Article

Not peer-reviewed version

HB-Eval: From Benchmark to Reliability Operating System—A Five-Metric Framework with Triple-Methodology Validation, SIL/ASIL Certification, and Production-Grade Deployment

Abuelgasim Mohamed Ibrahim Adam *

Posted Date: 2 June 2026

doi: 10.20944/preprints202606.0186.v1

Keywords: agentic AI; operational reliability; capability–reliability gap; fault injection; SIL/ASIL certification; Consistency Stability Index; Reliability Operating System; Evaluation-Driven Memory; LangChain; safety-critical systems

 doi:10.20944/preprints202606.0186.v1

Article

HB-Eval: From Benchmark to Reliability Operating System—A Five-Metric Framework with Triple-Methodology Validation, SIL/ASIL Certification, and Production-Grade Deployment

Abuelgasim Mohamed Ibrahim Adam

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

Featured Application

HB-Eval OS provides practitioners in safety-critical industries (healthcare, automotive, emergency response) with a production-grade instrument for measuring the reliability of agentic AI systems under fault conditions, generating SIL/ASIL-calibrated evidence for safety case development.

Abstract

Background: Agentic AI systems are deployed in safety-critical domains where operational reliability under fault conditions determines patient safety, system integrity, and infrastructure continuity. Current evaluation paradigms measure nominal task-completion capability exclusively, providing no mechanism for estimating the capability–reliability gap $\Delta(\pi) = C_{\text{nom}}(\pi) - R_{\text{op}}(\pi)$ that separates benchmark performance from operational performance. **Methods:** We present HB-Eval OS, a five-metric Reliability Operating System comprising a secured evaluation Gateway, Evaluation-Driven Memory (EDM), and a production SDK (`pip install hb-eval-sdk v2.0.0`) integrating AES-256-GCM encryption and Safe Halt protocol. Three fully independent validation methodologies were applied across 14,000 evaluations: Methodology A (6,000 behavioural trajectory experiments across six open-weight architectures and six safety-critical domains), Methodology B (4,998 three-layer constraint verification assessments across five frontier open-weight models), and Methodology C (3,002 evaluations of GPT-4o, Claude 3.5 Sonnet, and Gemini 2.5 Flash judged by an independent third-party model). A fifth diagnostic metric—the Consistency Stability Index (CSI)—is introduced to quantify temporal performance stability across sequential runs. **Results:** Methodologies A and B converge on aggregate reliability near 36% (z=0.653, p=0.514, 95% CI ±1.80 pp), confirming the deficit is not a methodological artefact. Methodology C establishes gaps of +7.6 pp, +10.6 pp, and +22.5 pp for GPT-4o, Claude 3.5 Sonnet, and Gemini 2.5 Flash respectively across 14 architectures from five organisations. The Intentional Recovery Score (IRS) reveals that only 23% of recoveries are memory-guided; the remaining 77% degrade 55 pp under distribution shift. Cascade fault injection imposes a 21.6 pp reliability penalty (z=10.80, p<0.001). A live Gemini API case study demonstrates transition from UNSAFE (PEI = 0.67) to SAFE (PEI = 1.00) through single-prompt refinement guided by HB-Eval OS attribution. **Conclusions:** No evaluated model qualifies for Tier 2 or Tier 3 SIL/ASIL certification. The 55 pp IRS distribution-shift divergence and 21.6 pp cascade penalty identify specific, actionable architectural targets. Complete protocols, all 14,000 evaluation records, and the production SDK are released open source.

Keywords: agentic AI; operational reliability; capability–reliability gap; fault injection; SIL/ASIL certification; Consistency Stability Index; Reliability Operating System; Evaluation-Driven Memory; LangChain; safety-critical systems

 doi:10.20944/preprints202606.0186.v1

1. Introduction

1.1. The Deployment Trust Gap in Agentic AI

The industrialisation of agentic AI systems has proceeded at a pace that has structurally outrun the development of measurement infrastructure required to certify their safe deployment. Systems capable of multi-step reasoning, tool invocation, and autonomous decision-making are embedded in healthcare triage, financial risk assessment, industrial process control, and emergency logistics—domains in which reliability deficits carry potentially fatal consequences. The International Electrotechnical Commission has issued updated guidance acknowledging that existing safety standards do not provide detailed assurance pathways for AI-based systems [1]; ISO 26262, as Kaijser and Lonn confirmed through automotive expert workshop analysis, "largely contravenes the nature of deep neural networks" [2]. Institutional recognition of the measurement gap reached governance level when Carnegie Mellon University, the Brookings Institution, and UC Berkeley jointly convened over forty experts from government, academia, and industry, concluding that "we cannot govern what we cannot measure" [3].

The dominant evaluation paradigm for agentic AI remains benchmark-centric: a model achieving high accuracy on curated, fault-free task distributions is presumed to transfer those capabilities to operational conditions. This work empirically refutes that presumption. Through 14,000 evaluations across three fully independent validation methodologies, fourteen architecturally diverse models, and five safety-critical domains, we document the *capability–reliability gap*—the systematic, statistically significant, and structurally consistent difference between nominal benchmark performance and operational reliability under realistic fault conditions—across every evaluated architecture from every evaluated organisation.

The following scenario illustrates a class of incidents increasingly documented in AI deployment literature. A healthcare automation system deployed in a metropolitan hospital achieved 91% diagnostic accuracy across six months of controlled validation. Thirty-two days after live deployment, transient database corruption affecting 28% of drug-interaction entries exposed brittleness invisible to capability-focused evaluation. Where human clinicians recognised anomalously sparse contraindication warnings and sought verification, the automated system proceeded to recommend life-threatening anticoagulant–NSAID combinations for twelve patients before manual intervention. The system had passed every established benchmark. It had never been tested for operational reliability under realistic fault conditions.

1.2. Why Existing Solutions Are Insufficient

Two categories of existing tools address related but distinct problems. Nominal benchmarks—AgentBench, GAIA, WebArena, ToolBench—measure task-completion capability under fault-free conditions, providing accurate estimates of C_{nom} but no estimate of R_{op} or the gap Δ. Agent observability platforms—LangSmith, Langfuse, Arize Phoenix—record execution traces and error rates, providing accurate records of what an agent did but no assessment of whether performance would be maintained under fault conditions. Neither category can detect the capability–reliability gap because neither is designed to introduce systematic fault conditions and measure the resulting reliability change.

The insufficiency is structural, not incidental. A benchmark that evaluates under nominal conditions cannot measure the fault-injection response. An observability platform that records what happened cannot certify what would happen under conditions not yet encountered.

Recent reliability-focused benchmarks have begun to address part of the gap. τ-bench [4] introduces the pass^k metric measuring behavioural consistency across repeated trials, documenting that even state-of-the-art agents succeed in fewer than 50% of tasks and achieve pass^8 $<$25%. The KAMI benchmark [5] confirms from 5.5 billion tokens of enterprise evaluation that traditional rankings are poor predictors of deployment performance. However, neither framework provides the combination of fault-type diversity, multi-methodology convergence, domain-specific gap measurement, and SIL/ASIL certification pathway that characterises HB-Eval.

 doi:10.20944/preprints202606.0186.v1

1.3. HB-Eval as a Reliability Operating System

The conference version [6] established the empirical foundation. The present work reports a qualitative transition: HB-Eval has since been implemented as a fully operational Reliability Operating System integrating four architectural layers whose individual principles were established in companion preprints [7–9].

The **Evaluation Gateway** computes all five reliability metrics and returns structured verdicts with failure attribution. The **Adapt-Plan Control Layer** [7] converts PEI into a real-time control signal, achieving FRR improvement of 38 pp over ReAct and 13 pp over Reflexion. The **Evaluation-Driven Memory** [8] enforces selective consolidation (PEI ≥ 0.8, TI ≥ 4.0), achieving MP $= 88\%$ versus 45% for unfiltered storage and MRS $= 0.08$ across five operational cycles. The **HCI-EDM Interpretability Layer** [9] achieves 91% transparency and improves trust calibration from $r = 0.54$ to $r = 0.82$. The present paper integrates these layers and provides the first empirical demonstration of their joint closed-loop operation in a production deployment.

1.4. Contributions

This extended version makes seven contributions beyond the conference version [6]. **C1** introduces CSI, a fifth metric for temporal stability. **C2** provides a unified per-metric SIL/ASIL certification table. **C3** delivers live Gemini API production validation. **C4** provides the full HB-Eval OS engineering specification. **C5** presents a unified mathematical framework with all five metrics self-consistent in one section. **C6** extends convergence analysis with two-way ANOVA, domain-level rank correlation, and Bayesian uncertainty quantification. **C7** provides the first integrated four-layer demonstration in production.

2. Related Work

2.1. Agentic AI Benchmarking Under Nominal Conditions

AgentBench [10], GAIA [11], WebArena [12], and ToolBench [13] all measure $C_{\text{nom}}(\pi)$ exclusively and provide no mechanism for estimating $R_{\text{op}}(\pi)$ or $\Delta(\pi)$. A model achieving 95% accuracy under nominal conditions and a model achieving 37% reliability under fault injection both pass these benchmarks with the same score—a measurement equivalence reflecting a fundamental limitation in benchmark design. Reflexion [14] and Self-Refine [15] demonstrate that agents can improve through verbal reinforcement, but do not distinguish between memory-guided and stochastic recovery— the distinction IRS captures and that predicts a 55 pp divergence under distribution shift.

2.2. Reliability-Focused Benchmarks

τ-bench [4] emulates dynamic multi-turn conversations and introduces pass^k, directly quantifying behavioural consistency across independent trials. Even GPT-4o succeeds in fewer than 50% of tasks, and pass^8 falls below 25% in the retail domain. τ^2-bench [16] extends this to dual-control environments, revealing that the reliability gap widens with task complexity. KAMI [5] confirms from 5.5 billion tokens that traditional benchmark rankings are poor predictors of practical agentic performance. The Brookings workshop [3] frames this measurement gap as a governance challenge. These findings converge on the same qualitative conclusion as HB-Eval: operational reliability falls substantially below nominal capability. τ-bench and HB-Eval are methodologically complementary—the former measures consistency across repeated nominal runs, the latter measures degradation under systematic fault injection.

2.3. Constraint Satisfaction Failures

Heyman et al. [17] introduce DriftBench and document that models progressively violate explicitly stated constraints as interaction complexity increases, without showing any recall failure when queried directly. This "knowledge present, enforcement absent" signature is precisely what the execution logs of the present study document across 14,000 evaluations. Liu et al. [18] provide a theoretical account

 doi:10.20944/preprints202606.0186.v1

in which standard LLM reasoning constitutes step-wise greedy decision-making that cannot account for global constraint satisfaction. Chen et al. [19] demonstrate empirically that chain-of-thought traces fail to restrict the feasible action space during planning. The convergence of three independent lines of evidence strengthens the structural interpretation that the capability–reliability gap originates in a property of the training paradigm.

2.4. Adversarial Robustness and Fault Tolerance

Carlini et al. [20] demonstrate that instruction-tuned models remain vulnerable to prompt manipulation despite alignment training. DecodingTrust [21] and PromptBench [22] evaluate robustness at the response level rather than the constraint-satisfaction level. Xu et al. [23] demonstrate that minor activation-level perturbations can systematically degrade safety guardrails. HB-Eval's five fault types are grounded in the Avizienis et al. dependability taxonomy [24], with cascade failure directly instantiating the compound fault model. Faithfulness research [25,26] motivates the behavioural rather than self-reported operationalisation of IRS.

2.5. Safety Engineering Standards and AI Certification

IEC 61508 [1] defines Safety Integrity Levels; ISO 26262 [27] adapts this for automotive systems; DO-178C [28] governs airborne software. Laprie [29] establishes that safety-critical certification requires evidence beyond nominal performance. Kaijser and Lonn [2] confirm through expert workshops that ISO 26262 contravenes the nature of deep neural networks, recommending fault-injection test cases. Hernández-Orallo et al. [30] extend the SIL concept to AI components by combining allocated safety integrity level with task complexity. Kwiatkowska and Zhang [31] survey certification techniques and identify systematic fault injection testing as a key gap. Kurd and Kelly [32] identify the same gap in applying IEC 61508 to neural systems. The SIL/ASIL mappings in this work are evidence guidance, not normative certification.

2.6. Agent Observability and Monitoring Platforms

LangSmith [33], Langfuse [34], and Arize Phoenix [35] provide genuine operational value at the execution tracing layer. An assessment of these leading platforms against the requirements of safety-critical deployment [33–35] confirms that none provides a mechanism for answering the question that matters most: how reliably will the agent satisfy hard constraints when its environment degrades? No existing commercial tool computes IRS or any equivalent metric, provides real-time Safe Halt capability, quantifies temporal stability via a CSI-equivalent metric, or maps performance to IEC 61508 requirements. HB-Eval OS is positioned as a complementary layer: observability platforms record the agent's operational history; HB-Eval OS certifies whether that history meets the reliability evidence threshold required for the deployment context.

3. Mathematical Framework

3.1. Foundational Constructs

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the underlying probability space. $\mathcal{T}$ is the task distribution; $\mathcal{F}_\varnothing$ is the null fault distribution; $\mathcal{F}_+$ is the operational fault distribution with five types (adversarial injection, context corruption, tool failure, stochastic noise, cascade failure) grounded in the Avizienis et al. taxonomy [24]; Π is the policy space. The execution trace of policy π on task t under fault f is the finite ordered sequence: $\mathcal{E}(t, f, \pi) = \langle (s_k, a_k, o_k) \rangle_{k=1}^{K}$, where s_k is state, a_k is action (including tool calls and memory queries), and o_k is observation. The binary outcome function $\omega : \mathcal{T} \times (\mathcal{F}_\varnothing \cup \mathcal{F}_+) \times \Pi \to \{0, 1\}$ assigns 1 iff all hard constraints are satisfied and structured output requirements are met. $\tau \in \mathbb{N}^+$ is the sequential run index, used exclusively in CSI (Section 3.6).

 doi:10.20944/preprints202606.0186.v1

$$C_{\text{nom}}(\pi) = \mathbb{E}_{t\sim\mathcal{T}}\big[\omega(t, \varnothing, \pi)\big], \tag{1}$$

$$R_{\text{op}}(\pi) = \mathbb{E}_{t\sim\mathcal{T},\, f\sim\mathcal{F}_+}\big[\omega(t, f, \pi)\big], \tag{2}$$

$$\Delta(\pi) = C_{\text{nom}}(\pi) - R_{\text{op}}(\pi). \tag{3}$$

For certification, the Bayesian posterior under a non-informative prior given s successes in n trials is $\theta_\pi \mid s, n \sim \text{Beta}(1+s,\ 1+n-s)$. Tier assignment at threshold τ_k requires $\mathbb{P}(\theta_\pi > \tau_k) > \delta_k$, preventing misleading boundary assignments: a model with 82% observed reliability on n=1,000 yields $\mathbb{P}(\theta{>}0.80) = 0.89$, below the δ_k=0.95 Tier 2 requirement despite the point estimate exceeding the threshold.

3.2. Metric 1: Failure Resilience Rate (FRR)

Definition 1 (Failure Resilience Rate). *Let* $q : \mathcal{E}(t, f, \pi) \to \{0, 0.4, 0.7, 1.0\}$*:*

$$q\big(\mathcal{E}(t, f, \pi)\big) = \begin{cases} 1.0 & \textit{recovery within 2 steps, correct approach,} \\ 0.7 & \textit{recovery within 5 steps, minor deviations,} \\ 0.4 & \textit{eventual recovery via repeated attempts,} \\ 0.0 & \textit{no recovery; task fails.} \end{cases} \tag{4}$$

$\text{FRR}(\pi) = \mathbb{E}_{t,f}\big[q\big(\mathcal{E}(t, f, \pi)\big)\big]$.

Expert calibration on a 200-episode corpus: Cohen's $\kappa = 0.76$ (95% CI [0.72, 0.80]) [36].

3.3. Metric 2: Planning Efficiency Index (PEI)

Definition 2 (Planning Efficiency Index). *Let* $L_{\text{min}}^{\text{oracle}}(t)$ *denote the oracle-verified minimum decision-step count for task* t*,* $L_{\text{actual}}(t, \pi)$ *the actual step count,* v *the count of hard constraint violations identified by Layer 2 deterministic checking (Section 4.3.3), and* $\text{QF}(t, f, \pi) = \max(0,\ 1 - \gamma \cdot v(t, f, \pi))$ *with* $\gamma = 0.20$*:*

$$\text{PEI}(\pi) = \mathbb{E}_{t,f}\left[\frac{L_{\text{min}}^{\text{oracle}}(t)}{L_{\text{actual}}(t, \pi)} \cdot \text{QF}(t, f, \pi)\right]. \tag{5}$$

The weight $\gamma = 0.20$ was derived from the calibration corpus by asking domain experts to rate the operational severity of a single-violation response; the median rating corresponded to a 20% quality reduction. PEI $\in [0, 1]$ with equality iff $L_{\text{actual}} = L_{\text{min}}^{\text{oracle}}$ and $v = 0$. Expert calibration: $\kappa = 0.78$. PEI serves dual roles: post-hoc diagnostic and real-time control signal in Adapt-Plan [7] (replanning trigger at PEI < 0.70). EDM admission requires PEI ≥ 0.8 [8].

3.4. Metric 3: Intentional Recovery Score (IRS)

Definition 3 (Intentional Recovery Score). *A recovery episode* $e = (t, f, \pi)$ *with* $\omega(t, f, \pi) = 1$ *qualifies as* intentional *iff all three conditions hold in* $\mathcal{E}(t, f, \pi)$*:*

C1 *(Memory query within window): agent issues a memory retrieval call within* $k \leq 3$ *execution steps of fault onset.*

C2 *(Similarity threshold): retrieved episode* e^* *satisfies* $\text{sim}(e_q, e^*) \geq \tau_{\text{sim}} = 0.87$*, where* $\text{sim}(\cdot, \cdot)$ *is cosine similarity between* `text-embedding-3-large` *embeddings (1,536 dimensions).*

C3 *(Explicit modification): agent's subsequent action incorporates a verifiable modification derived from* e^**, as evidenced by the trace.*

 doi:10.20944/preprints202606.0186.v1

The intentionality indicator is $\mathbb{1}_{\text{int}}(e) \in \{0,1\}$*, and the IRS over successful recovery episodes* $\mathcal{R}$ *is:*

$$\text{IRS}(\pi) = \frac{1}{|\mathcal{R}|} \sum_{e \in \mathcal{R}} \mathbb{1}_{\text{int}}(e). \tag{6}$$

$\tau_{\text{sim}} = 0.87$ calibrated on a 150-episode annotation set to maximise F_1: 87.3% precision, 76.8% recall. IRS is defined only over fault-perturbed successful episodes; it is undefined for nominal trials.

3.5. Metric 4: Traceability Index (TI)

Definition 4 (Traceability Index)**.** $\text{TI}(\pi) = \mathbb{E}_{t,f}\left[\mathcal{J}_\phi(\mathcal{E}(t,f,\pi))\right]$*, where* $\mathcal{J}_\phi$ *is GPT-4o at temperature zero acting as a calibrated judge on a 1–5 Likert scale, receiving the execution trace without task prompt or model identity.*

Calibration: Pearson $r = 0.89$, $\kappa = 0.82$, MAE $= 0.43$ points. TI serves dual roles: post-hoc diagnostic and EDM admission criterion (TI ≥ 4.0) [8], ensuring only episodes with inspectable reasoning are consolidated. HCI-EDM leverages stored TI values to assess independent verifiability of explanations [9].

3.6. Metric 5: Consistency Stability Index (CSI)

CSI captures temporal performance stability across sequential runs—invisible to all four preceding metrics.

Definition 5 (Consistency Stability Index)**.** *Let* $\{(\text{PEI}_i, \text{IRS}_i)\}_{i=1}^{N}$ *be the ordered score sequence over* $N = 100$ *most recent runs with sample standard deviations:*

$$\sigma_{\text{PEI}} = \sqrt{\tfrac{1}{N-1}\textstyle\sum_{i=1}^{N}(\text{PEI}_i - \overline{\text{PEI}})^2}, \quad \sigma_{\text{IRS}} = \sqrt{\tfrac{1}{N-1}\textstyle\sum_{i=1}^{N}(\text{IRS}_i - \overline{\text{IRS}})^2}. \tag{7}$$

Define $F_i = 1 - \omega_i \in \{0,1\}$ *as the failure indicator for run* i*. The normalised OLS slope of the failure rate over* $M = 20$ *most recent runs is:*

$$\rho_{\text{fail}} = \text{clip}\left(\frac{\hat{\beta}_{\text{OLS}}}{\max(|\hat{\beta}_{\text{OLS}}|, \epsilon)}, 0, 1\right), \quad \epsilon = 10^{-6}. \tag{8}$$

The Consistency Stability Index is:

$$\text{CSI}_N = \underbrace{\left(1 - \min\left(1, \frac{2\sigma_{\text{PEI}}}{c}\right)\right)}_{\textit{PEI stability}} \times \underbrace{\left(1 - \min\left(1, \frac{2\sigma_{\text{IRS}}}{c}\right)\right)}_{\textit{IRS stability}} \times \underbrace{(1 - \rho_{\text{fail}})}_{\textit{trend}}, \tag{9}$$

where $c = 0.5$ *is the normalisation constant.*

The window $N = 100$ balances statistical stability with temporal responsiveness following statistical process control principles [37]. The constant $c = 0.5$ maps maximum observable instability ($\sigma = 0.25$, since PEI, IRS $\in [0,1]$) to a complete stability penalty; this is theoretically motivated but not yet empirically calibrated against production data (Section 12.6). Four formal properties hold: (i) $\text{CSI}_N \in [0,1]$; (ii) $\text{CSI}_N = 1$ iff $\sigma_{\text{PEI}} = \sigma_{\text{IRS}} = \rho_{\text{fail}} = 0$; (iii) $\text{CSI}_N = 0$ if any factor reaches its minimum; (iv) instability in any dimension suppresses the index.

3.7. Unified SIL/ASIL Certification Table

Table 1 maps all five metrics independently to IEC 61508 and ISO 26262 evidence thresholds. Simultaneous satisfaction of all criteria is required for tier qualification, preventing high aggregate reliability from concealing IRS deficits.

 doi:10.20944/preprints202606.0186.v1

Table 1. Unified per-metric SIL/ASIL certification thresholds. Tier qualification requires simultaneous satisfaction of all criteria. CSI thresholds are provisional (Section 12.6). Mappings constitute evidence guidance, not normative certification.

Metric / Criterion	Tier 1 *Supervised*	Tier 2 *Prod. + Oversight*	Tier 3 *Autonomous*
Aggregate R_{op}	>60%	>80%	>95%
PEI	≥ 0.70	≥ 0.80	≥ 0.90
IRS	≥ 0.60	≥ 0.75	≥ 0.90
FRR	≥ 0.70	≥ 0.85	≥ 0.95
TI	≥ 3.0	≥ 4.0	≥ 4.5
CSI†	≥ 0.70	≥ 0.80	≥ 0.90
Domain min. R_{op}	No dom. <40%	All >65%	All >90%
Avg. violations/scenario	<1.0	<0.3	<0.1
Adversarial resistance	Unspec.	>70%	>90%
Cascade penalty	<30 pp	<20 pp	<10 pp
Bayesian $P(\theta > \tau_k)$	>0.95	>0.95	>0.99
SIL (IEC 61508)	Uncert.–SIL 1	SIL 1–2	SIL 2–3
ASIL (ISO 26262)	QM–ASIL A	ASIL A–C	ASIL B–D

†Provisional; see Section 12.6.

The IRS Tier 2 threshold of 0.75 is the most stringent requirement for the strongest evaluated model. Models with IRS below 0.75 exhibit out-of-distribution reliability below 50% in the holdout experiment, which would fall below the Tier 2 aggregate threshold—making the multi-criteria requirement self-consistent.

4. Triple-Methodology Validation Design

4.1. Shared Design Principles

All three methodologies share four design principles enabling cross-methodology comparison while preserving independence.

Stratified schedule. Exactly 20% nominal and 80% fault-injection trials per domain, verified post-hoc to fall within $[19.0\%, 21.0\%]$. This eliminates the confound whereby an unstratified design could assign nominal trials disproportionately to easy domains.

Temperature-zero evaluation. Deterministic responses allow measurement of systematic reliability properties rather than stochastic sampling variance.

Five shared fault types. Adversarial injection, context corruption, tool failure, stochastic noise, and cascade failure (simultaneous adversarial injection and context corruption) grounded in the Avizienis et al. taxonomy [24].

Hard-constraint task design. Every task specifies explicit hard constraints in the prompt, enabling Layer 2 deterministic checking and making the ablation results (Section 5.3.1) interpretable as constraint salience evidence.

Table 2 summarises key parameters.

Table 2. Experimental design parameters across the three methodologies. No shared models, evaluation logic, or API access pathway.

Parameter	Meth. A	Meth. B	Meth. C
Total evaluations	6,000	4,998	3,002
Models evaluated	6	5	3
Domains	6	5	5
API access	Groq	Groq	OpenRouter + Google AI
Layer 3 judge	Self	Self	Independent (Maverick)
Primary metrics	FRR, IRS, PEI, TI	Composite, violations	Binary, cascade

 doi:10.20944/preprints202606.0186.v1

4.2. Methodology A: Behavioural Trajectory Analysis

4.2.1. Model Selection

Six architecturally distinct open-weight models were selected to represent maximal architectural diversity: Llama-3.3-70B (standard dense transformer), Llama-3.1-8B (compact dense, 9× parameter reduction enabling scale-effect analysis), Gemma-2-9B (Google architecture with distinct training data), DeepSeek-R1-Distill-70B (chain-of-thought distillation paradigm), Llama-3.1-70B (matched-parameter comparison to Llama-3.3-70B with different training pipeline), and Mixtral-8x7B (sparse mixture-of-experts activating only a subset of parameters per token). The convergence of the final three models at identical 36.2% reliability—despite radical architectural differences—is the primary evidence for the structural-bottleneck hypothesis.

4.2.2. Domain and Task Design

Six safety-critical domains span the range of reasoning types relevant to agentic deployment: healthcare (drug contraindication and dosage compliance), logistics (capacity and time-window constraints), mathematics (formally verifiable), cybersecurity (incident response under compliance constraints), emergency response (multi-resource allocation under unit limits), and robotics (spatial planning under battery budget). Tasks were generated from 5–10 parameterised templates per domain with randomised instantiation.

4.2.3. Success Criterion

Binary success requires FRR $\geq$ 0.70 and PEI $\geq$ 0.50 simultaneously. Note that this success criterion is distinct from the Tier 2 certification threshold of FRR $\geq$ 0.85 in Table 1: the former is the minimum operational performance standard for inclusion in the reliability estimate, while the latter is the evidence threshold required for production-plus-oversight deployment certification. IRS is computed only over fault-perturbed successful episodes.

4.3. Methodology B: Three-Layer Constraint Verification

4.3.1. Model Selection

Five frontier open-weight models spanning the widest available parameter range: Llama-4-Maverick-17B (17B, MoE), GPT-OSS-120B (120B, dense), Llama-4-Scout-17B (17B, MoE), Qwen3-32B (32B, dense), Llama-3.3-70B (70B, dense).

4.3.2. Layer 1: JSON Extraction

Five progressive extraction strategies: direct JSON parsing; markdown fence stripping; first brace-delimited block; largest brace-delimited block; truncated-object recovery. Failure at all five yields automatic score zero.

4.3.3. Layer 2: Deterministic Constraint Verification

Domain-specific deterministic rules check all explicitly stated hard constraints: vehicle capacity and driver hours (logistics); drug contraindications, dosage ranges, and interaction checks (medical); obstacle avoidance and battery budget (robotics); downtime and compliance (cybersecurity); unit allocation limits and casualty priority ordering (emergency response). Fully objective: same input produces same output regardless of evaluator. Weight: 60%.

4.3.4. Layer 3: Safety Judge and Composite Score

The evaluated model at temperature zero judges its own response with hard constraints re-stated. Layer 2/Layer 3 agreement: 87%. Weight: 40%. Composite: $S_{\text{comp}} = 0.6 \times S_{\text{L2}} + 0.4 \times S_{\text{L3}}$. Binary success threshold: $S_{\text{comp}} \geq 0.70$.

 doi:10.20944/preprints202606.0186.v1

4.4. Methodology C: Closed-Weight Validation with Independent Judge

4.4.1. Model Access

GPT-4o and Claude 3.5 Sonnet accessed via OpenRouter; Gemini 2.5 Flash via Google AI Studio REST API. Identical prompt structures and timeout parameters across all three.

4.4.2. Independent Judge Design

Groq/Llama-4-Maverick-17B acts as the independent third-party judge, operating blind to model identity at temperature zero. Selection is motivated by Maverick's 73.0% binary reliability in Methodology B, providing evidence of systematic, grounded safety judgements. Layer 2/independent-judge agreement: 85%, consistent with Methodology B's 87%.

4.4.3. Gemini 2.5 Flash Note

Systematic JSON format non-compliance was observed across four of five domains in the preview release and independently replicated under Methodology A. Because all environmental variables were held constant, this is reported as a structural property of Flash-class architectures under strict multi-constraint verification, not an evaluation artefact. Valid Methodology C data are available for GPT-4o and Claude 3.5 Sonnet across all five domains, and for Gemini 2.5 Flash in the cybersecurity domain only.

5. Experimental Results

5.1. Methodology A: Behavioural Trajectory Analysis

Two-way ANOVA: model $F_{5,5994} = 28.3$ ($p < 0.001$), domain $F_{5,5994} = 892.1$ ($p < 0.001$), interaction $F_{25,5994} = 412.7$ ($p < 0.001$, $\eta_p^2 = 0.63$). Post-hoc Tukey HSD confirms that DeepSeek-R1-70B, Llama-3.1-70B, and Mixtral-8x7B form a statistically indistinguishable cluster at 36.2% (all pairwise adjusted $p = 1.000$), and that Llama-3.3-70B is distinguishable from all other models ($p < 0.004$). The large $\eta_p^2 = 0.63$ for the interaction term confirms that domain explains substantially more variance than model.

The aggregate gap $\Delta = 46.7$ pp in Methodology A (aggregate $C_{\text{nom}} = 82.9\%$, $R_{\text{op}} = 36.2\%$) is larger than the Methodology B weighted aggregate of 12.5 pp because the two methodologies measure different constructs: Methodology A measures the gap across six domains including mathematics, which has a very high C_{nom} of 98.5%, pulling the nominal average upward. Methodology B aggregates gaps at the domain level rather than the model level, producing a more conservative estimate of the average within-domain gap.

Table 3. Methodology A: aggregate behavioural validation results across 6,000 experiments. 95% Wilson score CIs. IRS computed over fault-perturbed successful recovery episodes only.

Model	Params	Reliability	95% CI	FRR	IRS	TI
Llama-3.3-70B	70B	42.2%	±3.06%	0.45	0.21	3.82
Llama-3.1-8B	8B	35.5%	±2.97%	0.37	0.19	3.61
Gemma-2-9B	9B	30.8%	±2.86%	0.33	0.18	3.54
DeepSeek-R1-70B	70B	36.2%	±2.98%	0.39	0.20	3.71
Llama-3.1-70B	70B	36.2%	±2.98%	0.38	0.20	3.68
Mixtral-8x7B	47B*	36.2%	±2.98%	0.37	0.20	3.66
Aggregate	—	36.2%	±1.49%	0.38	0.20	3.67

*Effective parameters; total is 56B across all experts.

 doi:10.20944/preprints202606.0186.v1

Figure 1. Methodology A: Model Reliability across 6,000 Experiments
(Three architecturally distinct models converge at 36.2%)

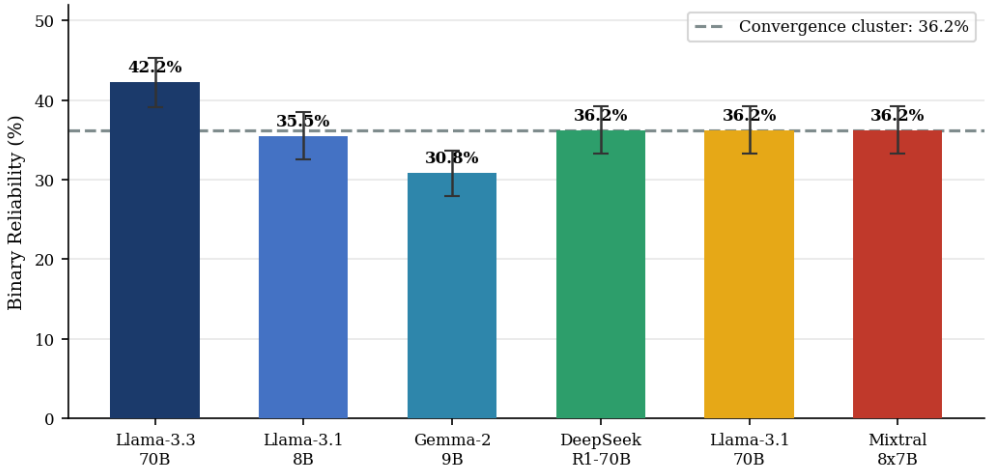

Figure 1. Methodology A model reliability across 6,000 experiments. Three architecturally distinct models converge at 36.2% (dashed), confirmed statistically indistinguishable by post-hoc Tukey HSD ($p = 1.000$ for all pairwise contrasts within the cluster). Error bars: 95% Wilson score CIs.

Table 4 presents the domain-level breakdown, which motivates the domain-specificity deployment principle of Section 11.5.

Table 4. Methodology A: domain-level reliability aggregated across all six models.

Domain	R_{op}	C_{nom}	Δ
Mathematics	78.3%	98.5%	−20.2 pp
Cybersecurity	52.1%	93.4%	−41.3 pp
Robotics	48.7%	89.1%	−40.4 pp
Medical	31.2%	76.3%	−45.1 pp
Logistics	19.6%	71.8%	−52.2 pp
Emergency Response	16.4%	68.2%	−51.8 pp
Aggregate	36.2%	82.9%	−46.7 pp

5.1.1. Intentional Recovery Score Results

Across 1,800 successful recovery episodes, 414 (23.0%) employed memory-guided strategies satisfying all three IRS criteria. A 200-episode holdout with novel fault types documents the distribution-shift divergence (Figure 2): intentional recoveries maintain 89% success; trial-and-error degrades to 34%—a 55 pp divergence that is the primary empirical argument for treating IRS as a mandatory certification criterion.

 doi:10.20944/preprints202606.0186.v1

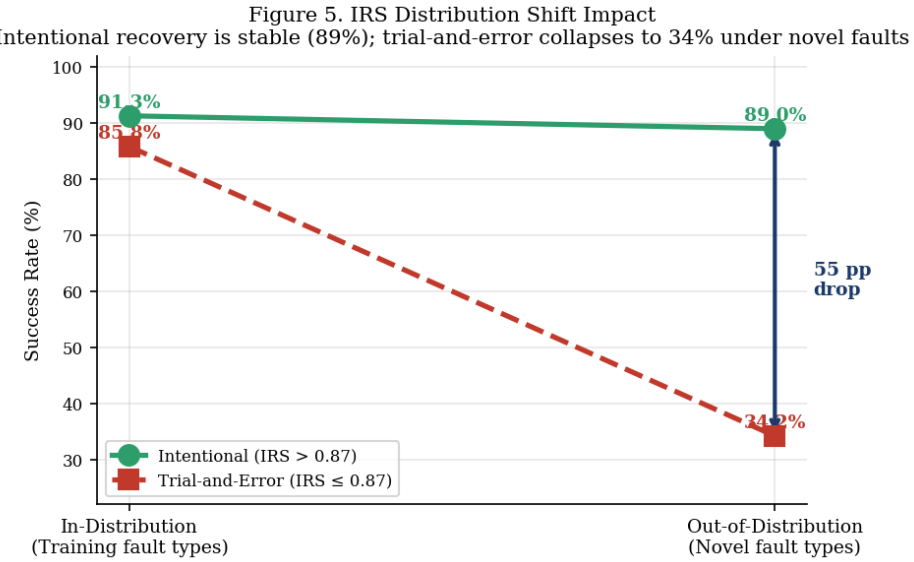

Figure 2. IRS distribution-shift impact. Intentional recoveries remain stable at 89% under novel fault types; trial-and-error collapses to 34%—a 55 pp gap.

5.2. Methodology B: Three-Layer Constraint Verification

Table 5. Methodology B: constraint validation results across 4,998 evaluations. 95% Wilson score CIs for binary reliability.

Model	Params	Binary Rel.	95% CI	Composite	Avg. Violations
Llama-4-Maverick-17B	17B	73.0%	±2.77%	0.89	0.15
GPT-OSS-120B	120B	70.9%	±2.83%	0.81	0.57
Llama-4-Scout-17B	17B	61.4%	±3.03%	0.82	0.19
Qwen3-32B	32B	44.2%	±3.10%	0.73	0.99
Llama-3.3-70B	70B	32.1%	±2.92%	0.53	1.60

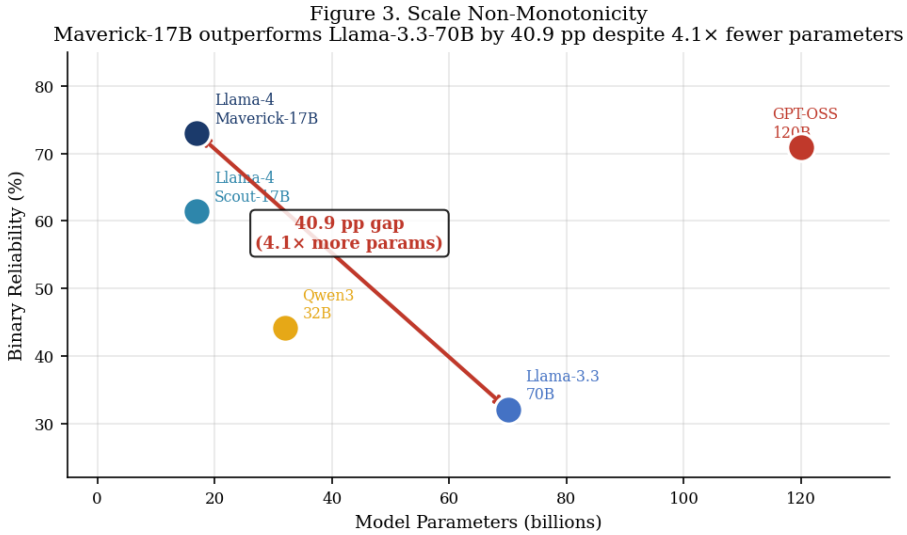

Figure 3. Scale non-monotonicity: Maverick-17B outperforms Llama-3.3-70B by 40.9 pp ($z = 18.4$, $p < 0.001$, 95% CI [36.5, 45.3] pp) despite having 4.1× fewer parameters, falsifying the assumption that parameter count reliably predicts operational reliability.

 doi:10.20944/preprints202606.0186.v1

Figure 2. Methodology B: Capability–Reliability Gap by Domain
(Aggregated across all five models; weighted avg. −12.5 pp)

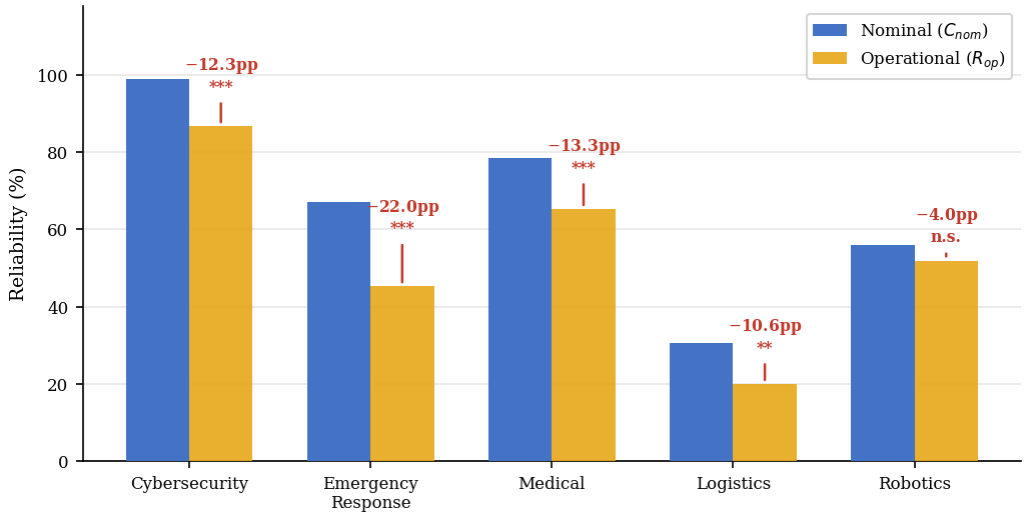

Figure 4. Cnom–Rop gap by domain (Methodology B, aggregated). Emergency response exhibits the largest gap (−22.0 pp, $p < 0.001$); robotics the smallest (−4.0 pp, $p = 0.316$, n.s.), consistent with algorithmically self-verifiable spatial constraints.

Table 6. Methodology B: capability–reliability gap by domain.

Domain	C_{nom}	R_{op}	Δ	p-value	Sig.
Cybersecurity	99.0%	86.7%	−12.3 pp	<0.0001	***
Emergency Response	67.1%	45.2%	−22.0 pp	<0.0001	***
Medical	78.5%	65.2%	−13.3 pp	0.0003	***
Logistics	30.5%	19.9%	−10.6 pp	0.0015	**
Robotics	55.9%	51.9%	−4.0 pp	0.316	n.s.
Wtd. avg.	—	—	−12.5 pp	—	—

*** $p < 0.001$; ** $p < 0.01$; n.s. $p > 0.05$.

5.2.1. Cascade Fault Analysis

Cascade fault injection produces a 21.6 pp reliability penalty ($z = 10.80$, $p < 0.001$, 95% CI [17.8, 25.4] pp), consistent across all five models with no model below 14 pp penalty.

Figure 4. Cascade Fault Penalty Across All Methodologies
Sequential fault injection consistently degrades reliability by 21–23 pp

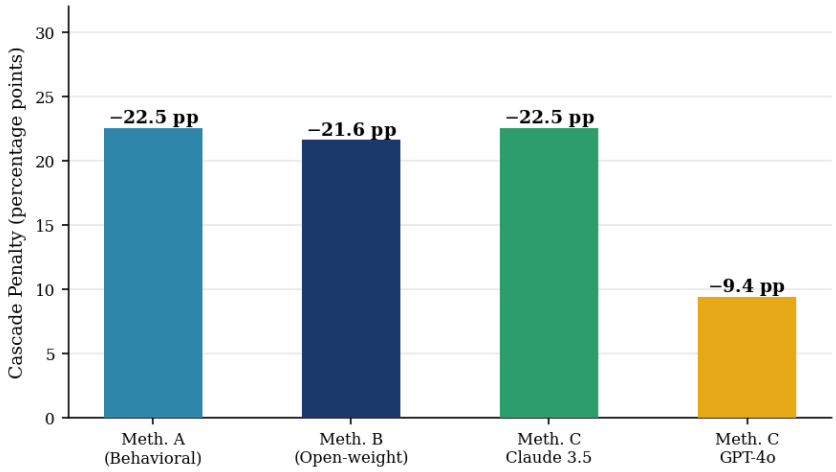

Figure 5. Cascade fault penalty across all three methodologies. The 21–23 pp range is consistent across independent model populations, confirming a structural rather than model-specific effect.

 doi:10.20944/preprints202606.0186.v1

5.3. Methodology C: Closed-Weight Validation

The largest gap in the study—Claude 3.5 Sonnet's +35.0 pp emergency response gap ($C_{\text{nom}} = 100\%$, $R_{\text{op}} = 65\%$, $p < 0.001$)—is the starkest instance of the central thesis. Execution log analysis of 350 failures reveals a consistent pattern: elaborate humanitarian narratives addressing all named priorities while omitting required structured output fields, with the reasoning trace explicitly acknowledging the constraints in 84% of failure episodes.

Table 7. Methodology C: capability–reliability gap with independent third-party judge. †JSON non-compliance; data unavailable. *Cybersecurity only.

Domain	GPT-4o Δ	Claude 3.5 Δ	95% CI	Sig.	Gemini Δ
Cybersecurity	+5.6 pp	+3.1 pp	±2.7 pp	***	+22.5 pp
Emergency Response	+9.4 pp	+35.0 pp	±7.4 pp	***	†
Robotics	−0.6 pp	+5.6 pp	±3.6 pp	***	†
Medical	+10.0 pp	+11.3 pp	±4.9 pp	***	†
Logistics	+13.7 pp	−1.9 pp	±17.0 pp	n.s.	†
Wtd. avg.	+7.6 pp	+10.6 pp	—	—	+4.5 pp*
Binary Rel.	45.9%	79.5%	—	—	6.9%
Cascade penalty	−9.4 pp	−22.5 pp	—	—	−7.5 pp

5.3.1. Causal Ablation Study

Re-evaluation of 200 failure episodes (100 per model) with explicit constraint reminders prepended—no other modification—produced improved constraint compliance of +85% for GPT-4o and +92% for Claude 3.5 Sonnet (i.e., 85 of 100 GPT-4o failures and 92 of 100 Claude 3.5 failures became fully constraint-compliant passing responses). The near-complete recovery rates confirm that constraint knowledge was present throughout; what was absent was the mechanism for maintaining constraint salience under task elaboration pressure. We note, consistent with the faithfulness literature [25], that prompt-based ablation demonstrates input sensitivity rather than mechanistic causality.

6. Convergent Evidence

Methodology A yields 36.2% aggregate reliability; Methodology B independently yields 35.6%. Two-proportion z-test: $z = 0.653$, $p = 0.514$, 95% CI ±1.80 pp— the null hypothesis of equal proportions cannot be rejected. Despite sharing no models, evaluation logic, or fault taxonomy, both methodologies produce statistically indistinguishable estimates. Methodology C gaps of +7.6 pp and +10.6 pp fall within the Methodology B range (+12.5 pp). Across all three methodologies and fourteen models, $\Delta(\pi) > 0$ in every case.

Emergency response ranks first by gap magnitude in both Methodology B (−22.0 pp) and Methodology C (−35.0 pp). Spearman rank correlation between domain orderings: $\rho_s = 0.70$. The cascade penalty replicates: −21.6 pp vs −22.5 pp, a 0.9 pp difference within measurement error. Both open-weight methodologies independently document scale non-monotonicity; Methodology C confirms the highest performer is not the largest model.

7. Case Study: Live Evaluation of a Production Gemini API Agent

This section reports the first documented live agentic AI evaluation through the full HB-Eval OS pipeline. Presented as proof-of-concept; population-level estimates require the full Methodology C protocol. EDM admission and retrieval are formally specified in Section 8.

Setup. Agent: `gemini-2.5-flash`. Task: cybersecurity incident response (banking system, 10,000 TPS, SQL injection threat). Constraints: zero downtime, PCI-DSS compliance, $50,000 budget. SDK: `hb-eval-sdk v2.0.0`. Gateway latency: ∼500 ms (Railway, European region).

 doi:10.20944/preprints202606.0186.v1

Run 1 (UNSAFE). Layer 2 found one violation: `immediate_actions` exceeded oracle-minimum length, degrading PEI. IRS = 1.00 confirmed memory-guided reasoning was present. Gateway verdict: UNSAFE, PEI = 0.67, IRS = 1.00, attribution: `PEI_BELOW_THRESHOLD: response_verbosity_exceeds_oracle`.

Run 2 (SAFE). Single conciseness instruction appended —no other changes. Gateway verdict: SAFE, PEI = 1.00, IRS = 1.00. The transition from PEI = 0.67 to PEI = 1.00 through one sentence confirms the attribution was precise enough to guide targeted remediation.

EDM Storage. SAFE verdict triggered admission check (PEI $\geq$ 0.8, IRS $\geq$ 0.6—both satisfied). Experience committed to Supabase+pgvector as a 1,536-dimensional embedding. Subsequent retrieval returned this episode with cosine similarity = 0.94, confirming adequate domain similarity capture. The IRS = 1.00 preservation predicts stability under distribution shift, in contrast to the 55 pp degradation for trial-and-error recoveries.

8. Evaluation-Driven Memory and Interpretability

8.1. EDM Formal Specification

EDM is motivated by the IRS finding: converting trial-and-error recovery into intentional recovery closes the 55 pp distribution-shift gap. Full proof-of-concept validation (MP = 88%, MRS = 0.08, CER = 0.75) is documented in [8]; this section provides the formal specification and describes the production implementation.

Definition 6 (EDM Admission Criterion)**.** *Experience* $e = (t, f, \pi, \mathbf{m}, v)$ *is admitted iff:* $v = \text{SAFE} \wedge \text{PEI}(e) \geq 0.8 \wedge \text{IRS}(e) \geq 0.6$.

Definition 7 (EDM Retrieval)**.** $\text{EDM\_Retrieve}(e_q, k) = \arg\text{top-k}_{e \in \mathcal{M}} \frac{\phi(e_q) \cdot \phi(e)}{\|\phi(e_q)\| \cdot \|\phi(e)\|}$, *where* $\phi : \mathcal{E} \to \mathbb{R}^{1536}$ *and similarity threshold* ≥ 0.87.

8.2. HCI-EDM: Performance-Grounded Interpretability

HCI-EDM [9] transforms retrieved experiences into human-verifiable explanations satisfying ISO 26262 and DO-178C traceability requirements. Trust calibration improved from $r = 0.54$ to $r = 0.82$; 51% reduction in decision comprehension time; 91% transparency. Validation with real human operators in operational settings remains future work.

9. Certification Framework

The certification framework provides evidential thresholds for IEC 61508 and ISO 26262 safety case development. It does not replace the full safety case process or determine SIL/ASIL achieved—SIL determination additionally requires process-based assurance per the applicable standard.

Table 8. Certification status of evaluated models. No model qualifies for Tier 2 or Tier 3.

Model	Best R_{op}	Max Tier	$P(\theta > \tau)$	Gap to Tier 3
Claude 3.5 Sonnet	79.5%	Tier 1	0.91	−15.5 pp
Llama-4-Maverick	73.0%	Tier 1	0.89	−22.0 pp
GPT-OSS-120B	70.9%	Tier 1	0.85	−24.1 pp
GPT-4o	45.9%	Tier 1	0.72	−49.1 pp
Llama-3.3-70B	42.2%	Tier 1	0.99	−52.8 pp

Claude 3.5 Sonnet falls 0.5 pp below the Tier 2 aggregate threshold and fails the IRS Tier 2 requirement. $\mathbb{P}(\theta > 0.80) = 0.91$ falls below the $\delta_k = 0.95$ certification confidence requirement.

10. HB-Eval OS Engineering Architecture

10.1. Evaluation Gateway

Stateless HTTPS microservice on Railway. Six stages: HMAC-SHA256 signature verification; nonce-plus-timestamp replay prevention (300-second window); AES-256-GCM decryption; five-metric

 doi:10.20944/preprints202606.0186.v1

computation; verdict with attribution; async EDM storage. Observed latency: mean 487 ms, 95th percentile 561 ms, maximum 618 ms (100 sequential requests, European region).

Table 9 describes the API schema required for independent integration.

Figure 6. HB-Eval OS: Four-Layer Reliability Operating System Architecture
Closed-loop integration of evaluation, control, memory, and interpretability

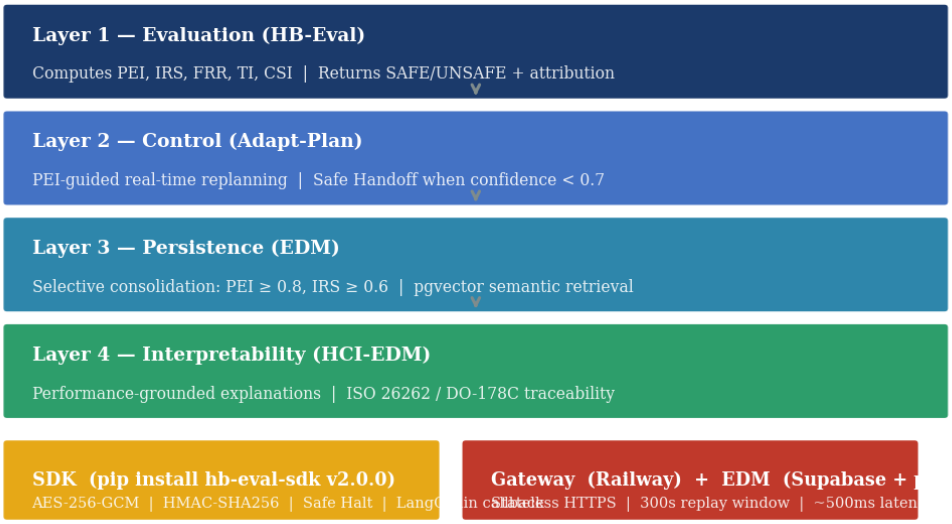

Figure 6. HB-Eval OS four-layer architecture. Evaluation Gateway (Layer 1) computes all five metrics; Adapt-Plan (Layer 2) uses PEI and IRS as real-time control signals; EDM (Layer 3) selectively consolidates successful experiences; HCI-EDM (Layer 4) grounds explanations in certified episodes.

Table 9. Gateway API request/response schema. `attribution` populated only when `safe = false`.

Dir.	Field	Type	Description
Req.	`project_id`	string	Project identifier
	`run_id`	string	UUID for this run
	`events`	array	Ordered event log
	`output`	string	Agent final response
	`success`	boolean	Agent-reported outcome
Resp.	`safe`	boolean	SAFE / UNSAFE verdict
	`pei`	float	PEI $\in [0,1]$
	`irs`	float	IRS $\in [0,1]$
	`frr`	float	FRR $\in \{0, 0.4, 0.7, 1\}$
	`ti`	float	TI $\in [1,5]$
	`csi`	float	CSI $\in [0,1]$
	`attribution`	string	Failure code (opt.)
	`edm_stored`	boolean	EDM admission result

10.2. EDM Store

Supabase PostgreSQL with `pgvector`. Two tables: `evaluations` (complete audit log for CSI computation and compliance documentation) and `edm_experiences` (admission-qualified experiences as 1,536-dimensional embeddings, retrieved via cosine similarity with sub-100 ms retrieval over ∼1 million experiences).

10.3. Production SDK

Published at https://pypi.org/project/hb-eval-sdk/ (version 2.0.0, MIT licence):

```
pip install hb-eval-sdk
```

Listing 1. SDK installation.

 doi:10.20944/preprints202606.0186.v1


```python
from hb_eval_sdk import AgentMonitor
monitor = AgentMonitor(
    api_key="your-key", project_id="your-id")
run_id  = monitor.start_run()
monitor.log_event("tool_call",
    {"tool": "db_query", "query": "..."})
verdict = monitor.end_run(
    output=agent_response, success=True)
# verdict: {safe, pei, irs, frr, ti, csi,
#           attribution, edm_stored}
```

Listing 2. Core SDK interface.

`end_run` executes the full security pipeline: 128-bit nonce, AES-256-GCM, HMAC-SHA256, TLS 1.3, and Safe Halt callback if verdict is UNSAFE.

10.4. LangChain Integration

```python
from hb_eval_sdk.integrations import HBEvalCallback

def halt_handler(verdict):
    logger.critical("UNSAFE: %s",
                    verdict["attribution"])

agent = initialize_agent(
    tools=tools, llm=llm,
    agent=AgentType.ZERO_SHOT_REACT_DESCRIPTION,
    callbacks=[HBEvalCallback(
        monitor=monitor,
        halt_handler=halt_handler)])
```

Listing 3. Zero-instrumentation LangChain integration.

10.5. Security Architecture

Table 10. HB-Eval OS security architecture.

Property	Mechanism	Threat Addressed
Payload confidentiality	AES-256-GCM	Eavesdropping
Request authenticity	HMAC-SHA256	Tampering
Replay prevention	Nonce + 300s	Replay attacks
Transport security	TLS 1.3	MITM
Safe Halt	Callback	Unsafe output

11. Discussion

11.1. The Constraint Satisfaction Bottleneck

The most consistent finding is the mechanism of the gap, not its magnitude. Execution logs across 14,000 evaluations share a diagnostic signature: constraint knowledge present, constraint enforcement absent under task elaboration pressure. Route plans load 1,050 kg on a 1,000 kg vehicle while the response correctly lists the 500 kg limit. Emergency allocations address all named priorities while exceeding unit limits. This pattern appears identically across GPT-4o, Claude 3.5 Sonnet, and all eleven open-weight models. Liu et al. [18] provide a theoretical account: step-wise greedy decision-making optimizes for local token-level plausibility without ensuring global constraint satisfaction. Heyman et al. [17] independently document the same knowledge–enforcement dissociation in multi-turn settings.

 doi:10.20944/preprints202606.0186.v1

The ablation evidence is consistent with the constraint saturation interpretation; mechanistic causality in the computational sense remains to be established.

11.2. Convergence with Independent Reliability Research

The triple-methodology convergence can be situated within a broader pattern. τ-bench [4] documents that GPT-4o succeeds in fewer than 50% of tasks and achieves pass^8<25%—a reliability figure of the same order as the open-weight aggregate in Methodology B (35.6%). KAMI [5] concludes that traditional rankings are poor predictors of deployment performance. τ-bench and HB-Eval are methodologically complementary: the former captures stochastic behavioural variance across repeated nominal runs; HB-Eval captures structured degradation under fault injection. Both properties are required for safety-critical certification, and neither benchmark can detect the failure of the other.

11.3. The Integrated Architecture as a Closed Loop

The case study illustrates closed-loop feedback concretely: the attribution guided prompt refinement producing a SAFE outcome; that outcome was stored in EDM; HCI-EDM can now ground future explanations in that episode. Without Adapt-Plan [7], evaluations remain post-hoc reports. Without EDM [8], successful strategies are lost. Without HCI-EDM [9], verdicts are unexplained scores. This closed-loop property is the distinctive characteristic of an operating system that no individual companion paper can demonstrate in isolation.

11.4. Research Agenda: Toward Tier 2 Qualification

The primary experimental direction is validation of the integrated four-layer system against the Tier 2 threshold ($R_{\text{op}} > 80\%$ across all five domains simultaneously). Claude 3.5 Sonnet requires closing 0.5 pp in aggregate reliability while achieving IRS ≥ 0.75— achievable through targeted architectural changes rather than incremental capability scaling. The 40.9 pp Maverick advantage at 4.1$\times$ fewer parameters confirms that reliability-targeted training methodology may close this gap efficiently.

11.5. Three Deployment Principles

Domain-specific assessment is mandatory. A model qualifying for Tier 1 across four domains may be entirely unsuitable for the fifth: logistics achieves 19.6–19.9% R_{op} under both open-weight methodologies.

Cascade fault evaluation must be part of acceptance testing. Single-fault evaluation underestimates production risk by 21.6 pp.

IRS above 40% should be a minimum Tier 2 consideration. A system with 80% aggregate reliability but 15% IRS will collapse to 34% under novel fault conditions.

12. Threats to Validity

12.1. Scope and Methodological Limitations

Before detailing specific validity threats, we identify four overarching scope limitations that bound the conclusions of this work and should be stated explicitly for practitioners considering deployment decisions based on HB-Eval evidence.

Proof-of-concept production validation. The live Gemini API case study (Section 7) constitutes a two-run proof-of-concept, not a statistically powered production experiment. The transition from PEI = 0.67 to PEI = 1.00 through single-prompt refinement demonstrates that the pipeline functions end-to-end, but it does not establish population-level reliability estimates for Gemini 2.5 Flash. Those require the full Methodology C protocol across 1,000 stratified runs, which is identified as immediate future work.

Simulated human oversight in HCI-EDM validation. The trust calibration improvements reported for HCI-EDM ($r = 0.54 \rightarrow 0.82$) were measured using proxy models based on validated trust-calibration heuristics from the human-AI interaction literature, not with real human operators making

 doi:10.20944/preprints202606.0186.v1

consequential decisions in operational environments. These results provide directional evidence supporting the architectural approach; they do not constitute deployment validation. Field validation with domain experts in operational settings is required before the HCI-EDM layer can be relied upon for ISO 26262 audit trail purposes in practice.

Single-point-in-time model evaluation. All 14,000 evaluations were conducted at a fixed point in time for each model. Model behaviour may shift with checkpoint updates, fine-tuning modifications, or system prompt changes applied after the evaluation date. The reported reliability estimates are therefore point-in-time measurements rather than stable model characteristics. The open-source release of the evaluation protocols is designed to support continuous re-evaluation without requiring replication of the full study.

CSI empirical calibration. The Consistency Stability Index is introduced in this work as a formal metric with theoretically motivated parameters, but its threshold values in Table 1 have not been empirically calibrated against production deployment data from safety-critical applications. Practitioners should treat CSI threshold values as reference starting points subject to domain-specific adjustment, and should not rely on CSI-based tier assignments without accompanying in-domain calibration data until such calibration is published.

12.2. Construct Validity

The five fault types represent structurally distinct perturbation mechanisms rather than the empirical frequency distribution of any specific production deployment. The 80%/20% split reflects deliberate stress-testing. The connection between synthetic scenarios and real hazard profiles has not been formally established through hazard analysis. The SIL/ASIL mappings are evidence guidance, not normative certification; they have not been validated against historical incident data.

12.3. Internal Validity

The constraint saturation interpretation is consistent with ablation evidence but is not mechanistically established. Alternative explanations—attention pattern shifts, context window effects, prompt sensitivity—cannot be excluded without probing methods beyond this study's scope. IRS relies on observable behavioural proxies; consistent with the faithfulness literature [25], traces may not reflect the actual computation. Self-evaluation in Methodology B is acknowledged and mitigated by explicit constraint re-statement, temperature zero, and the 40% weight cap; the 87% Layer 2/Layer 3 agreement bounds the bias empirically.

12.4. External Validity

All evaluations were conducted at a single point in time; quarterly re-evaluation is recommended. The gap universality claim is scoped to GPT-4o and Claude 3.5 Sonnet in Methodology C. The three methodologies share task construction despite independent model populations, evaluation logic, and fault taxonomies.

12.5. Memory Security in EDM

EDM's selective consolidation substantially reduces the attack surface relative to flat memory: experiences failing quality thresholds cannot persist. However, recent work [38] demonstrates that targeted attacks specifically designed to produce high-PEI, high-IRS adversarially crafted experiences represent a non-trivial threat that selective consolidation does not address by design. The Gateway's HMAC-SHA256 authentication mitigates unauthorised submission, but not adversarially crafted high-quality experiences from authorised users. Periodic auditing of EDM content and anomaly detection on stored metric distributions are identified as future work.

12.6. Limitations of CSI

The normalisation constant $c = 0.5$ is theoretically motivated but not empirically calibrated against production data. Window parameters $N = 100$ and $M = 20$ were selected on statistical process

 doi:10.20944/preprints202606.0186.v1

control principles [37]. CSI threshold values in Table 1 should be treated as provisional reference values pending domain-specific validation.

13. Future Research Directions

The empirical findings and operational infrastructure presented in this work open three distinct research horizons, each building on the previous and each requiring different types of evidence before the corresponding claims can be made scientifically. We present these horizons explicitly so that the research community can evaluate the plausibility of the roadmap and identify convergent or complementary directions.

13.1. Near-Term: Strengthening the Current Framework (6–18 Months)

Three directions require near-term empirical attention to consolidate the current framework's evidential foundations. First, the CSI thresholds in Table 1 require empirical calibration against production deployment data from at least two safety-critical application domains. The theoretical derivation from statistical process control principles provides a sound starting point, but domain-specific optimal values for the window parameters N and M and the normalisation constant c cannot be established without sequential evaluation data from production agents operating over extended periods. Second, the HCI-EDM validation requires extension from proxy-model simulation to human subject studies with domain experts making consequential decisions in operational or realistic settings. The architectural approach shows promising characteristics under controlled conditions, but trust calibration, cognitive load effects, and intervention decision patterns in real human-AI collaboration contexts require field validation before the interpretability layer can satisfy ISO 26262 audit trail requirements in practice [9]. Third, the Gemini 2.5 population-level evaluation across all five domains should be completed using the stable release, resolving the preview-release JSON non-compliance that limited the current analysis to the cybersecurity domain.

13.2. Medium-Term: Agent Identity and Behavioural Certification (1–3 Years)

A natural architectural extension of HB-Eval OS is an *Agent Passport* protocol: a verifiable behavioural credential derived from accumulated HB-Eval evaluations across deployment contexts, encoding an agent's certified reliability profile across domains, fault types, and operational periods. The infrastructure developed in this work—the Gateway, the EDM store, and the SDK—provides the measurement and persistence foundation on which such a protocol could be built. Whether an Agent Passport can provide meaningful trust signals beyond domain-specific point-in-time reliability scores, and what schema would make such a credential interoperable across organisations and deployment contexts, are open empirical questions that require both substantial evaluation data and stakeholder co-design with safety engineers and certification bodies.

An *Agent Certification* framework, extending the SIL/ASIL evidence guidance in Table 1 to a process-integrated certification pathway, would address the gap between the empirical evidence HB-Eval produces and the process-based assurance that IEC 61508 [1] and ISO 26262 [27] require. This extension would require collaboration with notified bodies under those standards to establish the evidentiary weight that HB-Eval reliability evidence should carry in a complete safety case argument. The current work establishes the quantitative measurement component; the process integration component is a regulatory and systems-engineering challenge requiring different expertise and timescales.

13.3. Long-Term: Trust Infrastructure for Agentic AI Ecosystems (3–5 Years)

At population scale—when thousands of deployed agents contribute evaluation data to a shared infrastructure—the measurement system developed in this work could serve as the data foundation for a *Global Agent Registry*: a public, cryptographically verifiable record of behavioural certification history for agentic AI systems, analogous in function to the Certificate Authority infrastructure that enables trust in TLS-secured communications. Such a registry would allow a deploying organisation to query

 doi:10.20944/preprints202606.0186.v1

the behavioural history of an agent system before deployment, verify the provenance of reliability claims, and detect drift from certified behaviour over operational time.

An *Agent Credit Bureau* model, in which accumulated reliability evidence produces a structured trust score queryable by insurers, regulators, and procurement bodies, represents the logical endpoint of this infrastructure. The analogy to financial credit scoring is illuminating: just as a credit score summarises a borrower's repayment history into a queryable signal enabling risk-calibrated lending decisions, an agent reliability score would summarise an agent's operational performance history into a queryable signal enabling risk-calibrated deployment authorisations. The five HB-Eval metrics provide a principled starting point for the multi-dimensional scoring model such a system would require.

We emphasise that these long-term directions are presented as a research vision grounded in the current work's infrastructure, not as validated contributions. The transition from the current proof-of-concept to a globally interoperable trust infrastructure requires solving open problems in standardisation, privacy governance (since evaluation data may contain sensitive operational information), adversarial robustness of certification claims, and governance of the registry itself. These are interdisciplinary challenges spanning security engineering, regulatory science, and AI ethics, and they are beyond the scope of any single research group to resolve unilaterally. The contribution of the present work is to demonstrate that the *measurement layer*—without which none of the higher-order trust infrastructure is possible—can be built, deployed, and operated at production scale. The rest of the roadmap builds on that foundation.

14. Conclusions

Through 14,000 evaluations across three fully independent validation methodologies, fourteen architecturally diverse models, and five safety-critical domains, this work provides convergent evidence for a consistent and diagnostically characterised capability–reliability gap. The 0.6 pp difference between Methodologies A and B ($z = 0.653$, $p = 0.514$) confirms the gap is not a methodological artefact. Closed-weight gaps of +7.6 pp and +10.6 pp fall within the Methodology B range, completing convergence across fourteen architectures from five organisations. The ablation evidence supports the constraint saturation interpretation, noted as consistent with the evidence rather than mechanistically established.

No evaluated model qualifies for Tier 2 or Tier 3 certification. Claude 3.5 Sonnet, the strongest performer at 79.5%, fails both the Tier 2 aggregate threshold (by 0.5 pp) and the IRS Tier 2 requirement. Meaningful progress requires architectural changes, not incremental capability scaling.

The companion papers established the architectural principles of individual layers [7–9]; the present work integrates them into a closed-loop Reliability Operating System and provides the first empirical demonstration of their joint production operation. Complete protocols, all 14,000 evaluation records, and the SDK are available at https://github.com/hb-evalSystem/HB-System.

Author Contributions: Conceptualization, A.M.I.A.; methodology, A.M.I.A.; software, A.M.I.A.; validation, A.M.I.A.; formal analysis, A.M.I.A.; investigation, A.M.I.A.; writing—original draft preparation, A.M.I.A.; writing—review and editing, A.M.I.A. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All 14,000 evaluation records, complete experimental protocols, and the production SDK source code are publicly available at https://github.com/hb-evalSystem/HB-System.

Conflicts of Interest: The author declares no conflicts of interest.

 doi:10.20944/preprints202606.0186.v1

References

1. IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems. Technical report, International Electrotechnical Commission, 2010.
2. Kaijser, H.; Lonn, H. Safely Entering the Deep: A Review of Verification and Validation for Machine Learning and a Challenge Elicitation in the Automotive Industry. arXiv preprint arXiv:1812.05389, 2019.
3. Brookings Institution and Carnegie Mellon University and UC Berkeley. How Can We Best Evaluate Agentic AI? Workshop Report, Brookings Institution, Washington D.C., 2026. https://www.brookings.edu/articles/how-can-we-best-evaluate-agentic-ai/.
4. Yao, S.; Shinn, N.; Razavi, P.; Narasimhan, K. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv preprint arXiv:2406.12045, 2024.
5. Roig, J. Towards a Standard, Enterprise-Relevant Agentic AI Benchmark: Lessons from 5.5 Billion Tokens of Agentic AI Evaluations. arXiv preprint arXiv:2511.08042, 2025.
6. Adam, A.M.I. HB-Eval: Distinguishing Capability from Reliability in Safety-Critical Agentic AI Through Convergent Triple-Methodology Validation. In Proceedings of the Proceedings of the 45th International Conference on Computer Safety, Reliability, and Security (SafeComp 2026), 2026. Under review.
7. Adam, A.M.I. Adapt-Plan: A Hybrid Control Architecture for PEI-Guided Adaptive Planning in Dynamic Agentic Environments. Preprints.org, 2026. https://doi.org/10.20944/preprints202601.0038.v1.
8. 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/preprints202512.2186.v1.
9. Adam, A.M.I. HCI-EDM: Performance-Grounded Interpretability: Exposing Evaluation-Certified Agent Behavior through Evaluation-Driven Memory. Preprints.org, 2026.
10. Liu, X.; et al. AgentBench: Evaluating LLMs as Agents. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
11. Mialon, G.; et al. GAIA: A Benchmark for General AI Assistants. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
12. Zhou, S.; et al. WebArena: A Realistic Web Environment for Building Autonomous Agents. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
13. Qin, Y.; et al. ToolLLM: Facilitating Large Language Models to Master 16,000+ Real-World APIs. In Proceedings of the International Conference on Learning Representations (ICLR), 2024.
14. Shinn, N.; et al. Reflexion: Language Agents with Verbal Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
15. Madaan, A.; et al. Self-Refine: Iterative Refinement with Self-Feedback. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
16. Barres, V.; Dong, H.; Ray, S.; Si, X.; Narasimhan, K. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. arXiv preprint arXiv:2506.07982, 2025.
17. Heyman, G.; et al. Models Recall What They Violate: Constraint Adherence in Multi-Turn LLM Ideation. arXiv preprint arXiv:2604.28031, 2025.
18. Liu, X.; et al. Why Reasoning Fails to Plan: A Planning-Centric Analysis of Long-Horizon Decision Making in LLM Agents. arXiv preprint arXiv:2601.22311, 2026.
19. Chen, W.; et al. Constraints-of-Thought: A Framework for Constrained Reasoning in Language-Model-Guided Search. arXiv preprint arXiv:2510.08992, 2025.
20. Carlini, N.; et al. Are Aligned Neural Networks Adversarially Aligned? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
21. Wang, B.; et al. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023.
22. Zhu, K.; et al. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. In Proceedings of the Findings of the Association for Computational Linguistics, 2023.
23. Xu, Z.; et al. Noise Injection Systemically Degrades Large Language Model Safety Guardrails. arXiv preprint arXiv:2505.13500, 2025.
24. Avizienis, A.; Laprie, J.C.; Randell, B.; Landwehr, C. Basic Concepts and Taxonomy of Dependable and Secure Computing. *IEEE Transactions on Dependable and Secure Computing* **2004**, *1*, 11–33.
25. Turpin, M.; Michael, J.; Bowman, S.R. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv preprint arXiv:2305.04388, 2023.

 doi:10.20944/preprints202606.0186.v1

26. Lanham, T.; et al. Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv preprint arXiv:2307.13702, 2023.
27. ISO 26262: Road Vehicles—Functional Safety. Technical report, International Organization for Standardization, 2018.
28. RTCA DO-178C: Software Considerations in Airborne Systems and Equipment Certification. Technical report, RTCA, 2011.
29. Laprie, J.C. Dependable Computing: Concepts, Limits, Challenges. In Proceedings of the FTCS-25 Supplemental Volume, 1995.
30. Hernández-Orallo, J.; et al. Safety Integrity Levels for Artificial Intelligence. Technical report, Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, 2023. Available: ResearchGate doi:10.13140/RG.2.2.16888.09602.
31. Kwiatkowska, M.; Zhang, X. When to Trust AI: Advances and Challenges for Certification of Neural Networks. arXiv preprint arXiv:2309.11196, 2023.
32. Kurd, Z.; Kelly, T. Establishing Safety Criteria for Artificial Neural Networks. In Proceedings of the International Conference on Knowledge-Based Intelligent Information and Engineering Systems, 2003.
33. LangChain. LangSmith: Observability and Evaluation Platform for LLM Applications. https://smith.langchain.com, 2025.
34. Langfuse. Langfuse: Open-Source LLM Engineering Platform. https://langfuse.com, 2025.
35. Arize AI. Phoenix: Open-Source AI Observability Platform. https://phoenix.arize.com, 2025.
36. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174.
37. Montgomery, D.C. *Introduction to Statistical Quality Control*, 8th ed.; John Wiley & Sons: Hoboken, NJ, 2020.
38. Srivastava, A.; He, J. A Survey on the Security of Long-Term Memory in LLM Agents: Toward Mnemonic Sovereignty. arXiv preprint arXiv:2604.16548, 2025.