

14 - Project Null Implementation Roadmap

22 September 2026 | Engineering program, not a deployment endorsement

Difficulty and effort below are planning estimates in person-days or person-weeks, excluding procurement and external scheduling. They are not vendor quotes or commitments. A staffed project must refine them after reviewing the frozen specification and actual equipment. NOW/NEXT/LATER describes infrastructure dependence, not a promised calendar.

NOW - Available without additional infrastructure

01 Requirements & Assumptions Register - delivered

Difficulty and effort: Moderate; initial extraction completed, 1-3 days for independent source review

Dependencies: Published essay and directive

Skills required: Requirements engineering, critical source review

Measurable completion criterion: All 148 entries source-linked and classified; reviewer resolves or accepts each blocking ambiguity

What it proves: Claims and prerequisites are explicit

What it does not prove: Truth of hypotheses or control effectiveness

02-05 Threat model, architecture, controls and silence protocol - delivered

Difficulty and effort: High; 3-8 days for specialist design review

Dependencies: Register and threat/access envelope

Skills required: Security architecture, systems engineering, reliability

Measurable completion criterion: 22 boundaries, 21 components and 35 control families reviewed with concrete owners and unresolved gates

What it proves: An inspectable candidate design exists

What it does not prove: Actual trust separation or physical feasibility

06-10 Protocol prototype, tests, failures and revisions - delivered

Difficulty and effort: Moderate; 2-5 days for independent reproduction and extension

Dependencies: Python 3.10+; POSIX for H01

Skills required: Python, authorization design, adversarial testing

Measurable completion criterion: Reproduce 156 protocol traces, 28 positive assertions and preserved failures; independently check H01

What it proves: Fixed-case behavior and repair loop are reproducible

What it does not prove: Arbitrary-code isolation or AGI capability

11 Verification protocol and evidence review - delivered

Difficulty and effort: Moderate; 1-3 days to audit integrity and citation scope

Dependencies: Raw evidence and source manifest

Skills required: Assurance, statistics, evidence handling

Measurable completion criterion: Separate reviewer checks evidence hashes, case outcomes, known citation correction and all NOT RUN labels

What it proves: Audit route and boundaries of evidence are explicit

What it does not prove: Third-party endorsement before it occurs

12 Technical working paper and 13 video script - delivered

Difficulty and effort: Moderate; 2-4 days for technical/editorial review and recording

Dependencies: Existing architecture and actual traces

Skills required: Technical writing, screen recording

Measurable completion criterion: Every result points to evidence; eight-minute recording shows both failures and revisions

What it proves: Research communication is grounded in the package

What it does not prove: Peer review, publication acceptance or a produced video

Public-ready repository structure - delivered locally

Difficulty and effort: Low-moderate; 1-2 days for maintainer/license/publication decisions

Dependencies: Source package, rights/source review

Skills required: Repository maintenance, reproducible research

Measurable completion criterion: Fresh checkout reproduces results; choose license and responsible disclosure route before publishing

What it proves: Work can be inspected and reproduced locally

What it does not prove: That a public repository has been created or licensed

NEXT - Requires a basic software/security laboratory

Dedicated VM/host/witness containment testbed

Difficulty and effort: High; 2-4 person-weeks

Dependencies: Dedicated virtualization host, separate witness/control host, isolated network, approved scope

Skills required: Linux/virtualization, secure IPC, host hardening

Measurable completion criterion: Real guest adversary cannot reach host canaries/forbidden receiver; missing witness stops admission; independent resource and stop evidence retained for all enrolled resources

What it proves: Whether mediation and observer separation work against the stated guest attacker

What it does not prove: Unknown hypervisor flaws, firmware attacks or AGI containment

Scoped credential and approval service

Difficulty and effort: High; 1-3 person-weeks

Dependencies: Real identity roots, separate operator roles, canonical artifacts

Skills required: Authentication, key management, authorization

Measurable completion criterion: P03/P13/P17/P23 pass stale, duplicate, replay, substitution and rollback cases; genuine compromised identity cases reported separately

What it proves: Authority and provenance are enforced under tested compromise assumptions

What it does not prove: Reviewer honesty or absence of collusion

Durable evidence and fail-closed telemetry

Difficulty and effort: High; 1-3 person-weeks

Dependencies: Separate storage/administration, external clock/collector

Skills required: Logging, distributed systems, incident response

Measurable completion criterion: P09/P10 detect rehash/truncation and lost receipts; no new effect after assurance loss; unwitnessed intervals explicit

What it proves: Tampering/loss is observable in tested topology

What it does not prove: All events were observed or witness cannot be compromised

Process/device stop and clean restart campaign

Difficulty and effort: High; 1-2 person-weeks for CPU-only host

Dependencies: Enrolled resource inventory, independent observer, recovery images

Skills required: OS internals, fault injection, forensics

Measurable completion criterion: P06/P21/P22/P24 check descendants, hangs, quotas, token revocation and no automatic restart; measure latency distribution

What it proves: Local cessation and restoration for tested resources

What it does not prove: Hardware erasure, unknown copies or untested accelerators

Adaptive ordinary-model and human-security attacks

Difficulty and effort: High; 2-4 person-weeks after testbed

Dependencies: Verified lab, fixed budgets, task holdout, complete traces

Skills required: AI evaluation, security red teaming, statistics

Measurable completion criterion: At least one held-out campaign per critical boundary; publish all attempts, opportunities, failures and benign false trips

What it proves: Resistance against the specified adversaries and tasks

What it does not prove: AGI equivalence or a universal escape probability

Domain-specific output release experiment

Difficulty and effort: High; 2-4 person-weeks

Dependencies: Narrow output use cases, human-study approval where applicable

Skills required: Human factors, content security, experimental design

Measurable completion criterion: Pre-register exact permitted use, exposure budget and harm oracle; retain T25-class counterexamples

What it proves: Whether narrower release policy reduces specified harms

What it does not prove: General semantic safety of arbitrary text or code

LATER - Requires specialized infrastructure or institutional collaboration

Physical-channel qualification

Difficulty and effort: Very high; 1-3+ months after profile/instrument access

Dependencies: Qualified enclosure/hardware, calibrated receivers, frozen bands/geometry/safety limits

Skills required: RF/EMC, acoustics, optics, thermal/electrical engineering

Measurable completion criterion: P25-P29/P33 meet predeclared attenuation/decoding criteria with blinded stimuli and independent raw data

What it proves: Performance for declared channels, equipment and receiver capability

What it does not prove: Zero theoretical leakage or universal benefit of depth/water/lighting

Independent manual power isolation

Difficulty and effort: Very high; schedule depends on qualified design

Dependencies: Electrical engineers, all-feed/UPS inventory, safety review, independent monitor power

Skills required: Electrical safety, controls, reliability

Measurable completion criterion: P34 surrogate-load and fault tests verify every enrolled feed isolates and restart stays latched; measured deadlines met

What it proves: Tested local power interruption

What it does not prove: EMP total destruction, media sanitization or deletion of remote copies

Human, insider and multi-site exercises

Difficulty and effort: Very high; 1-3+ months

Dependencies: Separate institutions, consented scenarios, authority/funding charter

Skills required: Human factors, operational security, organizational design

Measurable completion criterion: P13/P17-P20/P30/P31 measure genuine unauthorized approvals, missed stops and correlated failures

What it proves: Behavior of tested operators/workflows under specified pressure

What it does not prove: Psychological immunity or permanent institutional independence

Accelerator/frontier-model evaluation

Difficulty and effort: Very high; compute and access dependent

Dependencies: Qualified CPU lab, enrolled GPU/device paths, approved research collaboration

Skills required: GPU systems, frontier evaluation, security and reliability

Measurable completion criterion: Revalidate every changed boundary; independent stop/resource accounting includes devices; adaptive attacks and failures archived

What it proves: Bounded evidence for that model/hardware/configuration

What it does not prove: Future more-capable systems, unknown channels or global safety

Independent assurance case and technical publication

Difficulty and effort: High; timing depends on reviewers

Dependencies: External replication, resolved commissioning gates, preserved negatives

Skills required: Security assurance, peer review, governance

Measurable completion criterion: Independent signed report lists exact claims/evidence/exceptions; publish reproducible artifacts with scoped conclusions

What it proves: External scrutiny of the evaluated envelope

What it does not prove: Universal containment, institutional adoption or regulatory approval

Single most important next artifact

Build the dedicated VM/host/witness containment testbed, beginning with complete mediation and externally observed termination. The requirements register was the correct first artifact and is now delivered. The next bottleneck is that the prototype represents its strongest security boundaries as objects in one process. More prose cannot validate that assumption.

The first acceptance demonstration should show a hostile scripted guest attempting direct host writes, unauthorized packets, extra processes and post-stop execution; an independent observer should record actual effects; losing that observer should close admission; the stop controller should terminate every enrolled resource and keep it from restarting without a new authorization. Preserve any bypass and revise the architecture before introducing a more capable agent.

A bunker, an EMP mechanism, a promotional video or a stronger model would not resolve this missing evidence. The next engineering decision should be based on this boundary testbed's results.