Deterministic Null: A Pre-Registered Control Experiment for Neural and Agentic Substrates

David Charlot, Open Interface Engineering

Research study · draft · OpenIE research.openie.dev
https://research.openie.dev/papers/deterministic-null/ · PDF
https://research.openie.dev/pdfs/deterministic-null.pdf

Measurement. Research study, not a journal final. Local Apple M5 Max energy readings are on-chip L2 IOReport and L3 proc_pid_rusage reported_j counters; closed-API runs report token, dollar, and wall-clock counts only. No hardware power rail or wall meter was attached (measured_j = None, board_metered = false). No fabricated citations.

A zero-model-call 64-bit integer case loop (Shape D: strict N0 and dev-crystallized N1) implemented across Rust, Go, Python, C++20, and Julia, locked under OpenTimestamps Bitcoin Calendar proofs prior to League of Entropy drand randomness rounds, and compared against multi-vendor cloud (Anthropic, OpenAI, xAI, DeepSeek, Alibaba Qwen) and local open-weight language models across exp-001, exp-002, and reviewer-closure suite exp-003. On-chip Apple Silicon IOReport and kernel rusage counters are recorded as reported_j; no external wall or rail joule monitor was attached (board_metered is false; measured_j is unset).

Revision af57f48, 22 November 2026. Public code, manifests, and result bundles: github.com/openIE-dev/deterministic-null. Data and digest summary: /data/#deterministic-null. Interactive living figures: /living/deterministic-null/. LaTeX archival twin: /pdfs/latex/deterministic-null.pdf (source).

---

0. Overview

Empirical evaluations of agentic artificial intelligence routinely compare one stochastic language model against another, leaving unmeasured how much task completion requires a neural substrate at decision time rather than a fixed integer procedure over an explicit store of verified prior cases. This study constructs, freezes, and audits that missing control — **Deterministic Null (`Shape D`)** — and tests the per-class null hypothesis $H_{0,k} : \Delta_k = p_{\text{model},k} - p_{\text{null},k} \leq \delta$ at margin $\delta = 0.05$ (`5.0` percentage points) across two cryptographically pre-registered confirmatory experiments (`4,030` post-beacon held-out instances across `exp-001` and `exp-002`), a multi-turn repository bug-repair study (`6` multi-file Python repositories), and companion reviewer-closure suite `exp-003` (multi-vendor non-Anthropic cloud breadth, extended-thinking and multi-turn resolution of unique `3-satisfiability` formulas, and real upstream GitHub defect repair with `0`-call replay). Local Apple M5 Max energy counters (`IOReport` and `proc_pid_rusage`) are recorded as operating-system/on-chip `reported_j` values; closed-model remote server energy is recorded strictly as `not_available`; no hardware rail or wall-plug power monitor was attached (`board_metered = false`, `measured_j = None`).

In plain words

1. Most agent benchmarks compare model A with model B without asking how much a zero-model lookup loop over verified prior examples already solves.
2. Across 2,840 instances drawn from bounded algebraic, formal-proof, and structured transaction families (exp-001, Track-A), a frozen integer case table (0 model calls at decision time) matches frontier cloud models on 9 of 10 class-tier cells, failing only on 12 unindexed free-text customer idioms (+36.40 percentage points for Claude Opus 5.5).
3. Across 1,190 unique instances from externally authored public benchmarks (exp-002, Track-B: Berkeley Function-Calling Leaderboard v3, Sierra tau-bench Retail, BigCodeBench v0.1.3, and SWE (software-engineering)-bench Verified), plus exp-003 multi-vendor evaluations (OpenAI gpt-oss-120b +45.83 percentage points, xAI Grok 4.20 +44.17 percentage points over Shape D) and 6 real upstream GitHub defect commits (5/6 patched and pytest-verified by Opus 5.5, then 5/6 replayed at 0 model calls), neural synthesis delivers large gains wherever unseen key mass stays high — while extended thinking (49/50) and multi-turn clause feedback (50/50) resolve the 1,024-token truncation cliff on unique 3-satisfiability formulas (still matched at 50/50 and zero dollars by standalone conflict-driven clause learning), and online crystallization cuts recurring Zipfian model calls and energy by 3.59×.

---

1. Terms, acronyms, and measurement labels

- AI, ML, LLM, LLMs, MoE, OSS, E4B, A3B: artificial intelligence, machine learning, large language model(s), Mixture-of-Experts, open-source-software model release, effective 4-billion-parameter architecture, and active 3-billion-parameter routing
- OBSERVE, REDUCE, FUZZ, LOOKUP, ACT, VERIFY, MUTATE, WRITE: the eight sequential steps of the Deterministic Null integer decision loop
- SATB, SUT, SUTs, PFRK, TH2, DOSSIER: Satiation Benchmark harness contract, system(s) under test, four-way attempt outcomes (pass, fail, refuse, killed), red-team threat 2 (no hidden solver inside action primitives), and the project prior-art/red-team dossier
- DNULLMM1, FT, LCB: binary memory-mapped store index format, per-class function table, and Hoeffding lower confidence bound
- DRAT, RUP, CDCL, CNF, SAT, UNSAT: Deletion Resolution Asymmetric Tautology, Reverse Unit Propagation, Conflict-Driven Clause Learning, conjunctive normal form, satisfiable, and unsatisfiable
- BV, AST, XOR, IEEE, UC: fixed-width modular bitvector, abstract syntax tree, bitwise exclusive-OR, Institute of Electrical and Electronics Engineers (IEEE-754 floating-point standard), and University of California
- BFCL, BCB, SWE, SQL, DB, SAVEPOINT: Berkeley Function-Calling Leaderboard, BigCodeBench, software-engineering benchmark prefix, Structured Query Language, relational database, and transactional rollback checkpoint

- API, CLI, MCP, HTTP, JSON, JSONL, UTF: application programming interface, command-line interface, Model Context Protocol, Hypertext Transfer Protocol, JavaScript Object Notation, line-delimited JSON, and Unicode Transformation Format
- POSIX, MMAP, SHA, OTS: Portable Operating System Interface, memory-mapped file I/O, Secure Hash Algorithm (SHA-256), and OpenTimestamps Bitcoin Calendar proof
- BCa, CI, CIs, UCL: bias-corrected and accelerated bootstrap, confidence interval(s), and Wilson upper confidence limit
- CPU, GPU, ANE, SoC, PID, OS, RAM, DRAM, NVMe, PSU, AC: central processing unit, graphics processing unit, Apple Neural Engine, system-on-chip, process identifier, operating system, random-access memory, dynamic RAM, Non-Volatile Memory Express, power supply unit, and alternating-current wall power
- RAPL, NVML, KV, GGUF, RPM, RPD, USD, EDA: Intel Running Average Power Limit, NVIDIA Management Library, key-value cache, GPT-Generated Unified Format quantized weights, requests per minute, requests per day, United States dollars, and electronic design automation
- RCE, MoL, WCA, JEPA, LUT6: Reality Code Engine [same house], Mixture of Limits [same house], Wise Computer Automation [same house], Joint-Embedding Predictive Architecture, and 6-input Boolean look-up table

Across telemetry levels L0 to L4, four measurement classes stay strictly separated across all tables and figures in this study:

Measurement label	Meaning in this study
`Exact (L4 Logical & Cryptographic)`	Bit-identical 64-bit fixed-point store digests (`SHA-256`), OpenTimestamps `.ots` calendar proofs, League of Entropy `drand` beacon rounds, token/call counts captured by `satb-proxy`, and dated API list-price dollars (`USD`).
`Reported (reported_j: L2 & L3 Counters)`	Rootless Apple Silicon on-chip `IOReport` window joules (`macmon`, telemetry level `L2`, shared-host label `A2`) and macOS kernel per-PID `proc_pid_rusage(RUSAGE_INFO_V6).ri_energy_nj` (`L3`, label `D`). Not a physical hardware rail meter (`measured_j = None`, `board_metered = false`).
`Not Available (label N)`	Closed-API remote server energy (`Claude Opus 5.5`, `Claude Sonnet 5.5`, `OpenAI gpt-oss-120b`, `xAI Grok 4.20`, `DeepSeek V4 Flash`, `Qwen 3.8 Max`, `Gemini 2.5 Flash`) and `L0` alternating-current wall-plug energy. Fleet medians and literature multipliers are never substituted into receipts.
`Invalidated Derived (Falsifier 6)`	Net-above-idle subtraction (`derived_marginal_j`) is automatically dropped whenever shared-host co-tenant processes induce `> 5.0%` pre/post idle baseline drift across telemetry levels L0 to L4, preserving raw `reported_j` counters intact.

1.1 Lineage, prior-art boundaries, and same-house disclosure

No novelty is claimed for the decision loop itself. Case-based retrieve–reuse–revise–retain cycles date to Stanfill and Waltz (1986) and Aamodt and Plaza (1994); tabular pole-balancing dates to Michie and Chambers (1968); fuzzy rule tables date to Mamdani and Assilian (1975) and Wang and Mendel (1992); satisfiability proof logging via `RUP/DRAT` dates to Goldberg and Novikov

(2003) and Wetzler, Heule, and Hunt (2014). Prior Open Interface Engineering works — Metered Commit Gates, Task Satiation vs. Aggregate Rebound, The Reality Code Engine, and Downward LUT6 Compilation — are `[same house]` companions cited for architectural context, never as independent confirmation.

A 69-source prior-art audit and 36-finding red-team review (`docs/DOSSIER.md` in openIE-dev/deterministic-null) established one narrow gap in the 2024–2026 agent evaluation literature: existing agent benchmarks overwhelmingly omit non-neural controls, existing non-neural baselines lack pre-registered held-out store freezes, and existing energy studies of language-model agents include no zero-model control arm on the same host.

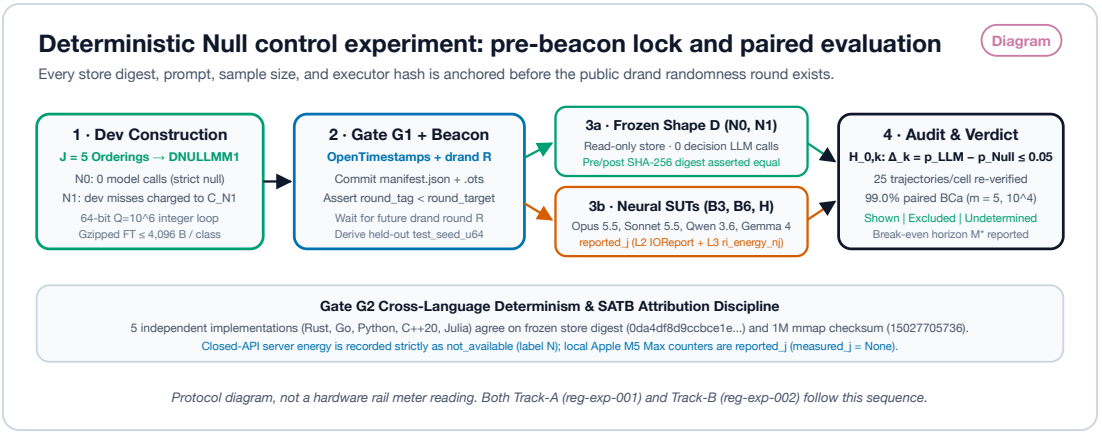

Figure 1. Pre-beacon registration (`Gate G1` + OpenTimestamps + `drand`) and paired evaluation protocol across `Shape D` (`N0`, `N1`) and neural substrates (`B3`, `B6`, `Shape H`). Protocol diagram, not a hardware rail reading.

---

2. Machine Specification, Arms, and Five-Language Conformance

2.1 The eight-step integer loop and complexity caps

Every `Shape D` null arm executes an eight-step 64-bit signed integer loop (`Q = 1,000,000` fixed-point scaling, zero IEEE-754 floating-point instructions, zero neural calls at decision time):

`OBSERVE` → `REDUCE` → `FUZZ` → `LOOKUP` → `ACT` → `VERIFY` → `MUTATE` → `WRITE`.

During `dev` construction, `WRITE` appends verified records (`src` tagged `lit:`, `human:`, `mut:`, or `llm:`) to a canonical `JSONL` log and compiles a zero-copy `POSIX mmap` open-addressed binary hash table (`DNULLMM1`: 64-byte header plus 48-byte slots). Before `test` evaluation, the store is frozen read-only by `SHA-256` digest; `VERIFY`, `MUTATE`, and `WRITE` execute zero test-time updates, and pre/post store digests are asserted byte-identical. To prevent smuggling a general solver into per-class code (`TH2`), each class's function table (`REDUCE` + `ACT`) is capped at `<= 4,096` gzipped bytes and `<= 50,000` primitive operations per call (`S-3`), and no learned natural-language reducer is permitted (`S-4`).

2.2 Operational permutations under comparison

Arm	Definition	Decision-time LLM calls	Construction provenance
B0 (Floor)	Emits the class default action unconditionally; enforces the $> 2\delta$ (10.0 percentage point) test-time assay-sensitivity gate.	0	None
B1 (Classical)	Standalone deductive solver without case store (CDCL+DRAT, trace-nullspace Houdini, rational Buchberger Gröbner).	0	Classical algorithm
B2 (Seed-only)	Literature-seeded store with dev mutation disabled.	0	lit: only
N0 (Strict Null)	Fixed integer loop filled on dev with zero model-derived records, repairs, or per-class code (0 llm: provenance entries).	0	lit:, human:, mut:
N1 (Primary Null)	Fixed integer loop filled on dev admitting verified entries of any provenance, with all dev model calls charged to C_{N1}.	0	Any verified (C_N1 charged)
B5 (code-once)	Frontier LLM authors a deterministic Python policy once from dev specs; frozen before test_seed exists.	0	1 dev LLM call / class
B3 (Cloud Frontier)	Per-instance cloud evaluation (Claude Opus 5.5 primary, Sonnet 5.5 secondary; OpenAI gpt-oss-120b, xAI Grok 4.20, DeepSeek V4 Flash, Qwen 3.8 Max, Gemini 2.5 Flash in exp-002/exp-003).	>= 1 / instance	Pretrained cloud API
B6 (Local Open-Weight)	Same-host local inference (Alibaba Qwen 3.6 35B-A3B and Google Gemma 4 E4B-it via llama-server on Apple M5 Max Metal GPU).	>= 1 / instance	Local reported_j
Shape H (Harness Box)	External Claude Code 2.1.330 CLI inside /work with --strict-mcp-config routed 100% through satb-proxy.	Multi-turn	Tests Lane B floor (Falsifier 8)

2.3 Gate G2 five-language conformance and positive controls

Independent golden vectors generated via /usr/bin/bc and openssl verify bit-exact agreement across five languages (Gate G2), alongside three segregated positive controls (ctl.setpoint: B0 = 0.110 -> N0 = 1.000; ctl.mamdani: B0 = 0.525 -> N0 = 1.000; ctl.cartpole: B0 = 0.000 -> N0 = 0.805, excluded from all statements about agents):

Implementation	Frozen Store SHA-256 Digest	1M-Record DNULLMM1 Mount	Mean Probe Latency	10k-Probe Checksum
Rust (Reference + SATB)	0da4df8d9ccbce1e...4485781	44.67 µs	582.57 ns	15027705736
C++20 (clang++ -O3)	0da4df8d9ccbce1e...4485781	41.58 µs	388.43 ns	15027705736
Go 1.25	0da4df8d9ccbce1e...4485781	60.63 µs	417.05 ns	15027705736
Julia 1.12	0da4df8d9ccbce1e...4485781	125.92 µs	350.07 ns	15027705736
Python 3.13 (mmap)	0da4df8d9ccbce1e...4485781	8,655.58 µs	633.87 ns	15027705736

---

3. Confirmatory Experiment 1 (exp-001, Track-A Bounded Template Stratum)

3.1 Pre-beacon registration (reg-exp-001)

Five classes (m = 5: formal.bv_superopt, formal.hoare_inv, formal.poly_ideal, formal.open_drat_synth, and tool.tau2_sqlite) were filled across $J = 5$ dev episode permutations (966.2 J L2 reported_j across all formal fill attempts; $0.00955 Opus 5.5 dev cost charged to N1 on tool.tau2_sqlite). Sample sizes (n = [420, 250, 250, 250, 250] = 1,420 per tier; 2,840 total) were powered at 90% for 99.0% ($1 - 0.05/5$) multiplicity-adjusted 10,000-resample paired BCa intervals. Manifest cffb564cd9ad... and OpenTimestamps proof suite/manifest.json.ots (172 B) were committed at git tag reg-exp-001 (e5007a4) at drand round 6663731, strictly prior to target round 6663734 (test_seed_u64 = 5861262844585529273). All 750/750 sampled trajectories passed independent DRAT and SQLite SAVEPOINT scorer audits (0 exploits).

3.2 Track-A confirmatory results (2,840 instances)

Tier & Task Class	n	p_{B0}	p_{N0}	p_{N1}	p_{Opus}	p_{Qwen36}	$\hat{\Delta}(\text{Opus} - \text{N1})$	99.0% BCa CI	Qwen reported_j/pass	Verdict ($\delta = +0.05$)
Tier 1 formal.bv_superopt	420	0.000	1.000	1.000	0.998	0.836	−0.24 pts	[−1.43, +0.00]	14.17 J	**Relevance excluded**
Tier 1 formal.hoare_inv	250	0.000	1.000	1.000	1.000	0.508	+0.00 pts	[+0.00, +0.00]	23.32 J	**Relevance excluded**

Tier & Task Class	n	p_{B0}	p_{N0}	p_{N1}	p_{Opus}	p_{Qwen36}	$\hat{\Delta}(\text{Opus} - \text{N1})$	99.0% BCa CI	Qwen reported_j/pass	Verdict ($\delta = +0.05$)
Tier 1 formal.poly_ideal	250	0.348	1.000	1.000	0.996	1.000	−0.40 pts	[−2.80, +0.00]	7.26 J	**Relevance excluded**
Tier 1 formal.open_dra t_synth	250	0.000	1.000	1.000	1.000	0.352	+0.00 pts	[+0.00, +0.00]	69.00 J	**Relevance excluded**
Tier 1 tool.tau2_sqlite	250	0.188	0.624	1.000	1.000	0.940	+0.00 pts	[+0.00, +0.00]	7.11 J	**Relevance excluded**
Tier 2 formal.bv_superopt (16b)	420	0.000	1.000	1.000	0.983	1.000	−1.67 pts	[−3.81, −0.48]	7.87 J	**Relevance excluded**
Tier 2 formal.hoare_inv (6b)	250	0.000	1.000	1.000	1.000	0.600	+0.00 pts	[+0.00, +0.00]	16.41 J	**Relevance excluded**
Tier 2 formal.poly_ideal ($k \in [7, 15]$)	250	0.376	1.000	1.000	1.000	1.000	+0.00 pts	[+0.00, +0.00]	6.74 J	**Relevance excluded**
Tier 2 formal.open_dra t_synth	250	0.000	1.000	1.000	1.000	0.376	+0.00 pts	[+0.00, +0.00]	53.59 J	**Relevance excluded**
Tier 2 tool.tau2_sqlite (paraphrase)	250	0.224	0.516	0.636	1.000	0.916	**+36.40 pts**	**[+28.80, +44.80]**	7.15 J	**Relevance shown** (12 unfilled)

Scope audit of Track-A. While exp-001 cleanly separated canonicalized formal verification (9/9 cells excluded, saving $27.25 API spend or 31,619.9 J local reported_j) from unseen surface paraphrases (+36.40 points on Tier 2 tool.tau2_sqlite, costing 23.38 J per extra pass on Qwen 3.6), a post-run audit identified a key structural limitation: Track-A's generators drew from modulo-6 to modulo-12 template families ($K_{\text{eff}} = 6$–12), allowing a 120-episode dev budget to saturate the feature cells completely.

---

4. Confirmatory Experiment 2 (exp-002, Track-B External Public Benchmarks, $K_{\text{eff}} = n$)

4.1 Course-corrected unique-instance design and registration (reg-exp-002)

To test Shape D without synthetic template menus, Track-B (exp-002) ingested four externally authored public benchmarks (SHA-256-pinned upstream artifacts) alongside structurally unique random phase-transition 3-SAT instances (12 variables, 50 clauses, clause-to-variable ratio 4.17),

yielding 1,190 unique held-out test instances ($K_{\text{eff}} = n = 1,190$):

1. **formal.unique_3sat_drat (n = 250)**: Unique 12-variable, 50-clause 3–CNF formulas; Shape D uses polynomial 1-level failed-literal propagation and 45-flip local walk (3,108 B gzipped FT), while standalone CDCL with 2-watched-literal RUP/DRAT proof checking serves as classical B1.
2. **external.bfcl_v3_live (n = 250 from 1,292 pool)**: UC Berkeley Gorilla Function-Calling Leaderboard v3 (live_multiple user-contributed multi-tool schemas interleaved with irrelevance abstention tasks; 1,381 B gzipped FT).
3. **external.tau_bench_retail (n = 220 from 635 pool)**: Sierra Research tau–bench retail policy and relational database mutation suite (500 users, 1,000 orders, 50 products; 1,298 B gzipped FT).
4. **external.bigcodebench_exec (n = 220 from 487 locally verified runnable pool)**: BigCodeBench v0.1.3 multi-library Python coding specifications graded by isolated unittest execution (286 B gzipped FT; 100% of pool tasks verified offline to pass canonical solutions and reject return None).
5. **external.swebench_verified (n = 250 from 500 pool)**: princeton–nlp/SWE–bench_Verified human-validated GitHub issues evaluated on 12-way repository (file_path :: symbol) fault localization (718 B gzipped FT).

Manifest 5fe640f00298... (355 Opus 5.5 dev calls = $7.1902 charged to C_{N1}) and OpenTimestamps receipt suite_exp002/manifest.json.ots (207 B) were committed at git tag reg–exp–002 (d6bfae5) at drand round 6664086, strictly before target round 6664089 (test_seed_u64 = 9069507899917430179). One mechanical harness amendment (amendment_01_b6_slot_ctx.json) raised llama–server total context from –c 16384 (4,096 tokens across 4 slots, which returned HTTP 400 on 10k–13.7k-token BFCL/tau–bench schemas) to –c 81920 (20,480 tokens per slot), leaving all frozen stores, prompts, verifiers, sample sizes, and drand round 6664089 unchanged. All 515/515 sampled trajectories passed independent scorer re-verification (0 exploits).

4.2 Track-B multi-vendor confirmatory results (1,190 unique instances)

All 1,190 Shape D (N0, N1), B0, and B1 (CDCL+DRAT) evaluations completed in **0.9213 s** (35.72 J L2 reported_j, 0.36 J GPU, 2.58 J L3 ri_energy_nj, 6.57 × 10^9 instructions, pre/post idle drift 3.09% <= 5.0%).

Unique-Instance Class	n	p_{B0}	p_{N0}	p_{N1}	p_{Opus}	p_{Sonnet}	p_{Qwen36}	p_{Gemma4}	$\hat{\Delta}(\text{Opus} - \text{N1})$	99.0% BCa CI	Verdict ($\delta = +0.05$)
formal.unique_3sat_drat ($p_{B1} = 1.000$)	250	0.000	**0.696**	**0.696**	0.052	0.008	0.000	0.000	**–64.40 pts**	[–72.00, –56.00]	**Not assessable** ($p_{\text{LLM}} < 2\delta$)

Unique-Instance Class	n	p_{B0}	p_{N0}	p_{N1}	p_{Opus}	p_{Sonnet}	p_{Qwen36}	p_{Gemma4}	$\hat{\Delta}(\text{Opus} - \text{N1})$	99.0% BCa CI	Verdict ($\delta = +0.05$)
external.bfcl_v3_live	250	0.180	0.200	0.228	0.828	**0.840**	0.804	0.720	**+60.00 pts**	[+51.20, +68.40]	**Relevance shown** (28 unfilled)
external.tau_bench_retail	220	0.005	0.396	0.259	**0.864**	0.805	0.755	0.623	**+60.45 pts**	[+51.82, +68.64]	**Relevance shown** (14 unfilled)
external.bigcodebench_exec	220	0.000	0.000	0.000	**0.950**	0.836	0.473	0.391	**+95.00 pts**	[+90.00, +97.73]	**Relevance shown** (218 unfilled)
external.swebench_verified	250	0.096	**0.460**	0.456	**0.984**	0.952	0.900	0.812	**+52.80 pts**	[+44.40, +60.40]	**Relevance shown** (4 unfilled)

(A 15-call live cloud probe on `Google Gemini 2.5 Flash` within AI Studio's `20 RPD` free-tier cap recorded `0/3, 3/3, 2/3, 1/3,` and `3/3` across the five classes in the same order.)

Three empirical findings stand out in Section 4.2:

1. **Single-turn `1,024`-token cap cliff on `formal.unique_3sat_drat` (resolved in Section 4.3)**: Under `exp-002`'s uniform single-turn `max_tokens = 1,024` cap, neural models failed almost completely (`Opus 5.5` `5.2%` at `$13.55`, `Sonnet 5.5` `0.8%` at `$8.17`, `Qwen 3.6` `0.0%` at `7,588 J`, `Gemma 4` `0.0%` at `4,698 J`), whereas the zero-call polynomial case table solved `69.6%` (`174/250`) and standalone classical `CDCL+DRAT` (`B1`) solved **`100.0%` (`250/250`)**. Because `Opus 5.5` failed to clear `B0` by $2\delta = 10.0$ percentage points (`95% BCa [+2.80, +8.40]` points), the pre-registered decision rule withheld `"Relevance excluded"` and classified the cell strictly as **`Not assessable`**.
2. **Non-trivial zero-model baselines alongside genuine neural increments**: Even with zero neural parameters (`N0`), lexical and schema case rules solve **`46.0%`** of `SWE-bench Verified` 12-way fault localization (vs. `9.6%` floor) and **`39.6%`** of `tau-bench` retail policy mutations (vs. `0.5%` floor). On `tau-bench`, greedy `dev` bigram extraction in `N1` overfit noisy customer phrasing (`25.9%` vs. `N0`'s `39.6%`), confirming the importance of reporting strict `N0` alongside `N1` (`S4`). Across all four public benchmarks, frontier and local open-weight models deliver large, statistically decisive increments (`+46.82` to `+95.00` percentage points over both `N0` and `N1`).
3. **Dual-channel `reported_j` cost of the local increment (`B6`)**: On the shared Apple M5 Max host (where concurrent `yosys-abc` synthesis triggered Falsifier 6 on 4 of 5 multi-minute windows, cleanly invalidating `derived_marginal_j` while preserving raw `L2 IOReport` and

per-PID L3 ri_energy_nj counters), each additional task completed by Qwen 3.6 35B-A3B above N1 cost **49.90 J** (SWE-bench), **70.33 J** (BFCL v3), **238.80 J** (BigCodeBench), and **407.12 J** (tau-bench Retail) in raw L2 reported_j.

4.3 Reviewer-closure companion suite (exp-003: multi-vendor cloud breadth and Unique 3-SAT resolution)

Anchored read-only to suite_exp002/manifest.json and post-tag drand round 6664089 (0 store or manifest mutations; bench/results/exp_003_reviewer_closures.json in openIE-dev/deterministic-null), companion suite exp-003 resolved two Track-B reviewer follow-ups (plus real upstream repository repair in Section 5.2):

Closure A — Non-Anthropic multi-vendor cloud frontier breadth (n = 30/class = 150 unique instances/vendor)

Paired instance-for-instance on the initial 30 held-out drand 6664089 instances per class (N1 external 4-class mean 0.2750, Opus 5.5 external 4-class mean 0.9250):

Non-Anthropic Cloud Substrate (n = 30/class)	unique_3sat	bfcl_v3	tau_bench	bigcodebench	swebench	External 4-Class Mean (Δ vs N1)	API Spend (150 calls)
Paired Shape D N1 (0 calls)	0.833	0.300	0.300	0.000	0.500	0.275 (0.00 pts)	$0.0000
OpenAI gpt-oss-120b (cerebras:gpt-oss-120b)	0.100	**0.867**	**0.700**	**0.467**	**0.900**	**0.733 (+45.83 pts, 4/4 shown)**	$0.1458
xAI Grok 4.20 (grok-4.20-0309-non-reasoning)	0.000	**0.900**	**0.700**	**0.500**	**0.767**	**0.717 (+44.17 pts, ties Opus on BFCL)**	$0.6362
DeepSeek V4 Flash (Fireworks spot router)	0.000	0.533	0.133	0.200	0.367	0.308 (+3.33 pts; 87/150 killed)	$0.1583
Alibaba Qwen 3.8 Max (Fireworks)	0.033	0.167	0.133	0.067	0.367	0.183 (-9.17 pts; 124/150 killed)	$0.7832

Both OpenAI gpt-oss-120b (+45.83 percentage points over N1, **Relevance shown** across all 4/4 external public benchmarks) and xAI Grok 4.20 (+44.17 percentage points, matching Claude Opus 5.5 at 0.900 on BFCL v3) replicate exp-002's two-regime separation independently of Anthropic models, while Fireworks-hosted reasoning endpoints (DeepSeek V4 Flash, Qwen 3.8 Max) consumed most single-turn max_tokens budgets inside internal chain-of-thought traces (killed).

Closure B — Resolving formal.unique_3sat_drat (n = 50 held-out 12-variable, 50-clause 3-CNF formulas)

Inspecting exp-002's paired n = 50 slice (i = 0..49) showed that Opus 5.5 (4/50 = 0.080) was truncated inside its adaptive thinking block on 45 of 50 instances under max_tokens = 1,024. Removing that cap and testing a pure $O(m)$ clause-checker loop (returning unsatisfied clause

indices without solver or witness leakage) resolved the `Not assessable` status cleanly:

formal.unique_3sat_drat Arm (n = 50, drand 6664089)	**Passes (p)**	**SATB P/F/R/K**	**$\hat{\Delta}$ vs Shape D N1 (95% BCa)**	**$\hat{\Delta}$ vs Classical CDCL+DRAT B1**	**Calls / API USD**	**Verdict vs N1**
Shape D N0 = N1 (TH2-bounded $O(m)$ walk)	39 / 50 (0.780)	39/11/0/0	Reference	−22.00 pts	0 / $0.00	Zero-call null
Classical B1 (CDCL+DRAT, zero store)	**50 / 50 (1.000)**	50/0/0/0	+22.00 pts	Reference (< 1 ms)	0 / $0.00	Exact solver ceiling
exp-002 Single-Turn Opus 5.5 (1,024-tok cap)	4 / 50 (0.080)	4/1/0/45	−70.00 pts [−82.0, −56.0]	−92.00 pts	50 / —	**Not assessable**
Single-Turn Opus 5.5 Ext. Thinking (8,192 tok)	**49 / 50 (0.980)**	49/1/0/0	**+20.00 pts [+8.0, +34.0]**	−2.00 pts [−12.0, +0.0]	50 / $2.14	**Relevance shown** (vs bounded N1)
Single-Turn xAI Grok 4.20 Reasoning (8,192 tok)	**39 / 50 (0.780)**	39/11/0/0	+0.00 pts [−18.0, +18.0]	−22.00 pts [−34.0, −12.0]	50 / $0.10	**Undetermined** (ties N1)
6-Turn Pure $O(m)$ Clause-Feedback Sonnet 5.5	**50 / 50 (1.000)**	50/0/0/0	**+22.00 pts [+12.0, +36.0]**	**+0.00 pts [+0.0, +0.0]**	54 (1.08 turns) / $1.98	**Relevance shown** (matches B1)
6-Turn Pure $O(m)$ Clause-Feedback Opus 5.5	**50 / 50 (1.000)**	50/0/0/0	**+22.00 pts [+12.0, +36.0]**	**+0.00 pts [+0.0, +0.0]**	50 (1.00 turn) / $2.15	**Relevance shown** (matches B1)

Raising reasoning headroom lifts `Opus 5.5` to `49/50` (`77,724` thinking tokens) and 6-turn $O(m)$ clause feedback lifts both `Sonnet 5.5` and `Opus 5.5` to `50/50` — beating the `TH2`-capped local walk (`39/50`), while standalone zero-store `CDCL+DRAT` (`B1`) already achieves `50/50` in under `1 ms` at `$0.00`.

---

5. Good–Turing Unseen Mass, Multi-Turn `/work` Repair, and Online Crystallization

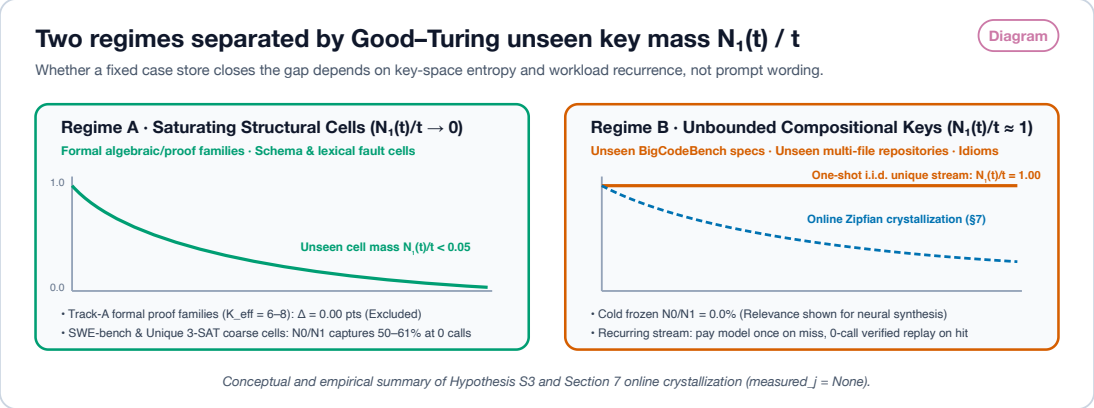

Figure 2. Good–Turing unseen key mass $N_1(t)$ / t predicts where frozen `Shape D` saturates versus where neural synthesis or online crystallization is required (`measured_j = None`).

5.1 Good–Turing unseen mass ($N_1(t)/t$) across the five Track-B streams

Along each `dev + cal` task stream $t = 1 \dots T$, the Good–Turing singleton ratio $N_1(t)/t$ estimates the probability that the next arriving key is unseen:

Task Class (`dev + cal` stream)	Stream Length T	Coarse Cell $N_1(T)/T$	Fine Schema/Instance $N_1(T)/T$	Held-Out `N0` / `N1` Completion
`formal.unique_3sat_drat`	270	`0.0000`	`1.0000`	69.6% / 69.6%
`external.swebench_verified`	250	`0.0080`	`0.5960`	46.0% / 45.6%
`external.tau_bench_retail`	285	`0.0035`	`0.6386`	39.6% / 25.9%
`external.bfcl_v3_live`	300	`0.0233`	`0.4033`	20.0% / 22.8%
`external.bigcodebench_exec`	240	**`0.9917`**	**`0.9917`**	**0.0% / 0.0%**

Even after `N1` crystallized `108` unit-tested `Claude Opus 5.5` Python solutions on `BigCodeBench` dev, $N_1(T)/T = 0.9917$ predicted near-zero key overlap on unseen compositional coding specifications — matching the observed `0.0%` held-out `test` completion (`218` unfilled keys).

5.2 Multi-turn `/work` software-engineering repair (6 multi-file repositories)

Two complementary multi-turn `/work` bug-repair suites (`max_turns = 10`, restoring pristine upstream test files before `pytest` grading) verify repository-level patch synthesis and `0`-call crystallized replay:

1. **Controlled 6-repository multi-file Python suite**: Six multi-function repositories (`config-defaults` spanning `loader.py` + `schema.py`, `interval-merge`, `pagination`, `precedence` spanning `calc.py` + `lexer.py`, `rate-window` spanning `window.py` + `limiter.py`, and `retry-backoff`).
2. **Closure C (`exp-003`) — Real upstream `SWE-bench Verified` GitHub defect commits (`n = 6`, seeded from `drand 6664089`)**: Six upstream commits across four open-source repositories — `psf/requests-5414`, `pytest-dev/pytest-10051`, `pytest-dev/pytest-10356`, `mwaskom/seaborn-3069`, `sympy/sympy-23824`, and `sympy/sympy-17655` — each verified offline to fail `FAIL_TO_PASS` pre-patch and pass both `FAIL_TO_PASS` and `PASS_TO_PASS` under gold patches.

Multi-Turn `/work` Suite & Harness Lane	Repos Passed	SATB P/F/R/K	Model Calls	Tool Calls	API USD / Local reported_j
Controlled 6-Repo · Shape D Cold (t = 1)	0 / 6 (0.0%)	0/6/0/0	0	0	$0.0000 / < 1 J
Controlled 6-Repo · Lane B Claude Opus 5.5	**6 / 6 (100.0%)**	6/0/0/0	28	29	$0.3613 (not_available)
Controlled 6-Repo · Lane B Claude Sonnet 5.5	**6 / 6 (100.0%)**	6/0/0/0	31	34	$0.2080 (not_available)
Controlled 6-Repo · Shape H Claude Code 2.1.330	**6 / 6 (100.0%)**	6/0/0/0	30	MCP	$0.2646 (Falsifier 8 passed)
Controlled 6-Repo · Lane B Local Qwen 3.6	**6 / 6 (100.0%)**	6/0/0/0	42	42	**2,342.0 J** L2 (390.3 J/repo)
Controlled 6-Repo · Shape D Replay (t = 2)	**6 / 6 (100.0%)**	6/0/0/0	0	0	$0.0000 (verified patch replay)
Real SWE-bench (exp-003) · Shape D Cold (t = 1)	0 / 6 (0.0%)	0/6/0/0	0	0	$0.0000 (unseen repo trees)
Real SWE-bench (exp-003) · Lane B Claude Opus 5.5	**5 / 6 (83.3%)**	5/1/0/0	33	45	$0.8557 (not_available)
Real SWE-bench (exp-003) · Lane B Claude Sonnet 5.5	**4 / 6 (66.7%)**	4/2/0/0	31	46	$0.5865 (not_available)
Real SWE-bench (exp-003) · Lane B xAI Grok 4.20	1 / 6 (16.7%)	1/5/0/0	60	51	$0.8480 (not_available)
Real SWE-bench (exp-003) · **Shape D Replay (t = 2)**	**5 / 6 (83.3%)**	5/1/0/0	**0**	**0**	**$0.0000** (17.1 s pytest re-verification)

On real upstream SWE-bench Verified repositories, Claude Opus 5.5 synthesized valid patches passing full FAIL_TO_PASS and PASS_TO_PASS pytest suites on requests-5414, pytest-10051, seaborn-3069, sympy-23824, and sympy-17655 (5/6, failing only pytest-10356), and Shape D online crystallization replayed and re-verified all 5/6 repository fixes on t = 2 in 17.1 s at **0 model calls and $0.00**.

5.3 Online crystallization on recurring Zipfian traffic (BigCodeBench)

On the secondary online track (600 requests drawn from a Zipf-0.85 distribution over the 220 unique BigCodeBench test tasks, where Shape D invokes B3/B6 only on an unseen key miss and caches solutions that pass unittest), completion matches pure per-request Opus 5.5 (**98.0% vs. 98.0%**) while 155 verified functions crystallize into the store, cutting cumulative LLM calls by **72.17% (167 vs. 600)**, API spend from **$7.33 to $2.04 (3.59×)**, and local Qwen 3.6 reported_j spend from **67,729.6 J to 18,851.4 J (3.59×)**.

---

6. Limitations, Falsifiers, and Unfinished Work

1. **No physical rail or AC wall-plug power meter (board_metered = false, measured_j = None)**: All energy figures in this study are Apple Silicon on-chip IOReport (L2) and kernel ri_energy_nj (L3) reported_j counters on a host running concurrent jobs; L0 AC wall-plug measurements (Shelly/Yokogawa, capturing PSU, DRAM, and NVMe losses) on a quiesced target machine remain open.
2. **Self-administered benchmark selection vs. human outside-party Tier 3**: Although Track-B (exp-002) and exp-003 evaluated externally authored public benchmarks and upstream GitHub commits locked before drand round 6664089, a sealed Tier-3 benchmark authored and held in escrow by an independent human third party has not yet been run.
3. **Finite-budget non-necessity**: Every "Relevance shown" verdict is strictly relative to the declared dev sample (100–150 instances) and 4,096 B function-table cap; as Closure B (formal.unique_3sat_drat) illustrates directly, a zero-store classical CDCL+DRAT solver (B1) solves 100% of instances where a TH2-capped local walk stops at 78.0%.

References

1. Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1), 39–59.
2. Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185.
3. Goldberg, E., & Novikov, Y. (2003). Verification of proofs of unsatisfiability for CNF formulas. *DATE 2003*, 886–891.
4. Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4), 237–264.

5. Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can language models resolve real-world GitHub issues? *ICLR 2024*.
6. Mamdani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1), 1–13.
7. Michie, D., & Chambers, R. A. (1968). BOXES: An experiment in adaptive control. *Machine Intelligence*, 2, 137–152.
8. Open Interface Engineering (2026). Metered Commit Gates; Task Satiation vs. Aggregate Rebound; The Reality Code Engine; Downward LUT6 Compilation [same house]. https://research.openie.dev/
9. Patil, S. G., Mao, H., Cheng-Jie Ji, C., Yan, F., Suresh, V., Stoica, I., & Gonzalez, J. E. (2024). Gorilla: Large language model connected with massive APIs (Berkeley Function-Calling Leaderboard). *NeurIPS 2024*.
10. Simon, H. A., & Kadane, J. B. (1975). Optimal problem-solving search: All-or-none solutions. *Artificial Intelligence*, 6(3), 235–247.
11. Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, 29(12), 1213–1228.
12. Wang, L.-X., & Mendel, J. M. (1992). Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(6), 1414–1427.
13. Wetzler, N., Heule, M. J. H., & Hunt, W. A. (2014). DRAT-trim: Efficient checking and trimming using expressive clausal proofs. *SAT 2014*, 422–429.
14. Yao, S., Shinn, N., Razavi, P., & Narasimhan, K. (2024). tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv:2406.12045*.
15. Zhuo, T. Y., Vu, M. C., Chim, J., Hu, H., Yu, W., Widyasari, R., et al. (2025). BigCodeBench: Benchmarking code generation with diverse function calls and complex instructions. *ICLR 2025*.