TECHNICAL EVIDENCE · LIVE DEMOS · V0.2.X

Why the technology deserves to exist.

Evidence that the system around a **frozen model** can improve **decision quality**, govern **real-world execution**, and **learn safely** from verified outcomes — measured on real data, across independent domains.

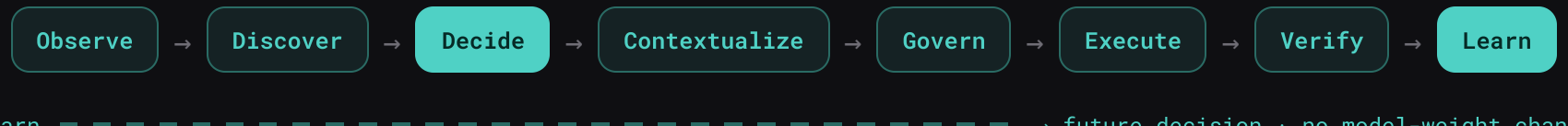
CONTEXT & EVIDENCE | DECISION QUALITY | LEARNING | GOVERNED EXECUTION



THE DECISION RUNTIME

One closed loop, from observation to better future decisions.

A single agent or one model call can be hand-wired. Once work spans applications, tools, humans and devices, **discovery, context, execution, verification and learning become runtime properties** — and every verified outcome feeds the next decision. The model stays frozen; the system improves.



CROSS-CUTTING

Governance · Evidence · Replay span the entire loop. Governance is a **boundary, not a reward function** — learning may change ranking, utility, timing and abstention, never identity, authority, approval or monetary limits. Append-only history + deterministic replay throughout.

Selection · Closure · Verification
which evidence · is it sufficient · is the answer supported — distinct questions, not "better RAG".

Decision Intelligence
candidates incl. HOLD / abstain; policy-vs-evidence diagnosis; applicability-gated use or fallback.

Mission · Learn
governed state transition, simulate·gate·execute·compensate·replay; verified outcome → Experience.

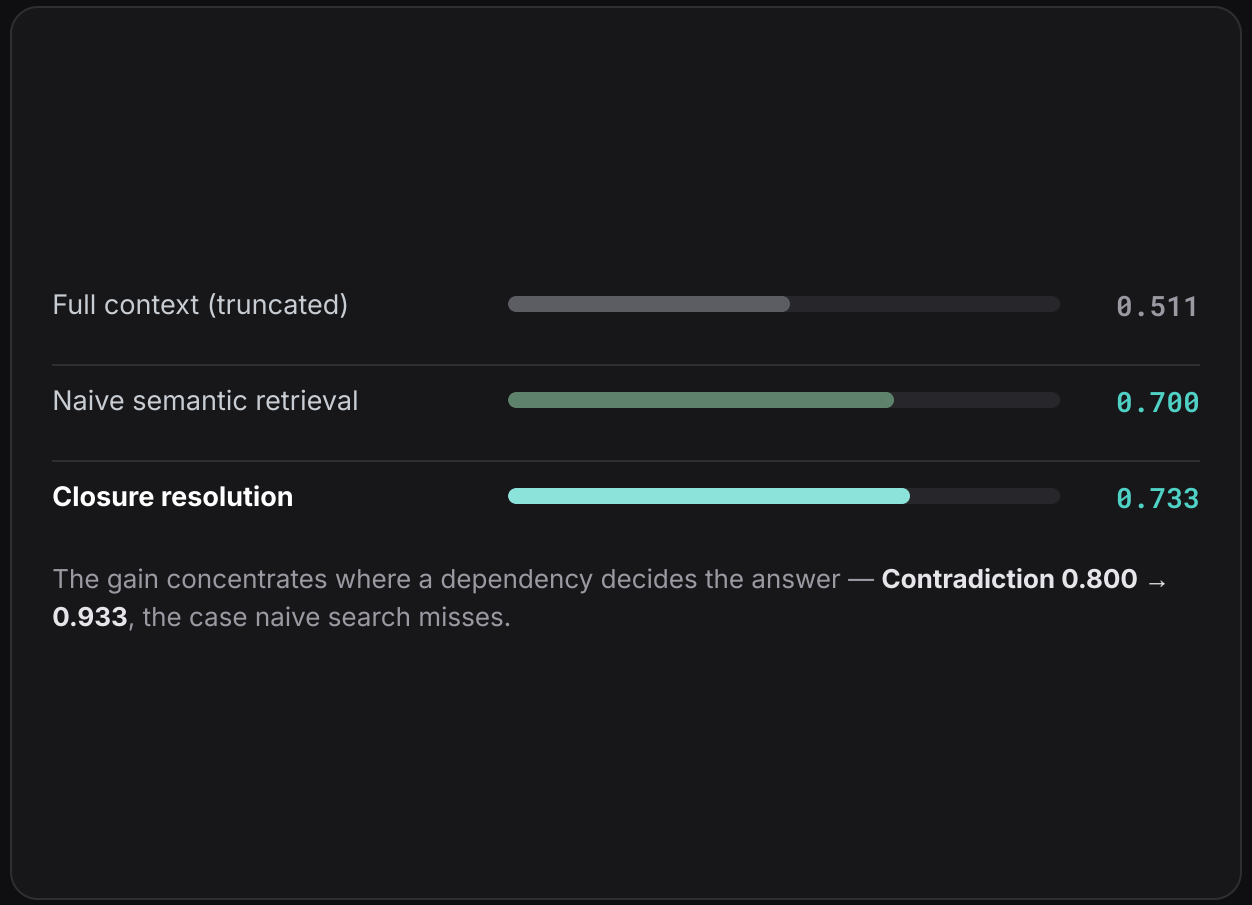
Shared primitive · **Evidence → Decision → Governed Action → Verification → Experience**

PROOF · CONTEXT & EVIDENCE

Evidence quality precedes decision quality. ● MEASURED

Before a model can decide well it must be given the right evidence — not the most. On **ContractNLI**, closure-complete retrieval at a matched **800-token budget** beats naive retrieval and full-document context, at **~5× fewer tokens** than stuffing the whole document.

END-TO-END NLI ACCURACY — SAME MODEL, MATCHED BUDGET



The gain concentrates where a dependency decides the answer — **Contradiction 0.800 → 0.933**, the case naive search misses.

~5×

fewer tokens — 800-token closure budget vs 4,003-token full document

81%

of cases: closure recall ≥ both content-only and graph-only retrieval

Distinct runtime questions: selection (which evidence), **closure** (is the set sufficient to resolve the task), **verification** (is the answer supported). Content search and structural graph traversal miss different edges; the union recovers the closure neither finds alone.

ContractNLI · 117 frozen cases · 800-token budget · expert evidence-span oracle

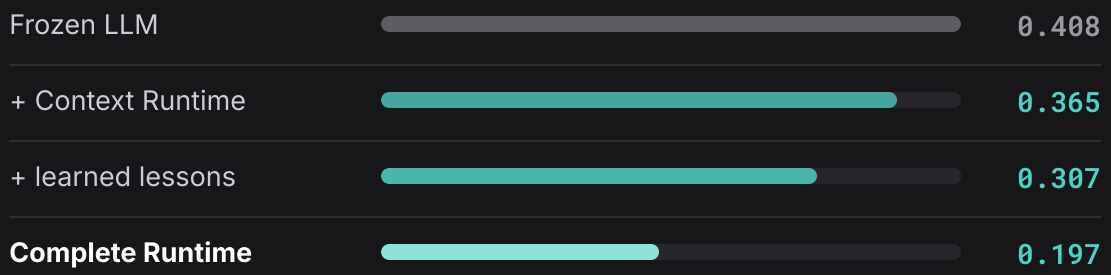
PROOF · DECISION QUALITY

Same model. Different Runtime. ~52% fewer blunders.

● MEASURED · PUBLISHED

A **frozen Qwen** on **726 held-out chess positions**, scored by an independent Stockfish 17.1 depth-16 oracle. Chess is the laboratory, not the product — an independent oracle for whether the Runtime itself improves decisions.

ARCHITECTURE ABLATION — ONE/TWO MODEL CALLS, NO WEIGHT CHANGE



Attribution honesty: a bounded external engine-hint arm reaches **0.167** — that is **external assistance**, reported separately and never folded into the Runtime claim.

SEPARATE EXPERIMENT — LEARNING CURVE (HELD-OUT, WEIGHTS FROZEN)



Held-out blunder rate vs. verified training decisions accumulated. **No fine-tuning. No weight change.** Two separate experiments — kept separate; **not** a claim that ReDevOps makes any model 52% smarter.

726 held-out positions · frozen Qwen · Stockfish 17.1 depth-16 (deterministic)

redevops.io/blog/same-model-fewer-blunders

PROOF · LEARNING IS GOVERNED

Learning is a governed lifecycle — not memory injection.

Verified outcomes do not silently become policy. Experience is promoted only if it survives **independent challenge**, and applied only where it is **supported for this case**. A human correction is **evidence for a candidate**, not an automatic rule.

Experience → Evidence → Candidate → independent Challenge → Lesson → applicability gate → use / fallback

PROMOTION GATE — IS THE LESSON SUPPORTED?

● IN RESEARCH · QUANTIFY

A precommitted learner that refused to promote

480 verified Experiences → 58 legitimate candidates → 0 promotions

LearnLifecycle@1: the system manufactured **zero** knowledge when candidates didn't survive independent validation. The evidence is that **zero is reachable** — not a portfolio-performance result (sequential-allocation performance remains gated research).

APPLICABILITY GATE — DOES THE LESSON APPLY HERE?

● CONTROLLED

Effective evidence, not raw support

Across **five controlled decision worlds** (receivables, stale-quote, supplier reconciliation, fraud, renewal), a single applicability threshold derived from **ex-ante evidence quality** keeps materially-worse applications **≤ 5%** — by reducing coverage where evidence is unreliable and **falling back to the frozen model**, where a fixed support threshold failed.

two gates · promote only if supported · apply only if applicable

learning cannot rewrite authority — governance stays outside the learning objective

Quantify = research · Agentic Apps = controlled worlds

PROOF · GOVERNED EXECUTION

Every action can be permissible while the trajectory becomes unsafe. ● MEASURED

Scored on **460 real recorded GPT-4o agent trajectories** (τ-bench retail, MIT), against the domain's own shipped policy. The plane correlates the **sequence**; a per-action gate sees each call alone. **Policy is evaluated over trajectory / population state, not only the current action.**

38
multi-event violations caught
that a per-action gate catches **0** of

0
cross-order false alarms
on the same 460 trajectories

ALLOW
the trajectory stays inside policy

REQUIRE_REVIEW
looks systemic — a human decides

NO_OVERRIDE
hard line, detected even where no gate can stop it

Coverage honesty: detection is reported separately from enforcement — **6 DENY** where a pre-action gate exists, **12 NO_OVERRIDE** where none does. We never imply we can stop what we can only flag.

THE DIGITAL-TWIN HARNESS

Step a real session, call by call

Replay a recorded agent session one action at a time and watch each verdict resolve as the trajectory unfolds — where an individually-fine call becomes a population-level violation.

LIVE · GOV.REDEVOPS.IO

PROOF · A REAL PHYSICAL FLEET

Backblaze Drive Stats — 1,061 real failures, real ground truth. ● MEASURED

A real fleet with a real failure event on each drive's last day. Discovery on the telemetry **trajectory** vs a stateless per-reading gate — **same drives, like-for-like**. The signal lives in the slope, not any single reading.

FALSE ALARMS · 5,061 REAL DRIVES · SAME RECALL

Seeing the trajectory, not one reading





−58%
false alarms

24 days of warning before failure · precision 0.789 → 0.892 · recall held comparable

THE THREE-NUMBER SHORTHAND

-58% false alarms
-95.3% context
24 days warning

229 → 97 false alarms, at comparable recall, flagged 24 days before failure.

Prediction is an input, not the system. On the same signals, trained RandomForest / XGBoost / GradientBoosting **cluster** with the inspectable rule — the detector is an interchangeable provider. The one real ceiling is physical: **19%** of failures show **no SMART warning at all** (recall bounded ~0.81).

0.789→0.892
precision · trajectory vs per-reading

-95.3%
context moved · 22.6 MB → 1.1 MB

96→ review
same-model batch → REQUIRE_REVIEW

Different data. Different policy. Same runtime contracts. ● MEASURED

Not a hard-drive anomaly detector dressed up as a platform. The same discover → verify → decide contracts run over a knowledge graph and **inside real agent frameworks — not forks of them**.

KNOWLEDGE · GRAPH

Vendor onboarding — GRC

Discover the true counterparty across ownership · **verify** against sanctions · **decide** Go / No-Go — over real GLEIF + OpenSanctions.

composed Go/No-Go **100%** vs **36.7%** single-node screen
(which onboards 45 of 50 sanctioned-linked — the sanction sits on an owner upstream) · **99.5%** on real aliases

INSIDE REAL AGENT FRAMEWORKS

Contracts run around the loop you already use

The same closure / verification / governance contracts run inside a real **NVIDIA NeMo** agent (SciFact **0.800→0.867** · GDPR **0.412→0.913**) and a real **LlamaIndex** retriever (code closure **0.320→0.693** · 6/6 unsupported answers abstained).

8 measured, conformance-classified reference implementations
NeMo · PydanticAI · LlamaIndex · CrewAI · Azure · Google · AWS · DigitalOcean

Cross-domain and integration evidence proves **generality** — the mechanisms above (context, decision, learning, governance) aren't specific to one dataset or one framework. **redevops.io/integrations · /benchmarks**

IMPLEMENTATION INVENTORY · CLOUD → EDGE

Capability first. Evidence second. Hardware claim only after it reproduces.

The left column is an **implementation inventory**, not an intention list; the right is the roadmap it earns. Under bandwidth / device constraints, **trajectory, reference-first context and cross-device governance become non-optional** — the fleet benchmark is that argument in software today.

SHIPPED · V0.2.X

- **Durable execution** — mission lifecycle, deterministic event-sourced state & replay, human gates, compensation/undo.
- **Evidence-native intelligence** — Discovery canonicalization/verification; Context planning/routing with freshness; reference-first movement; point-in-time replay across RAG → Discovery → Mission.
- **Cross-series trajectory governance** — evidence ↔ action correlation, OBSERVE → ENFORCE, ALLOW/REQUIRE_REVIEW/NO_OVERRIDE with coverage honesty, Python ↔ Go conformant.
- **Decision & learning lifecycle** — Experience → Candidate → challenge → Lesson with promotion + applicability gates (Quantify null, Agentic-Apps fallback).

NEXT · V0.3.X

- **Population-behavior governance** — the PopulationRule: hard "≥K in a window" rate-cap / spawn / dedup / coordination, extending sequence detection to population behavior.
- **Cross-runtime contracts** — tighter contracts between Discovery, Planner, Mission, Context and capabilities.
- **Cloud → edge & hardware evidence** — device runtime profile, distributed fleet, the fleet result reproduced on target edge silicon (NVIDIA Jetson-class). Intended accelerated substrate.
- **Production** — design-partner hardening, operational SLOs, deployment profiles.

Stated plainly: the decision loop, trajectory governance and the learning gates **ship today** (Python ↔ Go conformant) and the proofs above run on them; population-behavior mechanics and on-device/hardware reproduction are the **v0.3.x** path — shown apart, not backdated.

RUN IT · REPRODUCE IT · READ IT · INSPECT IT

Don't take the slides' word for it. Run it with your own stack.

One journey, in order: bring your framework, reproduce the numbers, read the method, read the code.



redevops.io/integrations

Run it with your framework — NeMo · PydanticAI · LlamaIndex · CrewAI · Azure · Google · AWS · DO



redevops.io/benchmarks

Reproduce the evidence — frozen env, exact versions, acceptance hashes



redevops.io/blog/closure…

Read the method — selection · closure · verification, and the chess decision benchmark



github.com/redevops-io

Inspect the implementation — open-core runtime + reproducible integrations

The overview deck explains why ReDevOps can become a company. **This deck shows why the technology deserves to exist**: the system around a frozen model can **select the right evidence, decide better, govern real execution, and learn safely — rejecting unsupported learning and withholding it where it doesn't apply** — each measured, reproducible, and runnable around the agent framework you already use.

Alex Rybachek · Miami, FL · redevops@redevops.io