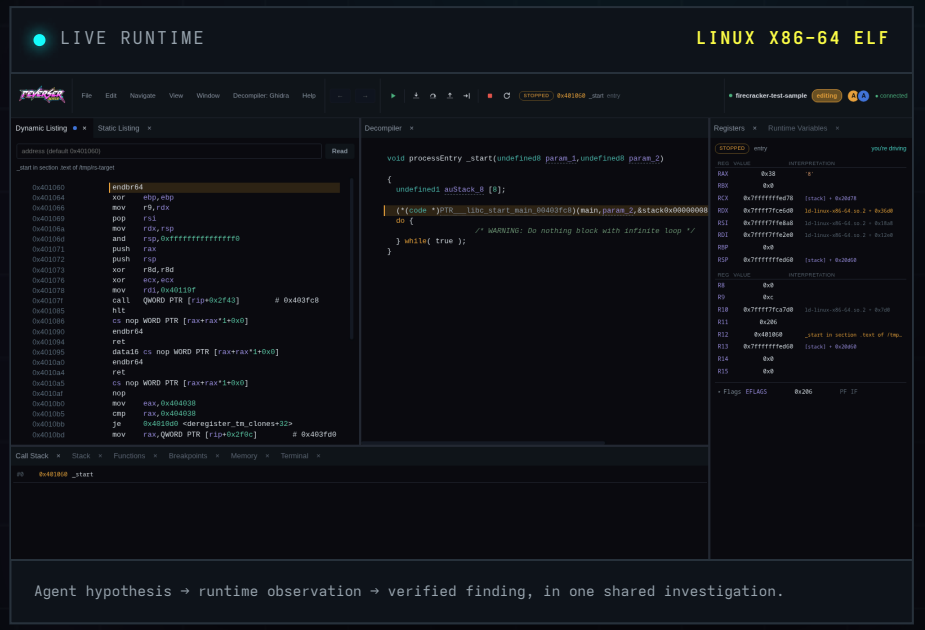
FOR AI LABS AND SECURITY EVALUATION TEAMS

EVALUATION ENVIRONMENT FOR REVERSE-ENGINEERING AGENTS.

Evaluate and improve agents with a persistent Ghidra-backed workspace, isolated Linux execution, expert intervention, and measurable trajectories.

SCOPE A DESIGN-PARTNER PILOT →

SEE AN AGENT ANALYZE A BINARY



AUTHORIZED BINARIES | PERSISTENT ANALYSIS | EXPERT INTERVENTION | EVALUATION-READY TRAJECTORIES

THE PROBLEM

A MODEL IS NOT AN ENVIRONMENT.

Prompt-and-tool agents lose analysis state, execute samples inconsistently, and leave behind transcripts that cannot show whether the reverse-engineering task was actually completed.

01

FRAGMENTED STATE

Symbols, annotations, runtime evidence, and analyst decisions are split across independent tools and conversations.

02

AD HOC EXECUTION

Runs without isolation, explicit limits, and reset semantics are difficult to reproduce or evaluate safely.

03

WEAK GROUND TRUTH

A plausible answer does not establish which actions were taken or whether the binary confirmed the conclusion.

04

FEEDBACK DISAPPEARS

Expert corrections remain trapped in chat instead of becoming attributed signals tied to the exact analysis state.

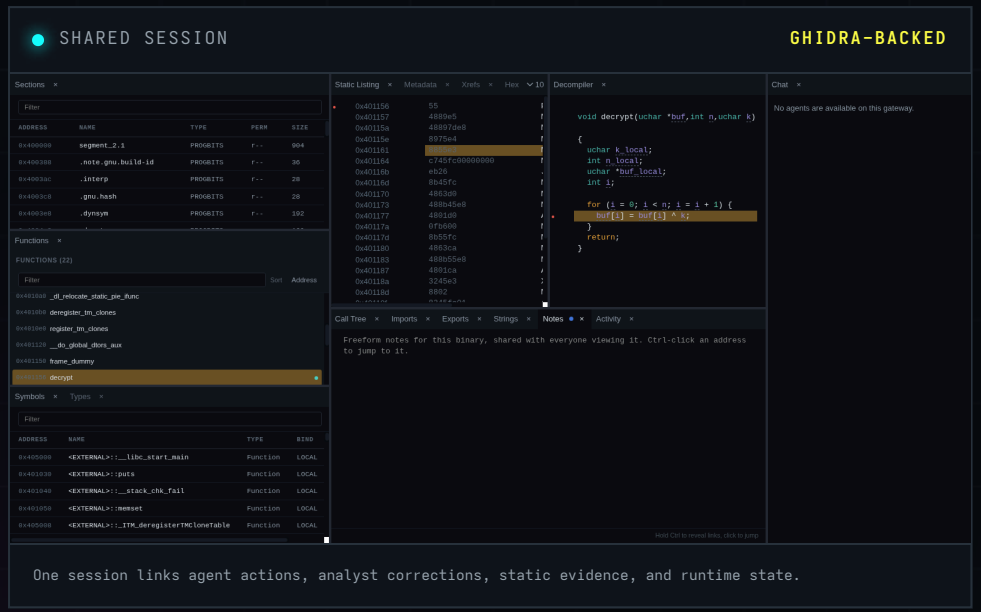
ILLUSTRATIVE ELF EPISODE

FROM SCOPED TASK TO VERIFIED FINDING.

AVAILABLE NOW

ONE ENVIRONMENT, SHARED BY HUMANS AND AGENTS.

Customer agents and human analysts share one durable investigation. They do not have to stitch together disconnected tool calls.



One session links agent actions, analyst corrections, static evidence, and runtime state.

ANALYSIS. Ghidra-backed decompilation, disassembly, graphs, symbols, strings, cross-references, hex, notes, and findings.

AGENTS. Customer agents connect through MCP and use session tools within the caller's permissions.

STATE. Sessions preserve analysis, notes, findings, and attributed activity for humans and agents.

RUNTIME. Live debugging exposes execution, registers, memory, breakpoints, modules, and terminal I/O.

POINT TOOLS	CHAT TRANSCRIPT	AD HOC VM
Independent calls and transient context	Answers without attributable evidence	Execution without episode semantics
→ ONE DURABLE INVESTIGATION	→ ACTIONS + CORRECTIONS + OUTCOMES	→ CONTROLLED, OBSERVABLE RUNS

MANAGED PILOT

YOUR AGENT. YOUR CORPUS. A DEDICATED EXECUTION BOUNDARY.

We integrate the customer's agent, run an authorized corpus inside a dedicated tenant, and deliver repeatable evaluation with structured trajectories.

WORKING DEPLOYMENT

YOUR AGENT IN THE ENVIRONMENT

+ Customer agent connected through MCP
+ Authorized corpus loaded into a dedicated tenant
+ Persistent Ghidra-backed analysis sessions
+ Disposable Linux x86-64 ELF execution
+ Fresh private VM with no public network path per run

MEASURABLE EVALUATION

A REPEATABLE EPISODE CONTRACT

+ Jointly defined tasks and acceptance criteria
+ Deterministic creation and reset semantics
+ Structured observations, actions, and expert corrections
+ Customer-defined or jointly designed evaluator
+ Scored runs with explicit termination reasons
+ Evaluation-ready trajectory export

CUSTOMER BRINGS

Model and agent framework / Authorized sample corpus / Evaluation goals and acceptance criteria

PILOT PRODUCES

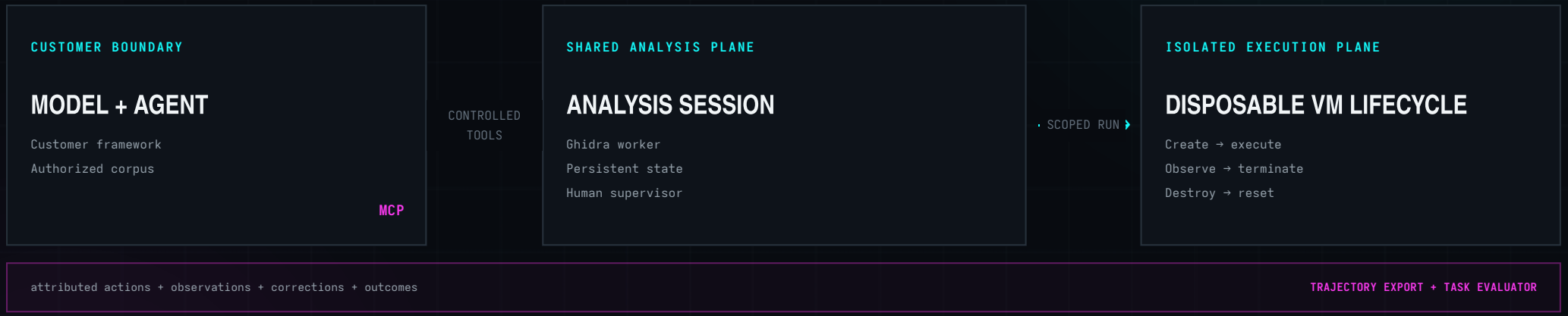
Working agent integration / Reproducible scored runs / Structured trajectory export

Definition of success. The customer's agent can complete agreed tasks against an authorized corpus in reproducible runs, with outcomes and expert interventions captured in an attributed export. Scope and acceptance thresholds are finalized with each design partner.

TECHNICAL APPENDIX

GROUND EVERY ACTION. ISOLATE EVERY RUN. KEEP EVERY SIGNAL.

The shared analysis plane preserves reasoning context. The isolated execution plane produces observable runtime outcomes. The evaluator turns both into a structured record.



Conceptual managed-pilot architecture. It intentionally omits infrastructure addresses, credentials, host configuration, and deployment secrets.

01

DATA ISOLATION

Customer binaries and trajectories remain isolated and are not reused to train shared models.

02

EXECUTION POLICY

Each managed-pilot run starts in a fresh private VM with no public address, service account, NAT path, or permitted egress.

03

HUMAN AUTHORITY

Operators can inspect the trajectory, correct the shared state, take control, and stop execution.

04

AUTHORIZED USE

Customers must provide samples they own or are explicitly authorized to analyze.

THE COMPOUNDING LAYER

THE ENVIRONMENT IS THE PRODUCT.

Every investigation produces grounded actions, observable outcomes, and expert corrections that can improve how reverse-engineering agents work.

01 **ONE ENVIRONMENT, FOUR JOBS**

Production investigation, evaluation, red-teaming, and reinforcement learning can share the same operational substrate.

02 **CORRECTIONS COMPOUND**

Expert interventions create attributed, tool-grounded signals tied to the exact state that required them.

03 **ENVIRONMENT OVER ABSTRACTION**

Persistent tools and observable consequences reveal what an agent can actually do, not only what it can say.

DESIGN PARTNERS

BRING THE AGENT AND AN AUTHORIZED CORPUS. LEAVE WITH MEASURABLE RUNS.

SCOPE A DESIGN-PARTNER PILOT →

See an agent analyze a binary

Current focus: native Linux x86-64 ELF for live debugging and managed-pilot detonation.

Pilot scope: integration, corpus, evaluator, acceptance thresholds, and operating boundaries are agreed with each design partner.