Karlsruhe Institute of Technology

KIT Karlsruhe Institute of Technology
KDD Workshop on Evaluation and Trustworthiness of Agentic AI (ETAAI'26) | Jeju, Korea

Evaluating governed LLM-driven ML experimentation: lifecycle, provenance, and failure metrics

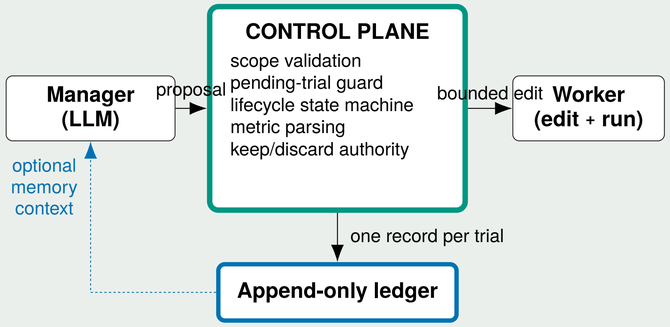
The LLM proposes; the control plane validates and adjudicates; the ledger makes every trial auditable.

Dowling Wong | Karlsruhe Institute of Technology (KIT)

1. Task scores evaluate results, not experimental conduct

Why outcome-only evaluation fails

- Did the agent edit only permitted files?
- Were crashes and invalid attempts preserved?
- Was acceptance decided independently of the agent?



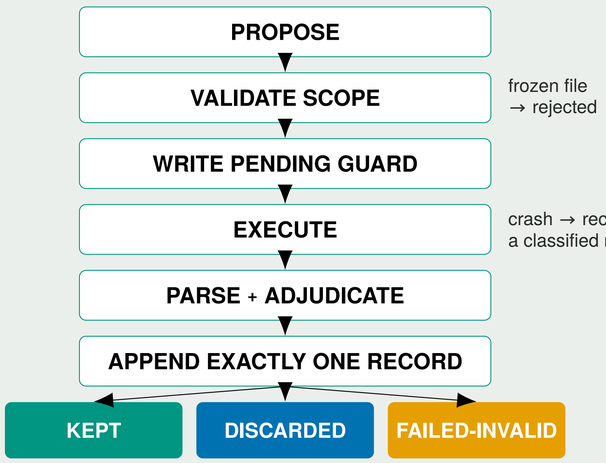
Memory is optional manager input; the ledger is the authoritative audit record — owned by the control plane, never by the agent.

Six governance metrics, reported alongside the task metric:

Acceptance	**Invalid actions**	**Repeated bad**
Provenance	**Artifact links**	**Failure taxonomy**

2. The governed trial contract

Lifecycle → exactly one immutable record per trial

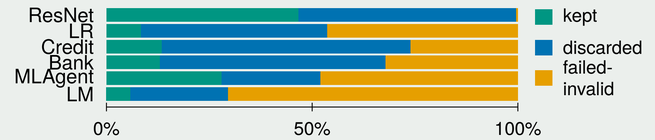


Every opened trial produces exactly one immutable terminal record before the next trial begins — the agent never owns lifecycle state or acceptance decisions.

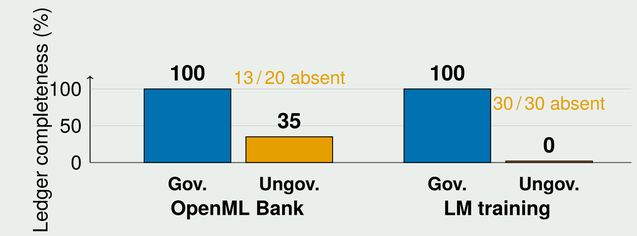
Each record: proposal ID · patch hash · run log · metric · decision · failure category

3. What the evidence shows

6 nodes · 1,445 trials · 100% provenance



Lifecycle outcomes per node — provenance complete on every node. ResNet is the verified 225-trial L40S replication; LM includes the 120-trial total-failure arm.



Matched pairs (blue = governed, orange = ungoverned); removed: pending guard, failure appends, blocking scope checks.

Without the control plane, 43 of 50 matched trials left no ledger record.

Code + ledgers + reproduction guide — all governance metrics are recomputable without a GPU. All 151 LM-training crashes classified and traceable.

Dowling Wong | dowling.wong@kit.edu
KDD ETAAI'26 | August 9, 2026 | ICC Jeju

Refs: [1] Trivedy, *The Anatomy of an Agent Harness* (2026); [2] Zaharia *et al.*, Compound AI Systems (2024).
Code + ledgers + reproduction guide: QR in panel 3 — all governance metrics recomputable without a GPU.