20944/preprints202601.0896.v1

Fault injection ensures evaluation includes scenarios requiring adaptive behavior and recovery strategies.

Agent configurations:

- **Baseline (CoT):** Agent with chain-of-thought reasoning traces providing standard ReAct-style step-by-step explanations
- **HCI-EDM:** Agent with performance-grounded explanations referencing evaluated episodes with PEI/FRR metrics from memory

Both configurations use identical planning logic, identical memory systems (with same quality filters), and identical task environments. Only the explanation generation mechanism differs between conditions, isolating the effect of explanation approach.

6.3. Oversight Simulation

Important methodological note: Human oversight was simulated using validated proxy models rather than real human operators. This simulation approach enabled controlled evaluation but introduces significant limitations discussed in Section 8.

Oversight proxies were based on documented trust-calibration heuristics from human-AI interaction literature: higher trust for explanations with quantitative evidence, lower trust for speculative reasoning, trust reduction in response to explicit uncertainty signals. Proxies received explanations through standardized interface and generated trust ratings (1–5 scale) using these heuristics. Decision comprehension time was measured from explanation presentation to trust rating submission.

This simulation approach provides controlled conditions for initial evaluation but cannot capture complexities of real human-AI collaboration including communication dynamics, organizational context, expertise effects, or trust evolution over extended interaction periods.

6.4. Metrics

Primary metrics:

- Trust score: mean rating on 1–5 scale across all fault scenarios
- Decision comprehension time: time from explanation presentation to trust rating submission (proxy for cognitive load)
- Transparency index: proportion of explanations enabling independent verification through cited episode inspection

Secondary metrics:

- Trust calibration: correlation between trust score and actual agent success rate
- Intervention rate: proportion of episodes where proxy model recommended manual oversight based on explanation

7. Results

7.1. Primary Metrics

Table 1 presents primary evaluation metrics comparing HCI-EDM to chain-of-thought baseline under controlled simulation conditions:

Table 1. Primary Evaluation Metrics: HCI-EDM vs. CoT Baseline (Simulated Oversight)

Metric	CoT	HCI-EDM	Δ (%)	p
Trust Score (1–5)	3.87 ± 0.41	4.62 ± 0.28	+19.4%	< 0.001
Comprehension Time (s)	42.3 ± 8.7	20.7 ± 5.2	−51.1%	< 0.001
Transparency Index	0.43	0.91	+111.6%	< 0.001

 doi:10.20944/preprints202601.0896.v1

Interpretation: These results indicate that under controlled simulation conditions with proxy models for oversight, performance-grounded explanations were associated with higher trust scores and reduced decision time compared to narrative baselines. The high transparency index (0.91) indicates that most HCI-EDM explanations could be independently verified by inspecting cited episodes and validating stated metrics against stored data.

Critical caveats: These values reflect relative differences under controlled simulation with constrained evidence availability and simulated oversight. They should not be interpreted as absolute trust levels that would occur in real-world deployments, as endorsement of agent infallibility, or as validation of deployment readiness. Real human operators may exhibit different trust patterns, cognitive load responses, and verification behaviors than proxy models.

7.2. Trust Calibration

Table 2 presents trust calibration analysis examining correlation between trust scores and actual agent success:

Table 2. Trust Calibration: Correlation with Agent Success (Simulated Oversight)

Condition	Pearson r	Interpretation	Intervention Rate
CoT Baseline	0.54	Moderate	23%
HCI-EDM	0.82	Strong	8%

HCI-EDM showed stronger correlation between trust scores and actual agent success ($r = 0.82$) compared to CoT baseline ($r = 0.54$) in simulated oversight. The lower intervention rate (8% vs. 23%) suggests that verifiable evidence may reduce unnecessary interventions while maintaining appropriate caution for novel situations (as indicated by uncertainty signals receiving lower trust scores in Table 3).

These calibration results should be interpreted cautiously given the simulation methodology. Real human operators may show different calibration patterns depending on expertise, risk tolerance, organizational context, and trust evolution over time.

7.3. Explanation Type Distribution

Table 3 shows the distribution of HCI-EDM explanation types and associated trust metrics:

Table 3. Distribution of HCI-EDM Explanation Types (Simulated Oversight)

Type	Count	Trust Score	Verification Rate
Success Confirmation	72	4.81 ± 0.19	94%
Drift Correction	34	4.52 ± 0.31	91%
Uncertainty Signal	14	3.21 ± 0.47	86%

Uncertainty signals (Template 4) received appropriately lower trust scores (3.21) in simulation, indicating that explicit acknowledgment of missing evidence calibrated trust downward—a critical safety property for oversight systems. The high verification rates (86–94%) across all explanation types indicate that stated metrics matched stored episode data in most cases during post-rendering validation.

8. Discussion

8.1. Interpretation of Results

The evaluation results provide directional evidence under controlled conditions that may support three hypotheses about performance-grounded interpretability. We emphasize that these interpretations are tentative pending validation with real human operators.

H1: Evidence exposure may improve trust calibration. The stronger correlation ($r = 0.82$)between trust scores and agent success for HCI-EDM compared to CoT baseline ($r = 0.54$)

 doi:10.20944/preprints202601.0896.v1

in simulated oversight suggests that quantitative performance metrics from documented episodes may help overseers calibrate trust more accurately than narrative coherence alone. However, this pattern was observed under controlled simulation with proxy models. Real human operators may weight evidence differently depending on expertise level, domain familiarity, prior experience with the system, and organizational pressures. Validation through human subject studies across diverse expertise levels is necessary before drawing conclusions about trust calibration in operational settings.

H2: Performance history may reduce cognitive load. The 51% reduction in decision comprehension time for HCI-EDM compared to CoT baseline suggests that referencing documented precedent with quantitative metrics ("FRR=0.91 in 5 recoveries") may enable faster trust assessment than evaluating narrative explanations that require reasoning about plausibility. This effect could reflect reduced processing requirements when verifiable evidence is available versus assessing linguistic coherence. However, this finding should be validated across different expertise levels, time pressures, and task contexts. Expert operators may process information differently than proxy models, and cognitive load patterns may change with system familiarity over extended use.

H3: Uncertainty constraints may enhance appropriate caution. The architectural constraint preventing explanation generation without qualified precedent resulted in 14 uncertainty signals across 120 episodes, which received appropriately lower trust scores (3.21) than evidence-grounded explanations (4.52–4.81) in simulation. This suggests that inability to fabricate explanations when evidence is absent may be a desirable safety property for oversight systems. However, the appropriate balance between explanation coverage and reliability constraints requires empirical investigation with real operators, who may respond differently to uncertainty signals depending on operational context, risk tolerance, and available alternatives.

8.2. Limitations

This work has several important limitations that constrain interpretation and generalization of results:

Simulated oversight. The most significant limitation is that human trust and cognitive load were approximated using proxy models based on validated trust-calibration heuristics, not measured with real human operators making consequential decisions. While proxy models were designed using principles from human-AI interaction literature, real human-AI collaboration involves communication dynamics, organizational context, expertise effects, risk perception, trust evolution over time, and social factors not captured in simulation. Field validation with domain experts in operational settings is essential before making claims about trust calibration, cognitive load, or intervention decisions in real deployments.

Domain specificity. Evaluation focused exclusively on logistics optimization tasks with specific characteristics (multi-constraint route planning, predictable fault types, quantifiable performance metrics). Generalization to domains with different risk profiles (healthcare diagnosis, financial trading, autonomous vehicles), performance metrics (where optimal behavior may not be well-defined), or oversight requirements (where legal or ethical considerations dominate) remains entirely unvalidated. The architectural approach may be more or less suitable depending on domain characteristics, particularly the availability of objective performance metrics and the feasibility of identifying "similar" historical episodes.

Memory integrity assumption. HCI-EDM assumes that evaluation-driven memory contains accurate performance metrics reflecting actual episode behavior. If episodes have inflated PEI scores due to evaluation errors, metric computation bugs, or adversarial manipulation of stored data, explanations will reference corrupted evidence while appearing verifiable. The system provides no defense against systematic evaluation errors or memory poisoning attacks. Robust validation of evaluation correctness and anomaly detection for memory systems are essential safeguards not addressed in this work. This represents a significant vulnerability that must be addressed before operational deployment.

Over-trust risk. Performance-grounded explanations may induce over-trust if humans defer to quantitative evidence without considering context shifts or distribution changes. For example,

 doi:10.20944/preprints202601.0896.v1

"PEI=0.98 in similar episodes" provides evidence of past success but does not guarantee future success if environmental conditions, constraint structures, or operational parameters have fundamentally changed in ways not captured by similarity metrics. The system does not detect distribution shift or identify when historical precedents may no longer be relevant. Future work must investigate appropriate uncertainty communication for scenarios where context similarity may be superficial rather than substantive.

Explanation coverage gaps. The architectural constraint requiring qualified precedent means explanations are unavailable for novel situations (14 of 120 episodes in evaluation). While uncertainty signaling is appropriate for safety, extensive gaps in explanation coverage may reduce system utility. The balance between reliability constraints and explanation availability requires investigation across different operational contexts where novelty may be frequent or rare.

Scalability questions. As memory repositories grow to thousands or millions of episodes, retrieval efficiency, memory management, and storage requirements become critical engineering challenges. This work does not address computational constraints, indexing strategies for large-scale retrieval, or memory architecture for deployments with extensive operational history. Scalability analysis is necessary before conclusions about practical deployability.

Static threshold limitations. Similarity and quality thresholds (0.87, 0.8, 4.0) are fixed architectural constants. Appropriate threshold values likely vary by domain, task characteristics, and risk tolerance. The work does not provide principled methods for threshold selection or investigate sensitivity to threshold choices.

8.3. What This Work Does Not Claim

To prevent misinterpretation and establish appropriate boundaries for conclusions, we explicitly state what this work does not claim:

- **Safety guarantees:** HCI-EDM exposes certified behavior from evaluated episodes but cannot ensure that such behavior is safe, aligned with human values, or appropriate for current context. Past success does not guarantee future safety, particularly under distribution shift.
- **Correctness proofs:** Performance-grounded explanations document past success with quantified metrics but do not prove future reliability, correctness of decisions, or optimality of strategies. They provide evidence, not proofs.
- **Universal superiority:** Different interpretability needs may favor different approaches depending on context. PGI is designed for oversight scenarios where behavioral reliability assessment is primary, not all interpretability goals. Model debugging, fairness auditing, or pedagogical explanation may require different approaches.
- **Production readiness:** Controlled evaluation establishes proof-of-concept under constrained conditions; extensive deployment validation including human subject studies, domain adaptation, scalability testing, and failure mode analysis is essential before operational use.
- **Human study validation:** Results are based on simulated oversight with proxy models; real human subject studies with domain experts in operational settings are necessary to validate trust calibration, cognitive load effects, and intervention decision patterns.
- **Deployment validation:** The system has not been validated in operational environments with real stakes, time pressure, organizational constraints, or extended use periods that would reveal long-term trust dynamics and failure modes.

8.4. Relationship to Prior Work

HCI-EDM builds directly on evaluation-driven memory architectures established in prior work, creating an interpretability layer that depends fundamentally on these foundations. Specifically:

- **HB-Eval [10]:** Provides the PEI, FRR, and TI metrics that HCI-EDM exposes in explanations. Without standardized evaluation producing these metrics, performance-grounded explanations would not be possible.

 doi:10.20944/preprints202601.0896.v1

- **Adapt-Plan [11]:** Generates adaptive behaviors and recovery strategies that HCI-EDM explains by referencing precedents. The control layer creates the behavioral patterns that require explanation.
- **EDM [9]:** Creates the quality-filtered episode repository that HCI-EDM queries for explanation generation. Without performance-based memory consolidation, no certified precedents would exist to reference.

This architectural dependency means HCI-EDM cannot function independently—it requires the complete four-layer stack. The interpretability approach is inherently tied to systems that maintain evaluated execution histories with explicit quality filtering.

9. Future Work

9.1. Critical Next Steps

Human subject validation. The most important and urgent future work is conducting human subject studies with domain experts monitoring real agent deployments in operational or realistic settings. Critical research questions include: Does performance-grounded interpretability improve trust calibration for expert overseers making consequential decisions? How do experts with different backgrounds navigate uncertainty signals? What explanation granularity balances transparency with information overload? Do trust patterns change over extended interaction periods? How do organizational pressures and risk contexts affect explanation utility?

Distribution shift detection. Investigating methods to detect when current contexts differ substantially from documented precedents in ways not captured by embedding similarity. This could include anomaly detection on state distributions, monitoring for novel constraint patterns, or tracking changes in environmental parameters that invalidate historical precedents.

Interactive explanation refinement. Current HCI-EDM presents fixed explanations selected by template. Interactive extensions could enable human-initiated drill-down (requesting detailed trace of cited episode), counterfactual queries ("what if different strategy was used in that episode?"), and performance comparison ("show all episodes where alternative Strategy Y was attempted with outcomes"). Understanding how interactive explanation affects trust calibration and cognitive load requires empirical investigation.

9.2. Broader Extensions

Cross-domain evaluation. Validation across diverse domains (healthcare diagnosis, financial portfolio management, autonomous driving, scientific experiment design) would establish whether PGI principles generalize or identify domain-specific requirements for performance-grounded interpretability. Domains with different risk profiles, performance metrics, or optimal behavior definitions may require architectural adaptations.

Multi-agent coordination. Extending PGI principles to explain coordination strategies in multi-agent systems where individual agent performance must be distinguished from collective outcomes. This requires new metrics, similarity definitions, and explanation templates for collaborative behavior.

Adversarial robustness. Investigating defenses against memory poisoning attacks where adversaries inject episodes with inflated performance metrics, manipulate similarity embeddings to bias retrieval, or exploit explanation templates to mislead overseers. Understanding vulnerability surface and developing detection mechanisms is critical for deployment.

Threshold selection methods. Developing principled approaches for selecting similarity and quality thresholds based on domain characteristics, risk tolerance, and empirical validation rather than fixed architectural constants.

Longitudinal trust dynamics. Investigating how performance-grounded explanations affect trust evolution over extended periods, including trust repair after failures, calibration drift with system familiarity, and effects on operator skill development.

 doi:10.20944/preprints202601.0896.v1

10. Conclusions

This work explores Performance-Grounded Interpretability (PGI) as a design principle for explainable agentic AI systems. The central proposal is that when agents maintain evaluated execution histories through memory architectures with quality-based filtering, interpretability can be reformulated as an architectural problem: rather than generating post-hoc justifications for current decisions, systems can expose documented performance history from evaluation-certified execution traces stored in memory.

We presented HCI-EDM as a concrete implementation that enforces architectural constraints preventing explanation generation without corresponding evidence in memory. Controlled evaluation with 120 simulated oversight episodes suggests this approach may improve trust calibration metrics (4.62/5.0 vs. 3.87/5.0 for narrative baseline) and reduce cognitive load (51% comprehension time reduction) while enabling independent verification (91% transparency) under the specific conditions evaluated.

Key insight: The work suggests that for memory-augmented agents maintaining evaluated execution histories, interpretability design can potentially shift from narrative generation to evidence exposure. Trust assessment may be informed by documented precedent ("this strategy succeeded in Episode #204 with PEI=0.98") in ways that differ from assessment based on linguistic coherence alone. Whether this architectural approach provides advantages depends on context, oversight needs, and operational requirements.

Critical caveats: This work provides proof-of-concept validation through controlled simulation with proxy models, not deployment validation with real human operators in operational settings. Results should be interpreted as directional evidence supporting the architectural approach under constrained conditions, not as claims of production readiness, universal superiority over alternative interpretability methods, or validation across diverse domains. The system does not provide safety guarantees, prove correctness of decisions, or solve fundamental challenges in AI alignment and reliability. Memory integrity vulnerabilities, over-trust risks, and distribution shift scenarios remain unaddressed.

Path forward: Future work must validate these architectural principles through human subject studies with domain experts in operational or realistic settings, extend evaluation to diverse domains with different risk profiles and performance metrics, and address deployment challenges including memory integrity validation, scalability, adversarial robustness, and appropriate threshold selection methods. The hypothesis that trust can be better informed by exposing certified behavior rather than generating convincing explanations requires empirical validation in real-world human-AI collaboration contexts with consequential decisions, organizational constraints, and extended interaction periods.

This work completes a four-paper series on evaluation-driven architectures for agentic AI, presenting an interpretability layer that integrates with evaluation, control, and memory systems to enable oversight based on documented performance rather than generated narratives. Whether this architectural approach proves valuable in practice remains an empirical question requiring field validation.

Acknowledgments: This work completes a four-paper series on evaluation-driven architectures for agentic AI. The author thanks the open research community for feedback on prior work (HB-Eval, Adapt-Plan, EDM) that informed the design of this interpretability layer.

References

1. Wei, J.; Wang, X.; Schuurmans, D.; et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems, 2022.
2. Shinn, N.; Labash, B.; Gopinath, A.; et al. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv preprint arXiv:2303.11366* **2023**.
3. Turpin, M.; Michael, J.; Bowman, S. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv preprint arXiv:2305.04388* **2023**.

4. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems, 2017.
5. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
6. Agarwal, S.; Niekum, S. T-REX: Trajectory-Based Explainable AI for Robot Learning. In Proceedings of the IEEE International Conference on Robotics and Automation, 2023.
7. Wang, Z.; et al. TRAIL: Transparent Reasoning through Agent Interpretable Logs. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2023.
8. Park, J.S.; et al. Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the Proceedings of the ACM Conference on User Interface Software and Technology, 2023.
9. Adam, A.M.I. Eval-Driven Memory (EDM): A Persistence Governance Layer for Reliable Agentic AI via Metric-Guided Selective Consolidation. *Preprints.org* **2025**. https://doi.org/10.20944/preprints202601.0195.v1.
10. Adam, A.M.I. HB-Eval: A System Level Reliability Evaluation and Certification Framework for Agentic AI. *Preprints.org* **2025**. https://doi.org/10.20944/preprints202512.2186.v1.
11. Adam, A.M.I. Adapt-Plan:A Hybrid Cotrol Architecture For PEI-Guided Reliable Adaptive Planning in Dynamic Agentic Enviromets. *Preprints.org* **2025**. https://doi.org/10.20944/preprints202601.0038.v1.