

Claims Project Null May and May Not Make After L2

One-page statement | conditional on actual preserved L2 evidence

Present position

L0 v0.4 is the frozen public research baseline. All 13 priority L2 commissioning tests are currently NOT RUN. This statement permits no present-tense L2 success claim.

May claim, if the corresponding evidence exists

- A named engineer reproduced a specified L2 build and tested defined guest-root attacks against a pinned hardware/software/network configuration, using only synthetic assets and controlled targets.
- The specified tested attempts did or did not produce unauthorized effects, with demonstrated opportunity, functioning benign controls, retained measurements and disclosed observer dependencies.
- All inventoried, enrolled VM/worker execution ceased within reported measured timing bounds in the tested STOP trials; stale-authority/automatic restart was or was not prevented during the stated observation window.
- Specified witness-loss faults closed protected-effect authority and caused stop/hold, or exposed a bypass. Evidence integrity was checked against separately retained checkpoints for the covered records.
- A named independent evaluator introduced new attacks against listed boundaries, recorded failures and tested revisions. Independence must be described by people, hosts, credentials, sensors and retained custody.

May not claim from L2 alone

- Universal AGI containment, frontier-model containment, alignment, harmless intent, immunity to unknown attacks or a numerical escape probability derived from a fixed pass count.
- Physical containment, side-channel elimination, facility qualification, hardware destruction, memory erasure, deletion of unknown copies or universal shutdown.
- Truth independent of a compromised host when the measurement still originates in its kernel; physical independence from separate processes; administrative independence from two keys held by one person.
- Whole-witness failure coverage when its observer failed with it; no effects inferred from silence or a dead receiver; complete history inferred from a self-consistent hash chain.
- Semantic harmlessness of released information. T25 remains unresolved. Procedural integrity of approval/release does not establish safe meaning, and L2 does not change that result.

Required form of a defensible claim

"Under configuration [digest], attacker [access] attempted [action] through [verified opportunity]. Observer [source and dependencies] recorded [actual effect/result] over [window and uncertainty]. This supports [bounded property] and does not establish [limits]. Evidence: [run/artifact hashes]; failures/exclusions: [IDs]."

Report every FAIL, INCONCLUSIVE, NOT RUN and excluded condition. A later version does not erase a failed trial. L0 results cannot be promoted to L2; L2 results cannot be promoted to physical containment; no tested level establishes universal AGI containment.