

Supplement to: In a Search Loop on Circle Packing, a Claude Code Subagent Keeps or Improves Its Parent, Not Its Zero-Shot Anchor

Anonymous authors

Paper under double-blind review

Section numbers without an S prefix refer to the main paper; their hyperlinks target the file `main.pdf`, the main paper’s name inside the artifact’s `paper/` directory, and the original arm predicates and frozen reports. The narrative uses F_N for the scalar family maximum (historically called the “family argmax”) and A_N for the anchor value. Post hoc clearance re-reads and audits are identified separately, with their sources and denominators; they do not rewrite the frozen baseline. Facts attributed to the companion manuscript are cited to it (Anonymous authors, 2026).

S1 Ledger of the arms this paper reports

Every arm reported in this paper, with the section that reports it (S4 for the four other-model arms), its proposer, design, registered test and outcome. Design gives cells $\times$ slots per cell. Registered tests are coded P (prediction), F (falsifier) and S (secondary prediction), each expanded in its arm’s section. These are the original runs; S7 reports the later loop replication separately. This paper is their primary report; the companion manuscript cites these findings as supporting evidence. The artifact bundle serves both manuscripts (`BUNDLE_README.md`) and also contains registrations for arms B and B2 (a numerical optimizer run without a model), CC, CC2 (with CCS), CCP, CL, CL-W, CN, CP, M, O, PP, R, RP, S and T, plus several wave registrations (`wave*_prereg*`). None of them is reported in this manuscript or counted in its 1,255-slot inventory.

Table S1: The eight registered arms this paper reports.

Arm	§	Proposer(s)	Design	Registered test	Outcome
MU (conditioning)	4	agent runtime	3 parents $\times$ 3 $N \times 15$	P-MU1–P-MU4 + F-MU1	keep-or-improve 26/26 under the maximizing parent, which entails its 0/26 anchor-match rate: the dissolution rule is met and F-MU1, which uses the same condition, fires; P-MU2 ($\geq 35\%$ anchor matches there) and P-MU3 (k -inheritance below 50%; observed 26/26) are missed; anchor-match rate 7/28 under the anchor parent, below P-MU1’s registered 50%, so the anchor is not self-stable; P-MU4 (modal off-family output on the k^* grid) is met at $N = 21$ and 31 and missed at $N = 13$, where the seven valid outputs fall in seven score buckets, a tie counted against it. The ledger <code>arm_mu_collect.jsonl</code> also contains the 45 rows of arm CH (recipe-family choice probe), registered in the same file and commit, whose frozen verdict is that P-CH1 is met in 2 of 3 cells; CH is reported in the companion manuscript and is outside this paper’s eight arms and its 1,255-slot inventory
L (iterated loop)	5	agent runtime	2 $N \times 2$ archives $\times$ 6 generations $\times$ 5	L1–L4 + F-L1	L2 is met; L1 and L3 missed ; GREEDY’s descendant anchor-match rate is lower in both cells, contrary to L3’s per-cell rule; historical pooled Fisher $p = 0.303$, odds ratio 0.375, descriptive only; L4 open: 0/49 descendants clear under the original reported analysis

Arm	§	Proposer(s)	Design	Registered test	Outcome
P (pinned API)	6	pinned API	15 cells $\times$ 15	P-P1: $T(k^*, N)$ modal at ≥ 5 of the 7 square cells, including ≥ 2 of the 4 discriminating ones, and at ≥ 4 of the 5 held-out cells; P-P2: modal at ≤ 3 of the 7, leaving exactly 4 of 7 unadjudicated, a registration design gap; P-P3: no valid program output at $N = 13, 21, 31$ clears the family maximum (valid at 10^{-9} with excess $> 10^{-6}$); P-P4: at least one clears; F-P1: pooled anchor-match rate over the four discriminating square cells below 40% of valid outputs	validity collapses: 0/105 valid at the square cells, 4/75 held-out; every direct-emission prediction UNSCOREABLE (a post hoc scoreability reading, S5); P-P3 is met (0 of 12 valid programs clear); same-day runtime control 6/6 valid, anchor matches 4/6 (post hoc)
P-D (serving configurations)	6	pinned API, $N = 13$	2 conditions $\times$ 15	P-PD1 vs P-PD2 vs P-PD3 + S-PD1	P-PD1's numeric rule is met (closed at 27/30 rows by amendment, verdict unchangeable): thinking on without a system prompt 12/14 valid, anchor modal; thinking off with the system line 0/13; the two factors are confounded
GM (gemini-flash-lite)	S4	2nd vendor, 55 first-party + 85 routed rows	7 $N \times 20$	GM-registration thresholds	UNDERPOWERED (20 valid / 140) under both serving-path accountings
GM2 (gemma, 4k budget)	S4	2nd vendor	7 $N \times 20$	GM-registration thresholds	0/140 parseable; budget confound, diagnosed by GM3

Arm	§	Proposer(s)	Design	Registered test	Outcome
GM3 (gemma, 16k budget)	S4	gemma-4- 26b	7 $N \times 20$	P-GM3.1-.3 + falsifier	P-GM3.2 met (pooled 46/80 = 57.5%); P-GM3.1 missed (anchor modal in 2 of 5 scoreable cells); P-GM3.3 missed (20/46), since an anchor at a truncating cell has a single radius; falsifier not fired (3 mode failures, short of 4); post hoc, the discriminating cells give 11/43 = 26%, and the 27 outputs there at the family maximum value carry its (k^*-1) two-radius signature
V (cross-vendor zero-shot)	S4	13 free-tier aliases	5 $N \times 5$	floor + V1/V2	2 aliases meet the V1 pooled threshold at point estimate (8/12 and 11/18), gemma-4-31b does not (0/15), V2 met at point estimate at each scoreable alias (0/4, 0/9, 0/3; every interval includes the threshold)

Registrations and commits. Conditioning arm (MU): `arm_mu_preregistration.txt` with `arm_mu_prompts.json`, committed at `2b7d202` before sampling; decision thresholds, power analysis, tie conventions and falsifier all registered. The power analysis used a no-parent baseline from the companion study’s data, not this paper’s seeds. The same registration names arithmetic tractability as an alternative explanation and registers arm CH to probe it. Second-vendor arms (GM, GM2, GM3): `arm_gm_preregistration.txt` at `37b3adb`, `arm_gm2_preregistration.txt` at `3019aab`, `arm_gm3_preregistration.txt` at `c0b5b97`; the serving-path deviations `arm_gm_amendment1_openrouter.md` at `a075024` and `arm_gm3_amendment1_openrouter.md` at `67cc4a1` were each registered before the affected cells were sampled. Cross-vendor arm (V): `arm_v_preregistration.md` at `5444f10`, ancestry-checked by its runners before any sampling, with three amendments each committed before the sampling it governs; the second of them, `arm_v_preregistration_amendment2.md` at `d579797` (2026-08-12), added the three aliases that later reached the floor. Loop arm (L): registered at `c9645fe` and scored once at `750bb32`. Pinned-API arm (P): registered at `aa473ad` before sampling, with its scoring code committed afterwards at `596b113`. Serving-configuration arm (P-D): `arm_pd_preregistration.txt` at `5554cc0`, closed at 27 of 30 rows by amendment 1 at `1932cc0`. The packaged MU, P and P-D registration files each assert a push to a public remote (L’s does not); the anonymized package cannot demonstrate any push, so every “before sampling” above rests on author-recorded commit ancestry. Row timestamps: the P, P-D, GM, GM2 and GM3 ledgers carry a UTC timestamp per row (the P and P-D first rows fall 79 s and 21 s after the recorded dates of `aa473ad` and `5554cc0`, both local records); the replication logs request and response times per slot; the MU and L ledgers and the V ledger (a run date only) carry no row times, so their ordering rests on commit ancestry alone.

Arm L’s original decision rules. L3 compares rates within each cell, pooling only generations 1–5: GREEDY must have a strictly higher anchor-match rate than DIVERSE in at least one cell and a lower rate in neither. At $N = 13$ the rates are 0/4 versus 5/16; at $N = 31$ they are 3/18 versus 3/11. GREEDY is lower in both, so L3 is missed. The historical comparison pooled across cells, 3/22 versus 8/27 (odds ratio 0.375, two-sided Fisher $p = 0.303$), is retained as a descriptive summary, not as the registered test or calibrated evidence about archive width. Proposals share adaptive parent histories within lineages, and there is only one lineage per cell and archive regime. Treating all proposals as independent observations ignores that

dependence; parent identity also differs. L2 requires a lower descendant anchor-match rate than generation 0 in both cells. F-L1 instead requires a rate at least as high in both cells. It is not L2’s logical negation: one non-decrease already negates L2, whereas F-L1 requires two. Mixed directions occupy an adjudicated region. The observed decreases, 4/9 to 5/20 and 7/8 to 6/29, meet L2, and F-L1 does not fire.

What arm P-D identifies. The registration names P-PD1 “the reasoning budget is the cause,” but its numerical decision rule and that causal interpretation are distinct. D1 enables thinking with no system prompt; D2 disables thinking and adds the system line. The observed 12/14 versus 0/13 meets P-PD1’s numeric thresholds, and S-PD1 is met because the anchor is modal among D1’s valid outputs. P-PD2’s thresholds are missed and the remaining P-PD3 pattern is not selected. This two-configuration contrast cannot identify the effect of thinking separately from the system prompt or their interaction. Arm P’s earlier thinking-off, no-system rows and the post hoc runtime control provide context, not a contemporaneous factorial control.

Ledgers, scorers and reports. Conditioning arm: raw rows verbatim in `arm_mu_collect.jsonl`, scoring `arm_mu_analysis.py`, frozen output `arm_mu_results.txt`; the post hoc clearance recomputation is `arm_mu_ceiling.py`. Loop arm: `arm_l_collect.jsonl`, stepper `arm_l_step.py`, frozen report `arm_l_report.json`. Pinned-API arm: prompts and hashes `arm_p_prompts.json`, runner `arm_p_run.py`, scorer `arm_p_analysis.py`, report `arm_p_report.json`, ledger `arm_p_collect.jsonl` with 225 rows, six refused by the serving path; post hoc diagnostics `arm_p_diag_temperature.jsonl`, written by `arm_p_diag_temperature.py`, whose header labels it post hoc, and `arm_p_diag_runtime.jsonl`, labelled post hoc in its header row. Serving-configuration arm: builder `arm_pd_build.py`, spec `arm_pd_prompts.json`, runner `arm_pd_run.py`, scorer `arm_pd_analysis.py` importing the pinned-API scorer unmodified, ledger `arm_pd_collect.jsonl` with 27 rows plus the refusal rows kept for resume. The post hoc clearance re-read of arms GM3 and V is `arm_ceiling_gm3_v.py`. Second-vendor arms: `arm_gm_gm3_checkpoint.jsonl` carries every row’s raw API response verbatim with serving path and provider tagged, scored by `arm_gm3_analysis.py` into `arm_gm3_report.json` with both serving-path accountings. Cross-vendor arm: every call appended live to `arm_v_candidates_raw.jsonl`, scored by `arm_v_score.py`, which imports unchanged the companion manuscript’s parse-validate-classify scorer, `arm_f_repro.py`, shipped with this bundle. One script cited here, `arm_mu_ceiling.py`, lives in the paper repository rather than the corpus root and is shipped under `paper_repo/loop/`.

The independent re-score. `arm_transfer_independent_rescore.py` recounts the parse, validity, anchor-match and clearance counts of every direct-emission cell of the eight original arms from the raw ledgers, with its own parser, its own validity scorer and a linear-programming oracle blind to the recipe, importing nothing from the arms’ scoring code; its report is `arm_transfer_independent_rescore.json`. For arm P’s three program-channel cells it does not execute the ledger programs; it only checks the frozen report’s sums against the recomputed family maximum. It also does not recount some fields, such as MU’s keep-or-improve and k -inheritance and P’s modal values and failure bins. Closed-form values and family maxima agree to at most 3.6×10^{-15} , and of 135 cell-level counts, the three program cells included, 129 agree. Its parser was revised once after a first run, which raised agreement from 120 to 129 cells; the revision accepts whitespace between nested brackets and first tries `literal_eval` on the whole fence-stripped text (the report’s `parser_revision` block). The six that differ are these. Three arm L clearance counts at $N = 31$: a proposal sitting on the maximum to within 1.3×10^{-15} counts as clearing under “excess > 0” and not under “excess > 10^{-6} ”, and no substantive clearance verdict moves. The latter threshold must not be retroactively described as arm L’s registered rule. Two arm GM validity counts, at $N = 21$ (7 against 6) and $N = 35$ (3 against 2), where a balanced-bracket scan recovers packings from completions the vendor truncated mid-prose. And arm P-D’s sampled-row count for D2 (16 against 13), where three rows logged as call errors with no content are counted as sampled by the re-score. The $N = 35$ recount crosses the GM registration’s floor of 3, so that cell becomes scoreable on the re-score’s own count, while arm GM’s UNDERPOWERED verdict, which applies when fewer than five cells are scoreable, stands either way: three cells are scoreable in the frozen report and four on the re-score’s count. The re-score also splits arm MU’s excess buckets as 18 beyond rounding, 9 within it and 61 not exceeding, the counts the MU clearance re-read in S2 reports (artifact `corrections_ledger.md`, item 36).

Harmonized clearance (post hoc analysis). For cross-arm comparison only, let F_N be the exact closed-form family maximum and s the sum recomputed from emitted radii. A harmonized clearance requires *both* geometric validity at 10^{-9} and $s - F_N > 10^{-6}$. The geometric tolerance and the score margin are separate conditions; neither the 2×10^{-3} matching window nor a positive floating-point excess defines clearance. This common predicate is explicitly post hoc. Original arm analyses, denominators, predicates and registered misses remain labelled as such and unchanged; a historical count under a different predicate is not silently substituted by the harmonized count. Failure to clear is a score statement, not evidence of conformance to the recipe family. Matching the maximum score or a two-radius signature likewise does not establish conformance without a coordinate-level test.

The post hoc replay in `revision_replay.json` applies this predicate to every arm and records zero harmonized clearances in each, including below-floor arm V aliases. It sums emitted radii without geometric repair or recorded-score rounding. At the 10^{-6} geometric tolerance, three MU outputs exceed F_N by more than 10^{-6} ; all three fail the 10^{-9} gate. Tiny positive excesses are retained as diagnostics: one GM3 output has a positive floating-point excess even at the 10^{-9} gate, but it does not exceed the score margin. Historical statements of no clearance therefore must not be strengthened into claims of no positive excess whatsoever.

S2 Deviations from registration

The rows below preserve the original study deviation ledger, restricted to the arms this paper reports. Arm GM2 ran under its own registration (3019aab), which restates GM’s design and thresholds for Gemma; it has no deviation of its own and no row. Arms GM and GM3 each carry one registered serving-path amendment, and the GM3 row also discloses a ledger repair. A separate GM3 row records the observed timing call excluded from the scored ledger, as disclosed in its registration.

Table S2: Deviations from registration, for the arms this paper reports.

Prediction	Registered rule	As registered?	Outcome / reason
GM serving path	first-party API only	amendment registered before resumption (arm_gm_ amendment1_ openrouter.md, a075024, committed 19 s before the first routed row)	55/140 rows first-party, 85/140 via the routed path with the provider logged as the vendor; UNDERPOWERED under both accountings: all rows, 20 valid, 3 scoreable cells, 1 mode match, 3/20 anchor matches; first-party rows alone, 6 valid, 1 scoreable cell, 0 mode matches; the falsifier fires under neither (S4)
GM3 serving path	first-party API only	amendment registered before resumption (arm_gm3_ amendment1_ openrouter.md, 67cc4a1)	41/140 rows via routed path, provider logged; 59.0% / 57.5% dual accounting, no verdict change (S4); ledger repair at 31a7fc4: two concurrent runners with stale completion snapshots wrote 3 duplicate rows at $N = 31$ (call indices 6–8), removed from 93 rows to 90 before the arm was completed to 140 unique rows, the pre-repair file kept under a PREDEDUP suffix

continued on next page

Prediction	Registered rule	As registered?	Outcome / reason
Arm V proposer set	3 aliases served through the opcode gateway, a third-party free-tier routing service	expanded 3 $\rightarrow$ 13 by amendments 1–2, each registered before the sampling it governs	verdicts in S4; one of the thirteen aliases is served under a name no vendor attests, and amendment 1 discloses it as unattributed
Arm V transport repair	single pass, <code>max_tokens</code> 8192	amendment 3 registered before resumption (mechanical: 24,576 budget, retry, censored-row accounting)	877 attempt rows over the 325 registered slots (314 carry rows); floor evaluated per slot (S4)
GM3 timing call	140 scored calls	excluded diagnostic, disclosed in <code>arm_gm3_pre_registration.txt</code> before the run	one $N = 13$ bare-prompt timing call followed two timeout-caused false starts (120 s and 420 s). Its output was observed: a uniform-radius packing, with the tail inspected for format compliance. It is outside the ledger and all analyses; the design and prediction thresholds were unchanged. “Before sampling” for the scored run therefore does not mean before any GM3 output was observed
GM3 maximum-radius check	—	post hoc , disclosed (<code>diagnostics_gm3_rival.py</code>)	all 27 discriminating-cell packings at the family maximum value carry its two-radius signature (S4); the registered signature is scoped to k^* and cannot see $(k^* - 1)$ radii. A score and radius signature alone do not establish coordinate conformance in the recipe family

continued on next page

Prediction	Registered rule	As registered?	Outcome / reason
Arm L (iterated loop)	L1–L4	registered before sampling (c9645fe), scored once (750bb32); L1 uses four per-lineage thresholds (the registration text describes the modal-seed match in both cells; the committed scorer applies it per lineage), while L3 requires GREEDY higher in at least one cell and lower in neither; one post-sampling scorer correction (corrections ledger item 34) replaced a seven-decimal maximum with its closed form in the 10^{-9} clearance analysis: an apparent 3.7×10^{-8} excess at $N = 31$ became 1.3×10^{-15} ; two runtime rejections, registered as excluded and counted, are shown as invalid slots, and no rate among valid outputs changes	L2 met ; L1 missed on 13-DIVERSE, a 2–2 tie at $n = 4$ decided against the prediction by the registered tie rule; L3 missed in its original per-cell scope; historical pooled Fisher $p = 0.303$ retained descriptively with lineage dependence (S1); L4 open and answered under the original report: 0/49 descendants clear (§5)

continued on next page

Prediction	Registered rule	As registered?	Outcome / reason
Arm P (pinned API)	P-P1-P-P4 + F-P1	registered before sampling (aa473ad); motivation disclosed in the registration file	225 calls, 219 returning content and 6 refused by the serving path and recorded as such; validity collapsed to 0/105 at the square cells; direct-emission predictions unscorable under the registered floor (post hoc scoreability reading; the frozen scorer recorded P-P2 as met, S5); P-P3 is met (0/12); F-P1 denominator empty, reported as such; the frozen scorer’s P-P2 Boolean and [0, 0] interval are annotated in S5 (§6)
Pinned-API re-draws of MU and L	protocols in replication_mu/ and replication_l_api/	registered in local commits (2026-09-16 and 2026-09-17, UTC)	not run: the MU pilot’s one call was refused for lack of account credit and returned no output; the L re-draw made no call; no outputs exist and no result is reported
Arm P temperature diagnostic	—	post hoc , not preregistered	0/6 valid at temperature 0.0, 0.5 and 1.0, which is 0/18 across the three settings (§6)
Arm P runtime control	—	post hoc , not preregistered	same prompt through the agent runtime, same day: 6/6 valid, anchor matches 4/6 (§6)
Arm P-D (serving configurations)	P-PD1/P-PD2/P-PD3 + S-PD1	registered before sampling (5554cc0), after arm P’s validity collapse and temperature diagnostic (§3); run stopped at 27 completed slots (14 D1, 13 D2), with one unresolved refused slot and two unattempted slots; closed there by amendment 1 (arm_pd_amendment1_closed_at_27.md), since no completion could change the numeric verdict	P-PD1’s numeric rule is met and completing the three outstanding slots could not change it: thinking on, no system prompt, 12/14 valid, modal $T(4, 13)$; thinking off with system line, 0/13; S-PD1 met. Thinking and system prompt change together, so a causal component is not identified (§6)

continued on next page

Prediction	Registered rule	As registered?	Outcome / reason
F-MU1 consequence	if F-MU1 fires, report the loop-relevance framing as refuted for the one-parent regime, with title and abstract restricted to unconditioned calls	not as registered	F-MU1 fires (§4); the non-transfer to parent-conditioned calls is reported as a finding in the abstract and §4, but the title and abstract are not restricted to unconditioned calls
Arm MU clearance re-read	—	post hoc , not preregistered	the clearance question was never registered for arm MU; of 88 valid at 10^{-6} , 18 exceed the maximum beyond the recorded 5×10^{-8} rounding scale, by at most 1.53×10^{-6} , and 9 more exceed it within that scale. All 18 fail the 10^{-9} gate on overlap, and 13 return the maximizing parent’s rounded sum unchanged (per-row table below); the historical re-read (<code>arm_mu_ceiling.py</code>) reports 63 valid at 10^{-9} and no clearance; §7 and Table 7 give the harmonized replay
GM3 discriminating- cell split	—	post hoc , not preregistered	the registered thresholds were the pooled rate, the per-cell modal count, the two-radius-signature share and the falsifier; no threshold was registered on the discriminating cells alone; read post hoc, those three scoreable cells give $11/43 = 26\%$, while the two scoreable cells where the anchor equals the family maximum contain 35 of the 46 anchor-matching packings (S4)
Arm GM3 and arm V clearance re-read	—	post hoc , not preregistered	the clearance question was registered for neither arm; historical re-read: GM3 80 valid at 10^{-6} , 79 at 10^{-9} ; V’s three scoreable aliases 45 and 39, with no reported clearance. The harmonized predicate also yields none, including all thirteen V aliases; a tiny positive GM3 excess does not meet its score margin (S4) (<code>arm_ceiling_gm3_v.py</code> , <code>revision_replay.json</code>)

The eighteen arm MU outputs beyond the recorded rounding scale. One row per output, from the arm MU clearance re-read (`arm_mu_ceiling.py`, overlaps recomputed from the verbatim coordinates). Excess is the recorded sum minus the closed-form maximum. Largest overlap is the largest $r_i + r_j - d_{ij}$ over pairs, the quantity the 10^{-9} test fails on. In every row the excess exceeds the largest overlap. These are different quantities: one is a sum-of-radii difference and the other a single constraint violation. Thirteen rows return the maximizing parent’s seven-decimal sum unchanged and carry its rounding; the other five sit above or below that parent sum. That maximizing parent itself fails the 10^{-9} test at all three cells (largest

overlaps 1.34×10^{-7} , 4.7×10^{-9} and 4.38×10^{-8} at $N = 13, 21$ and 31), and 4 of its 26 outputs valid at 10^{-6} pass 10^{-9} , against 26 of 28 under the anchor parent and 33 of 34 under the off-family parent. The recorded seven-decimal sums have a 5×10^{-8} rounding scale. These eighteen sit above the maximum by more than that scale. Nine further rows of the 88 valid at 10^{-6} exceed it only within that scale, and the remaining 61 do not. Thus 27 recorded sums are strictly positive excesses; 18 is the count beyond rounding, not the total positive count.

These are buckets of the *recorded seven-decimal sums*. Summing the original radii without that rounding in `revision_replay.json` gives 23 positive excesses among MU’s 88 outputs valid at 10^{-6} . The difference reflects the precision of the quantity being counted; it does not overwrite the historical 18/9/61 recorded-score partition. Only three of the full-precision excesses exceed 10^{-6} , and none passes the 10^{-9} gate.

The registrations use the parent codes A_anchor (anchor parent), B_rival (maximizing parent) and C_offfamily (off-family parent) and call an anchor match “on-prediction”; the per-row table below uses the first two codes.

N	Parent	Slot	Recorded sum	Excess over F_N	Largest overlap	Relation to parent sum
13	B_rival	1	1.7761427	3.25×10^{-7}	1.34×10^{-7}	equals parent sum
13	B_rival	3	1.7761427	3.25×10^{-7}	1.34×10^{-7}	equals parent sum
13	B_rival	5	1.7761435	1.13×10^{-6}	3.34×10^{-7}	above parent
13	B_rival	7	1.7761439	1.53×10^{-6}	4.34×10^{-7}	above parent
13	B_rival	9	1.7761427	3.25×10^{-7}	1.34×10^{-7}	equals parent sum
13	B_rival	13	1.7761427	3.25×10^{-7}	1.34×10^{-7}	equals parent sum
21	A_anchor	7	2.2588850	1.52×10^{-6}	3.05×10^{-7}	above parent
21	B_rival	11	2.2588840	5.24×10^{-7}	1.05×10^{-7}	above parent
31	B_rival	2	2.7485284	2.63×10^{-7}	4.38×10^{-8}	equals parent sum
31	B_rival	3	2.7485284	2.63×10^{-7}	4.38×10^{-8}	equals parent sum
31	B_rival	4	2.7485284	2.63×10^{-7}	4.38×10^{-8}	equals parent sum
31	B_rival	6	2.7485284	2.63×10^{-7}	4.38×10^{-8}	equals parent sum
31	B_rival	7	2.7485284	2.63×10^{-7}	4.38×10^{-8}	equals parent sum
31	B_rival	8	2.7485284	2.63×10^{-7}	4.38×10^{-8}	equals parent sum
31	B_rival	10	2.7485284	2.63×10^{-7}	4.38×10^{-8}	equals parent sum
31	B_rival	11	2.7485284	2.63×10^{-7}	4.38×10^{-8}	equals parent sum
31	B_rival	12	2.7485282	6.26×10^{-8}	3.76×10^{-9}	below parent
31	B_rival	13	2.7485284	2.63×10^{-7}	4.38×10^{-8}	equals parent sum

S3 Prompts and hashes

S3.1 Bare prompt ($\{n\}$ parameterized; the $N = 13$ instance hashes to 32db485bea625ff9f39f4723ebf1a01f337559a9e2cf567fb486928f71f7f8df):

Pack exactly $\{n\}$ non-overlapping circles inside the unit square $[0,1] \times [0,1]$ so that the sum of their radii is as large as possible.
 MUST hold: exactly $\{n\}$ circles; every circle fully inside the unit square $[0,1] \times [0,1]$ ($x-r \geq 0$, $x+r \leq 1$, same for y); no two circles overlap (distance between centers $\geq$ sum of the two radii; touching is allowed).
 Do not write or execute code - construct the packing by reasoning alone.
 Output ONLY the raw Python list of $\{n\}$ $[x, y, r]$ lists. No explanation, no code fences, no other text.

S3.2 Registered prompt hashes. bare $N=13$ 32db485b..., $N=17$ 8437df75..., $N=21$ a415425b..., $N=31$ a664d003..., $N=35$ 3b08c56e..., $N=37$ 2c0a8819.... The $N=43$ bare hash 1208e7d2... first appears in a registration in arm GM’s (37b3adb), and GM2 and GM3 adopt it. Arm P’s registration states that $N=43$ had never carried a registered hash; that statement is incorrect. The second-vendor arms and the cross-vendor arm send these same prompts byte-identical; every ledger row’s logged SHA-256 matches the list

above. The GM registration restates the hashes ($N = 35$ and 37 by reference to `arm_f_prompts.json`), GM2 and GM3 adopt GM’s list, and V’s restates those of its five cells. The runners recompute the hashes of the cells in `arm_f_prompts.json` before sampling and abort on mismatch; at $N = 21$ and 43 the GM runners log the hash without checking it. The conditioning arm’s mutation prompt is hashed in `arm_mu_prompts.json` and the loop arm reuses it verbatim for generations 1 to 5.

S3.3 Dispatch wrapper (the agent-runtime arms only, and not part of the hashed prompt): **Do not use any tools. Your entire final message must be the answer and nothing else.**

S3.4 Parser. Completions are parsed with `ast.literal_eval` after stripping code fences and taking the outermost bracketed span. A completion that does not yield a list of numeric $[x, y, r]$ triples is scored invalid, and its failure mode is logged (`no_output`, `no_bracketed_list`, `literal_eval_failed`, `row_not_xyr` and similar); a list with the wrong count fails validation as `count_m_expected_N`. Radii must be strictly positive. Containment and pairwise non-overlap are checked at the two registered tolerances.

S4 Second-vendor and cross-vendor arms: results and per-cell tables

Arms GM, GM2, GM3 and V: designs and registered results. The second-vendor results pair a met pooled-rate prediction with anchor-mode misses; §6 summarizes all four arms.

GM and GM2. Arms GM, GM2 and GM3 share one design under three registrations: seven cells with 20 calls each. Each predicts that the modal output matches the anchor in at least five scoreable cells, a pooled anchor-match rate of at least 30%, and the k^* two-radius signature in at least half of anchor-matching outputs. That signature pairs grid radius $1/(2k^*)$ with filler radius $(\sqrt{2} - 1)/(2k^*)$. The falsifier requires at least four scoreable mode failures. A cell is scoreable with at least three valid outputs, and an arm verdict needs five scoreable cells. GM, a gemini-flash-lite alias, returns 20 outputs valid at 10^{-6} out of 140 and only three scoreable cells: UNDERPOWERED. Its 55 first-party rows give six valid outputs and one scoreable cell. The falsifier fires under neither accounting, and an alternative parser adds two valid lists without changing the verdict (S1). GM2 gives Gemma-4-26b a 4,096-token output budget, which deliberation exhausts: zero of 140 responses parse. GM3 repeats the design with 16,384 tokens, an adaptation registered before the scored rerun.

GM3 registered results. GM3 meets the pooled-rate prediction and misses the other two. Of 140 rows, 123 parse and 80 are valid at 10^{-6} . A quota amendment substitutes 41 routed rows for first-party calls; they contribute two valid outputs and no anchor matches. The pooled anchor-match count is 46/80 (57.5%), above the registered 30% threshold; first-party-only it is 46/78 (59.0%). The per-cell anchor-mode prediction is missed, with matches in two of five scoreable cells (Figure S1). The falsifier does not fire (three failures, short of four). The radius-signature prediction is missed at 20/46. A truncated anchor has a single radius and cannot carry the signature, so the 15 anchor-matching outputs at $N = 35$ count against it.

Discriminating split (post hoc). Most anchor matches fall where the anchor equals the family maximum: 35 of the 46 lie at $N = 17$ and 35 . The three scoreable discriminating cells give 11/43 (26%), below 30%, though no threshold was registered for this subset. At each of these cells the modal score matches the family maximum. A post hoc radius diagnostic finds the two-radius decomposition of $k^* - 1$, not the registered k^* , in all 27 outputs matching that score. Radius signatures alone do not certify coordinate conformance. The coordinate audit finds conformance at 10^{-6} in 26 of these 27. The other rounds its filler radii beyond that tolerance (§7; S6).

V: thirteen free-tier aliases. Three of the thirteen aliases reach the floor, all added by amendment 2; the ten below it include all three originally proposed. Each alias has 25 registered slots: five cells ($N = 13, 17, 31, 35, 37$) with five slots each. An alias reaches the floor with at least ten valid slots. Eight aliases use the opencode gateway’s free tier and five use OpenRouter’s free tier. Three amendments expanded the design from three aliases to thirteen and changed transport handling and output budget. The ledger distinguishes 877 attempts from 325 slots (314 carry a row). Transport retries follow the amended rule, not score-based selection among valid packings (below).

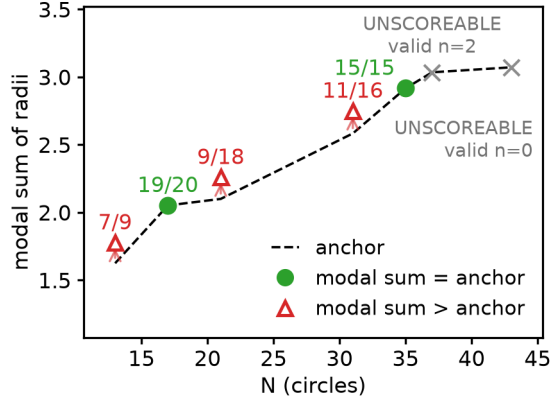


Figure S1: *Arm GM3 modal sums against the anchor.* The dashed anchor line is $V(k^*, m)$ where the anchor grid extends and $T(k^*, N)$ where it truncates. Filled green circles match it, with frequencies annotated. Open red triangles miss, all landing above the anchor at the family maximum *value*. Grey crosses are the two UNSCOREABLE cells, below the mode-analysis floor, with their valid counts. Discriminating cells ($A_N \neq F_N$) are $N = 13, 21, 31$ and 43 ; $N = 43$ is unscoreable.

At point estimates, two aliases meet the anchor-majority prediction (V1) and all three meet the family-maximum-rarity prediction (V2; Table S3). The first requires at least half of a scoreable alias’s valid outputs to match the anchor. The two met aliases’ intervals include 50%. Restricted to discriminating cells, north-mini-code gives 3/4 and gpt-oss-20b 5/9. The registration’s TRANSFERS label means only that this pooled threshold is met, not that an alias’s coordinate distribution replicates another model.

Table S3: Arm V’s three scoreable aliases. Columns score the anchor-majority prediction (V1) and, at the discriminating cells only, the family-maximum-rarity prediction (V2). Bracketed ranges are 95% Wilson intervals on the rate they follow.

alias	vendor	valid	V1 anchor matches	V1 threshold ($\geq 50\%$)	V2 family-maximum matches / valid at discriminating cells ($\leq 10\%$)
north-mini-code	Cohere	12	8/12 = 67% [39–86%]	MET (point est.)	0/4 [0–49%] MET
gpt-oss-20b	OpenAI	18	11/18 = 61% [39–80%]	MET (point est.)	0/9 [0–30%] MET
gemma-4-31b-it	Google	15	0/15 = 0% [0–20%]	MISSED	0/3 [0–56%] MET

V2 allows at most 10% of valid outputs to match the family maximum at the discriminating cells 13 and 31. All three aliases have zero such matches, but all three Wilson intervals include the 10% threshold.

A post hoc template-shape null (testing anchor matches against uniform choice among nearby grid orders; S5) changes no registered verdict.

No V output clears the family maximum. At discriminating cells, the closest score is 2.5×10^{-3} below F_{31} . Non-discriminating cells include matches to the maximum within 10^{-9} and a 1.2×10^{-4} gap at $N = 37$. Gemma-4-31b matches neither anchor nor family maximum in its scoreable V outputs, whereas GM3’s Gemma modes match the family maximum at discriminating cells; model variant and serving path change together.

The GM registration’s verdict table. The registered test column gives the thresholds that the GM, GM2 and GM3 registrations share. Arm V’s per-alias counts and intervals are in Table S3; its row here gives the registered tests, the threshold outcomes and the registration’s own labels.

Vendor	Model	Serving path	Registered test	Verdict
Google	gemma-4-26b	first-party API, 7 $N \times 20$ (41/140 rows routed under amendment; dual accounting)	pooled $\geq 30\%$; cell-level ≥ 5 ; falsifier ≥ 4 fails	pooled threshold met (57.5%); discriminating cells 11/43 = 26%, below the threshold (post hoc split, no threshold registered on those cells); cell-level short (modal in 2 of the 5 scoreable cells, against a threshold of 5; the falsifier did not fire, needing 4 failures); 27/43 match the family maximum value and its radius signature, which does not establish coordinate conformance; the structural threshold cannot be met at a truncate cell, whose anchor carries a single radius
Cohere; OpenAI (open- weight); Google	north-mini- code; gpt-oss-20b; gemma-4-31b	free-tier alias, 5 $N \times 5$	V1 anchor matches $\geq 50\%$ of valid; V2 family-maximum matches among valid outputs at discriminating cells $\leq 10\%$	Counts and intervals in Table S3. V1 threshold MET at point estimate for north-mini-code and gpt-oss-20b (both intervals include the threshold), the registration’s label TRANSFERS; MISSED for gemma-4-31b, the registration’s label DOES-NOT-TRANSFER. V2 point estimate met for all three
10 further aliases	(NVIDIA, Liquid, Poolside, inclusionAI, others; one unattributed)	free-tier	floor: ≥ 10 of 25 slots valid	BELOW-FLOOR; failure taxonomies logged, no claim in either direction

Arm GM3 per cell. Recomputable from `arm_gm3_checkpoint.jsonl` by `arm_gm3_analysis.py` with no arguments.

N	anchor (4 dp)	Gemma modal sum	mode freq	valid	anchor matches	MODE-MATCH
13	1.6250	1.7761	7/9	9	0	no (misses upward)
17	2.0518	2.0518	19/20	20	20	yes
21	2.1000	2.2589	9/18	18	8	no (misses upward)
31	2.5833	2.7485	11/16	16	3	no (misses upward)
35	2.9167	2.9167	15/15	15	15	yes
37	3.0345	—	—	2	0	UNSCOREABLE
43	3.0714	—	—	0	0	UNSCOREABLE

The two scoreable non-discriminating cells are $N = 17$ and $N = 35$, and 35 of the 46 anchor-matching packings sit in them. The three scoreable discriminating cells are $N = 13$, 21 and 31, where the pooled rate, read post hoc, is 11 of 43. All 20 two-radius signatures come from $N = 17$, the one extend cell among the matches, while $N = 35$ ’s 15 anchor-matching truncations carry no filler to detect.

Clearance of the family maximum (post hoc, disclosed). The rows of arms GM3 and V were re-scored by `arm_ceiling_gm3_v.py`, which imports the GM registration’s scorer and carries the closed-form maximum that `arm_mu_ceiling.py` uses. Of arm GM3’s 80 valid packings, 79 are valid at 10^{-9} . The best sum at each of its five scoreable cells equals the family maximum to within 10^{-9} ; the historical re-read

reports no clearance. The harmonized predicate likewise finds none. Arm V’s three scoreable aliases give 45 valid outputs at 10^{-6} (per alias in Table S3). Of those, 39 are valid at 10^{-9} , and none clears the maximum under that historical analysis. The closest result below the maximum among the discriminating cells is gpt-oss-20b, short by 2.5×10^{-3} at $N = 31$. Its gap at the non-discriminating $N = 37$ is smaller, 1.2×10^{-4} , so the former is not the closest gap over all cells. Its best at $N = 35$, and north-mini-code’s at $N = 17$, equal the maximum to within 10^{-9} .

Arm V per model. The three aliases that reached the registered floor are reported above. The other ten finished BELOW-FLOOR and claim nothing in either direction. Transport failures dominate for five of the eight opencode aliases. For deepseek-v4-flash-free and nemotron-3-ultra-free, about half of the ledger rows are transport failures and the rest are responses that fail to parse; for nemotron-3.5-lightning-free, 14 of 38 rows are transport failures and all 24 answered slots fail to parse. Of the two OpenRouter aliases below the floor, Liquid’s 25 completed slots carry no answer content, and 21 of NVIDIA’s 25 answered slots fail to parse. Retries mean the ledger contains far more attempt rows than slots: gemma-4-31b’s 25 slots took 190 attempt rows, 165 of them transport failures, and the floor is evaluated on the 25 slots.

Arm V in full (post hoc replay, all thirteen aliases). Table S4 is generated from the cell rows of `revision_replay.json`, produced by `revision_replay.py`. It includes all 65 model–cell combinations, including the ten aliases below the registered floor. Each cell has five planned slots. A slot with no ledger row is unattempted, not an imputed transport failure. The replay selects the first nonempty answer in source order; otherwise it retains the first non-error response, or the first failed attempt. A completed slot has a nonempty answer or a response without an explicit error; for CLI rows this is a ledger proxy, not provider attestation.

Columns give attempted slots (Att.), ledger attempt rows (Rows), completed slots (Comp.), slots with answer content (Ans.), valid counts at 10^{-6} and 10^{-9} (v_6 , v_9), anchor-match counts among outputs valid at 10^{-6} (Hit, $|s - A_N| < 2 \times 10^{-3}$), the best sum among outputs valid at 10^{-6} (Best), and harmonized clearances (C_9 : geometry 10^{-9} and score excess $> 10^{-6}$). A dash means no valid score. These post hoc counts do not replace the frozen V1/V2 verdicts; the alias floor remains at least 10 outputs valid at 10^{-6} over 25 planned slots.

Table S4: All arm V cells under the post hoc replay. Alias headings retain exact serving identifiers and show the original floor classification. Best sums are displayed to seven decimals; clearance uses the full-precision sum.

N	Att.	Rows	Comp.	Ans.	v_6	v_9	Hit	Best	C_9
cohere/north-mini-code:free: 12/25 valid, floor met									
13	5	34	5	2	2	1	1	1.6250000	0
17	5	25	5	3	3	3	2	2.0517767	0
31	5	14	5	5	2	1	2	2.5833333	0
35	5	9	5	5	3	3	3	2.9166667	0
37	5	7	5	5	2	1	0	2.6428571	0
google/gemma-4-31b-it:free: 15/25 valid, floor met									
13	5	40	5	5	2	2	0	1.3000000	0
17	5	37	5	5	5	5	0	1.7000000	0
31	5	32	5	5	1	1	0	1.5500000	0
35	5	38	5	5	4	4	0	0.8750000	0
37	5	43	5	5	3	3	0	1.8500000	0

continued on next page

N	Att.	Rows	Comp.	Ans.	v_6	v_9	Hit	Best	C_9
liquid/lfm-2.5-2.6b:free: 0/25 valid, BELOW-FLOOR									
13	5	26	5	0	0	0	0	—	0
17	5	22	5	0	0	0	0	—	0
31	5	20	5	0	0	0	0	—	0
35	5	20	5	0	0	0	0	—	0
37	5	20	5	0	0	0	0	—	0
nvidia/nemotron-3-ultra-550b-a55b:free: 3/25 valid, BELOW-FLOOR									
13	5	16	5	5	0	0	0	—	0
17	5	11	5	5	1	1	1	2.0517767	0
31	5	10	5	5	1	1	1	2.5833333	0
35	5	10	5	5	0	0	0	—	0
37	5	9	5	5	1	1	1	3.0345178	0
openai/gpt-oss-20b:free: 18/25 valid, floor met									
13	5	19	5	5	5	5	5	1.6250000	0
17	5	32	5	2	0	0	0	—	0
31	5	7	5	5	4	4	0	2.7460000	0
35	5	5	5	5	5	5	5	2.9166667	0
37	5	14	5	5	4	1	1	3.0344000	0
opencode/big-pickle: 0/25 valid, BELOW-FLOOR									
13	5	8	0	0	0	0	0	—	0
17	5	5	0	0	0	0	0	—	0
31	5	5	0	0	0	0	0	—	0
35	2	2	0	0	0	0	0	—	0
37	0	0	0	0	0	0	0	—	0
opencode/deepseek-v4-flash-free: 0/25 valid, BELOW-FLOOR									
13	5	12	4	4	0	0	0	—	0
17	5	8	5	5	0	0	0	—	0
31	5	6	5	5	0	0	0	—	0
35	5	13	4	4	0	0	0	—	0
37	5	8	3	3	0	0	0	—	0
opencode/hy3-free: 0/25 valid, BELOW-FLOOR									
13	5	15	0	0	0	0	0	—	0
17	5	11	0	0	0	0	0	—	0
31	5	10	0	0	0	0	0	—	0
35	5	10	0	0	0	0	0	—	0
37	5	5	0	0	0	0	0	—	0
opencode/laguna-s-2.1-free: 0/25 valid, BELOW-FLOOR									
13	5	22	1	1	0	0	0	—	0
17	5	13	2	2	0	0	0	—	0
31	5	6	0	0	0	0	0	—	0
35	5	5	0	0	0	0	0	—	0
37	2	2	0	0	0	0	0	—	0
opencode/ling-3.0-tiny-free: 0/25 valid, BELOW-FLOOR									
13	5	16	2	2	0	0	0	—	0
17	5	17	1	1	0	0	0	—	0
31	5	15	1	1	0	0	0	—	0
35	5	15	0	0	0	0	0	—	0
37	5	15	0	0	0	0	0	—	0

continued on next page

N	Att.	Rows	Comp.	Ans.	v_6	v_9	Hit	Best	C_9
opcode/mimo-v2.5-free: 0/25 valid, BELOW-FLOOR									
13	5	9	4	4	0	0	0	—	0
17	5	8	2	2	0	0	0	—	0
31	5	5	0	0	0	0	0	—	0
35	5	5	0	0	0	0	0	—	0
37	5	5	0	0	0	0	0	—	0
opcode/nemotron-3-ultra-free: 0/25 valid, BELOW-FLOOR									
13	5	13	5	5	0	0	0	—	0
17	5	9	5	5	0	0	0	—	0
31	5	9	4	4	0	0	0	—	0
35	5	7	4	4	0	0	0	—	0
37	5	5	5	5	0	0	0	—	0
opcode/nemotron-3.5-lightning-free: 0/25 valid, BELOW-FLOOR									
13	5	8	4	4	0	0	0	—	0
17	5	9	5	5	0	0	0	—	0
31	5	8	5	5	0	0	0	—	0
35	5	6	5	5	0	0	0	—	0
37	5	7	5	5	0	0	0	—	0

Across all thirteen aliases the replay gives 48 outputs valid at 10^{-6} and 42 at 10^{-9} , with 22 anchor-matching outputs valid at 10^{-6} and no harmonized clearances. The three scoreable aliases account for 45 and 39 valid outputs respectively; the other three valid outputs belong to the below-floor NVIDIA alias. Its three anchor-matching outputs are descriptive only and do not confer scoreability. Across 325 planned slots, 314 have ledger rows and 11 are unattempted. The 877 ledger attempt rows include 450 empty-output transport-failure rows, with four additional error-tagged rows carrying content and retained for scoring. Embedded retry counters are reported separately in the replay; these ledger-row counts are not an exhaustive count of underlying HTTP requests.

S5 Per-cell tables for arms P and P-D, and the template-shape null

The rows below report arms P and P-D per cell so that this submission can be checked on its own. This paper is their primary report (§3); the companion manuscript cites the findings as supporting evidence (Anonymous authors, 2026). These tables preserve the frozen counts.

Arm P per cell. Fifteen calls per cell. Refused rows are serving-path refusals and are counted invalid at the cell they were issued for. The seven square cells are the companion manuscript’s original zero-shot cells; the five held-out cells were excluded from its fit. This table retains the frozen 10^{-6} validity counts. The harmonized replay reproduces these 10^{-6} counts. At 10^{-9} it keeps 3 of the 4 valid programs at code cell $N = 13$ (the frozen report logs no 10^{-9} count for program cells), giving 15 valid at 10^{-9} outputs overall; at the direct-emission cells its 10^{-9} counts equal the frozen report’s.

channel	N	sampled	returned	refused	valid	note
square	13	15	15	0	0	all failures overlap
square	17	15	15	0	0	all failures overlap
square	21	15	15	0	0	all failures overlap
square	31	15	15	0	0	all failures overlap
square	35	15	15	0	0	one non-positive radius, rest overlap
square	37	15	15	0	0	one count mismatch, rest overlap
square	43	15	15	0	0	one malformed row, rest overlap
held-out	50	15	15	0	4	all four sum to 2.5, family maximum 3.5295867
held-out	58	15	15	0	0	count mismatches dominate
held-out	62	15	15	0	0	count mismatches dominate
held-out	65	15	15	0	0	count mismatches and overlaps
held-out	75	15	14	1	0	count mismatches and one refusal
code	13	15	13	2	4	below the floor of five
code	21	15	15	0	6	the one code cell above the floor
code	31	15	12	3	2	below the floor of five
total	—	225	219	6	16	0 of 105 square, 4 of 75 held-out, 12 of 45 code

No valid output at any cell clears the family maximum. The code cells were executed under a 10-second timeout, and P-P3 is met on their 12 valid programs.

Frozen scorer output for P-P2 and F-P1 (annotation; the report is unchanged). The frozen scorer `arm_p_analysis.py` wrote `arm_p_report.json` with `"evaluable": false` and `"hit": null` at every square and held-out cell, yet its `predictions` block records P-P2 (modal at ≤ 3 of 7 square cells) with `"observed": "square 0/7"` and `"holds": true`, and its `pooled_discriminating` block records `"valid": 0`, `"on_prediction": 0`, `"rate": 0.0` and `"wilson_95ci": [0.0, 0.0]`; F-P1 is recorded as `"fires": false` with observed `rate=0.0000 (0/0) 95%CI=(0.0, 0.0)`. The Boolean arises because a cell below the registered floor contributes no modal match, so zero matches satisfy “at most three” vacuously (scorer lines 191–196); the interval arises because the Wilson helper returns (0,0) when $n = 0$ (`arm_cn_analysis.py`, line 22). F-P1’s `fires` is correctly guarded by `valid > 0` (line 203); only its rate and interval misrepresent 0/0. The registration fixed a scoreability floor of five valid outputs per cell; no direct-emission cell reaches it, so P-P1 and P-P2 are unscoreable and the F-P1 rate is undefined. That is how Table 1, §6, and Tables S1 and S2 report them; it is a post hoc scoreability correction made in the manuscript, not the frozen scorer’s verdict. The report is preserved byte for byte (sha256 71167d34...) and `artifact/arm_p_report_NOTE.md` carries the same annotation. P-P3 and P-P4 are unaffected.

Arm P-D per cell. One cell, $N = 13$, two conditions, closed at 27 completed slots by a registered amendment after credit refusals. The ledger has three refused attempts, including one later-completed slot and two attempts on a single unresolved slot; two further planned slots were unattempted. Thus three refused attempts are not three distinct refused slots. The third row is arm P’s same-day runtime control at the same cell, post hoc and not preregistered, scored at 10^{-6} only.

condition	completed	valid at 10^{-6}	valid at 10^{-9}	anchor matches
D1 (thinking on, no system prompt)	14	12	11	5 of 12
D2 (thinking off, system line)	13	0	0	—
runtime control (post hoc)	6	6	not logged	4 of 6

Under D1 the median row spends 3160 reasoning tokens, a descriptive count carrying no registered threshold, $k = 4$ is modal at 6 of 12, and one output matches the family maximum value 1.7761424. The anchor $T(4, 13) = 1.625$ accounts for 5 of those 6. The sixth is a $k = 4$ grid at radius 0.125 filling only three of its four rows, twelve circles, with a thirteenth of radius 0.05 at the centre, summing to 1.55 rather than the 1.625 a full truncation gives. That extra radius is neither the anchor’s nor the family’s tangency filler at $k = 4$, which is 0.0517767, so the row has the anchor construction’s grid order without matching the anchor, and it is the whole difference between the 5 of 12 anchor matches and the 6 of 12 at $k = 4$. Under D2 every failure

is an overlap, which is arm P’s collapse. The two D1 failures overlap. The two runtime-control outputs that miss the anchor $T(4, 13) = 1.625$ are a 3×3 grid carrying corner fillers at 1.6144 and a centre-plus-border construction at 1.7500.

The family maximum over its full feasible domain. The post hoc audit `revision_math.json`, generated by `revision_math.py`, uses

$$F_N = \max \left(\{V(k, N - k^2) : k \geq 1, 0 \leq N - k^2 \leq (k - 1)^2\} \cup \{T(k, N) : k \geq 1, k^2 > N\} \right).$$

Every feasible extension order must be considered. Truncation values $N/(2k)$ decrease strictly with k , so only $k = \lfloor \sqrt{N} \rfloor + 1$ is needed from the second set. This leaves a finite exact maximization, rather than a general restriction to k^* and $k^* - 1$. For example, at $N = 60$ the feasible order $k = 6$ has $m = 24 \leq 25$ and $V(6, 24) = 1 + 2\sqrt{2} = 3.828427125$, larger than both $V(7, 11) = 3.825453513$ and $T(8, 60) = 3.75$. This counterexample invalidates the general two-candidate assertion. The exhaustive calculation preserves the reported maxima at all twelve studied cells, $N = 13, 17, 21, 31, 35, 37, 43, 50, 58, 62, 65$ and 75 . The audit records agreement with a brute-force comparison for $1 \leq N \leq 200$; the formula above supplies the domain beyond that finite check.

Which studied cells discriminate, and by how much. A cell discriminates when the anchor and the family maximum name different values, so that anchoring and family search can be told apart on it. At the studied discriminating cells the anchor takes the truncate branch, $T(k^*, N) = N/(2k^*)$, and exhaustive maximization gives the larger of that and the drop-and-fill value $V(k^* - 1, N - (k^* - 1)^2)$. This is a statement about these cells, not a general two-candidate theorem. Over the square cells the values differ at $N = 13, 21, 31$ and 43 , where the anchor and the maximum are 1.625 against 1.7761424, 2.1 against 2.2588835, 2.5833333 against 2.7485281, and 3.0714286 against 3.2416246. At $N = 35$ the truncated anchor is the larger of the two, and at $N = 17$ and 37 the branch rule extends and the anchor is itself the maximum, so those three cells cannot discriminate.

The feasible local-order null (post hoc sensitivity). A local-order sensitivity restricts attention, by assumption, to positive orders in $\{k^* - 1, k^*, k^* + 1\}$. Larger truncation orders are feasible too, so this restriction does not describe all templates the family can express. Retain an order k only when $N < k^2$ (truncation) or $0 \leq N - k^2 \leq (k - 1)^2$ (extension), and denote the resulting set by $\mathcal{K}_N$. Let $f_N(k)$ be that order’s branch value and A_N the anchor value. Conditional on a valid output, the sensitivity assumes uniform choice over $\mathcal{K}_N$, giving

$$p_N = \frac{\#\{k \in \mathcal{K}_N : |f_N(k) - A_N| < 2 \times 10^{-3}\}}{|\mathcal{K}_N|}.$$

Uniformity is over feasible local orders, not over values or coordinate packings. At $N = 17$, order 3 would require eight fillers where only four fit, so it is excluded: $\mathcal{K}_{17} = \{4, 5\}$ and $p_{17} = 1/2$. At each of the other six square cells the three local orders are feasible, exactly one falls within the matching window, and $p_N = 1/3$. In particular, at $N = 35$ the drop-and-fill value is about 0.0024531 below the anchor, outside the 0.002 window, so it does not add a second matching order.

Condition on each cell’s observed valid count v_N . Assuming independent choices across valid slots, the pooled count has the Poisson-binomial distribution with generating function $\prod_N (1 - p_N + p_N z)^{v_N}$. Its upper tail is obtained by summing coefficients from the observed hit count upward. The calculation in `revision_math.json` uses exact rational convolution. Invalid attempts are not counted as Bernoulli failures. Uniform weights conditional on validity are an assumption, not a fitted finding; a pre-emission interpretation would also require equal success probabilities across orders. Neither order-dependent validity nor retry dependence is estimated. These unadjusted post hoc tails do not change any registered verdict and do not establish recipe conformance.

arm / alias	subset	anchor matches / valid	feasible-null upper tail
V / north-mini-code	pooled	8/12	0.0372
V / gpt-oss-20b	pooled	11/18	0.0144
V / gemma-4-31b-it	pooled	0/15	1.000
V / north-mini-code	discriminating	3/4	0.111
V / gpt-oss-20b	discriminating	5/9	0.145
V / gemma-4-31b-it	discriminating	0/3	1.000
GM3 / all registered paths	pooled	46/80	1.91×10^{-4}
GM3 / first-party only	pooled	46/78	8.83×10^{-5}
GM3 / either accounting	discriminating	11/43	0.895

Constant-1/3 sensitivity. A simpler calculation, preserved in `revision_math.json` beside each feasible tail, assumes three choices at every cell and gives the following binomial upper tails. They are descriptive sensitivities, not the feasible-family null: the order-3 branch at $N = 17$ violates filler capacity. In the feasible calculation north-mini-code’s pooled tail rises from 0.019 to 0.0372; gpt-oss-20b’s is unchanged because it has no valid $N = 17$ outputs.

arm	alias	anchor matches / valid	exact upper tail
V	north-mini-code	8/12 pooled	0.019
V	gpt-oss-20b	11/18 pooled	0.014
V	north-mini-code	3/4 discriminating cells	0.111
V	gpt-oss-20b	5/9 discriminating cells	0.145
GM3	gemma-4-26b	11/43 discriminating cells	0.895

At an unadjusted 0.05 cutoff the two pooled V tails remain below the cutoff, whereas their discriminating-cell tails and GM3’s discriminating-cell tail remain above it. These comparisons depend on the stated local support, conditional weights and independence assumptions; they do not identify why pooled and discriminating results differ.

S6 Harmonized numerical and coordinate audit (post hoc)

This section supplies the definitions and complete arm-level results supporting §7 and Table 7. The numerical source is `revision_replay.json`; the coordinate source is `revision_membership.json`. The latter records the SHA-256 of the replay it classifies. These are post hoc analyses, not new registrations or replacements for the frozen original reports. Arm V includes every alias regardless of its floor, arm L includes generation 0 and all conditioned generations, and arm P includes both direct-emission and program cells. The replay re-executed arm P’s programs and did not substitute frozen results for its counts; the numbers below are taken from that report. Outputs invalid at 10^{-6} are excluded from the conformance denominator. The historical GM3 timing call remains excluded.

Numerical criterion. Recompute s from the emitted radii at full precision and F_N over all feasible orders as defined in S5. Harmonized clearance requires geometric validity at 10^{-9} and $s - F_N > 10^{-6}$. Positive excess with no score margin is a separate diagnostic. None of the eight arms has a harmonized clearance. This result says that no observed output meets the joint predicate; it does not say that every output belongs to the recipe family.

Coordinate conformance algorithm. For each output valid at 10^{-6} , the post hoc classifier searches candidate integer grid orders and requires absolute componentwise errors of at most 10^{-6} in x , y and radius. For $N < k^2$, every circle must match a distinct grid center with radius $1/(2k)$, and there are no fillers. For $N \geq k^2$, all k^2 grid centers must be occupied, with the remaining $N - k^2$ circles at distinct interior grid vertices with radius $(\sqrt{2} - 1)/(2k)$ and $N - k^2 \leq (k - 1)^2$. The grid-center and interior-vertex site sets already respect the square’s rotations and reflections. Scores do not enter these conformance decisions.

Candidate orders are obtained by intersecting, over observed radii r , the unions of integer intervals compatible with $|r - a/k| \leq 10^{-6}$, where $a = 1/2$ or $a = (\sqrt{2} - 1)/2$. Every candidate is checked against the coordinates and occupancy requirements. The implementation searches $1 \leq k \leq 10,000$. Within that bound, $k \times 10^{-6} \leq 0.01 < 1/4$, so the site tolerance neighborhoods are disjoint. If a radius-compatible interval extends beyond the bound, an output without a conformance witness is marked unresolved rather than rejected; conservative interval padding can also produce an unresolved classification. The report records no unresolved classifications in this corpus. A successful witness establishes approximate conformance at the stated tolerance, not exact mathematical conformance or exact feasibility. The table below repeats the first seven columns of Table 7, with an Unresolved column in place of D_9 .

Arm	Slots	v_6	v_9	C_9	Conform	Nonconform	Unresolved
MU	135	88	63	0	47	41	0
L	120	66	63	0	23	43	0
P	225	16	15	0	6	10	0
P-D	30	12	11	0	3	9	0
GM	140	20	20	0	11	9	0
GM2	140	0	0	0	0	0	0
GM3	140	80	79	0	72	8	0
V	325	48	42	0	30	18	0

Here v_6 and v_9 are valid counts at the two geometric tolerances, C_9 is the harmonized clearance count, and the last three columns partition v_6 by coordinate conformance at 10^{-6} . Planned slots include failures and unattempted slots and are not counts of returned content. These rows are separate serving-configuration and arm summaries; they are not pooled into a transfer rate. There are 192 approximately conforming and 138 nonconforming outputs among 330 outputs valid at 10^{-6} , with no unavailable validity or unresolved conformance cases in the reports. In particular, absence of clearance coexists with many coordinate-nonconforming packings. Conformance and clearance answer different questions.

An off-grid example. One MU anchor-parent output at $N = 13$ is valid at 10^{-9} and scores exactly 1.625, with all thirteen radii 0.125. Yet its center (0.5, 0.875) lies off the order-4 grid, so it does not conform. The next paragraph gives a rounding example.

GM3 maximum-score subset. The historical four-decimal score-match rule selects 27 outputs valid at 10^{-6} at discriminating cells, all with the coarse two-radius signature. The coordinate audit classifies 26 as conforming and one as nonconforming at 10^{-6} . The latter is **GM3/31/1**, source line 62 of `arm_gm_gm3_checkpoint.jsonl`. It is valid at 10^{-9} and occupies the expected 31 distinct sites, but emits filler radius 0.04142 instead of $(\sqrt{2} - 1)/10 = 0.041421356\dots$, an error of 1.356×10^{-6} . Its score 2.74852 matches $F_{31} = 2.748528137\dots$ after four-decimal rounding. This explains why a coarse radius signature can pass while the stated coordinate tolerance fails; nonconformance need not mean a qualitatively new construction.

Tolerance sensitivity. To test whether the 10^{-6} failures mostly reflect small discrepancies, we repeat the coordinate test at 10^{-4} and 10^{-3} (Table S5). The denominator stays fixed at the 330 outputs valid under the original 10^{-6} geometric check. Only the tolerance for matching family coordinates and radii changes; validity and clearance are not relaxed. The source is `revision_sensitivity.json`.

A further post hoc cross-tabulation joins the sensitivity rows to `revision_replay.json` by stable output identifier. Among the 129 failures at 10^{-3} , 116 also pass geometric validity at 10^{-9} ; all 116 score strictly below F_N . Main Table 7 reports their per-arm counts. The packaged `review_audits/check_strict_departure.py` regenerates this intersection and checks that the remaining 13 fail the 10^{-9} gate.

Of the 138 original failures, two conform at 10^{-4} and seven more at 10^{-3} ; 129 still fail. The GM3 rounded-filler example above is among the two recovered at 10^{-4} . The nine recovered cases establish small coordinate or radius discrepancies, not their cause. Conversely, the 129 remaining cases differ from this specified

Table S5: Coordinate conformance at three tolerances, on fixed outputs valid at 10^{-6} . Q_ϵ counts outputs conforming at coordinate tolerance ϵ ; U_3 counts those still nonconforming at 10^{-3} . No classification is unresolved. These are separate arm inventories, not a pooled transfer rate.

Arm	Valid	$Q_{10^{-6}}$	$Q_{10^{-4}}$	$Q_{10^{-3}}$	U_3
MU	88	47	47	47	41
L	66	23	23	27	39
P	16	6	6	6	10
P-D	12	3	3	3	9
GM	20	11	11	11	9
GM2	0	0	0	0	0
GM3	80	72	74	75	5
V	48	30	30	32	16

family beyond 10^{-3} under the test; this is not a count of distinct new algorithms or recipes. The original nonconformance result therefore is not mainly due to discrepancies that this looser tolerance absorbs.

For each tolerance ϵ , the classifier searches through an order bound satisfying $k\epsilon < 1/4$, keeping site neighborhoods disjoint. Radius-compatible orders beyond the bound remain explicitly unresolved rather than rejected. An independent exhaustive check uses the largest observed radius $r_{\max}$: every possible family match requires $k \leq [2(r_{\max} - \epsilon)]^{-1}$ when $r_{\max} > \epsilon$. This gives an exhaustive bound no larger than 36 here, including conservative endpoint padding. Enumerating sites explicitly agrees with the classifier on all 330 outputs at all three tolerances.

Where loop proposals leave the family. Table S6 separates newly proposed packings by lineage and generation. Nonconformance at 10^{-6} occurs in 8 of 17 valid seed proposals and 35 of 49 valid descendants. Across conditioned generations 1–5 the counts are 8/10, 8/10, 8/10, 5/7 and 6/12: failures persist throughout the run. At 10^{-3} these become 8/10, 7/10, 6/10, 5/7 and 5/12. The lineage pattern is uneven: all 18 valid descendants in 31-GREEDY fail even at 10^{-3} .

Retained best proposals. The audit reconstructs the terminal archive implied by the recorded proposals and the original stepper. That rule ranks proposals valid at 10^{-6} by the unrounded Python sum, uses stable ledger order for exact ties, and keeps one proposal per rounded score bucket, up to the archive’s capacity. All 100 logged conditioned parent scores agree with the reconstruction. The final state files contain only a done flag, so this is a reconstruction, not an independently saved terminal archive.

The selected best proposals are 13-GREEDY generation 0, slot 3; 13-DIVERSE generation 5, slot 4; 31-GREEDY generation 5, slot 2; and 31-DIVERSE generation 3, slot 5. Three match F_N within 10^{-9} and conform at 10^{-6} . The exception, 31-GREEDY, scores 2.6104167 and fails even at 10^{-3} . In 31-DIVERSE the displayed seven-decimal maximum first appears at generation 2 (Table 5); a later, slightly higher full-precision score determines the retained identity. The report includes every retained entry, score tie and source row. Thus nonconforming proposals coexist with conforming best proposals in three lineages. Four dependent trajectories do not identify a causal push away from the family followed by a selection effect back toward it.

Traceability and checks. The coordinate report records a source row identifier, selected ledger line, candidate orders, witnesses or failure reasons for each included output; the replay records selected slots, failed attempts and source hashes. Re-executing P’s 45 program slots reproduces the 12 outputs valid at 10^{-6} ; 11 are valid at 10^{-9} . At code cell $N = 13$, 3 of its 4 valid programs remain valid at 10^{-9} ; the frozen report logs no 10^{-9} count for program cells (S5). The reports record passing replay checks, 24 synthetic coordinate tests with no failures or errors, and agreement between the two reports’ sums and family maxima on all 330 compared rows. The sensitivity extension passes 11 additional tests; an independent explicit-site implementation agrees on all 990 output–tolerance comparisons. The full arm V cell table is Table S4; the exact feasible-null sensitivities are in S5. These checks support the stated implementation and arithmetic, not an inference that a score or a radius signature establishes recipe conformance.

Table S6: New arm L proposals, separated from retained archive entries. Each generation has five planned slots; v_6 counts valid proposals at geometric tolerance 10^{-6} . Q_ϵ counts coordinate conformance at tolerance ϵ among those proposals. Generation 0 is unconditioned. Zero-valid rows are counts, not imputed rates.

Lineage	Generation	v_6	$Q_{10^{-6}}$	$Q_{10^{-3}}$
13-GREEDY	0	5	2	2
13-GREEDY	1	1	1	1
13-GREEDY	2	0	0	0
13-GREEDY	3	0	0	0
13-GREEDY	4	1	0	0
13-GREEDY	5	2	2	2
13-DIVERSE	0	4	1	1
13-DIVERSE	1	4	1	1
13-DIVERSE	2	3	1	2
13-DIVERSE	3	5	0	2
13-DIVERSE	4	1	0	0
13-DIVERSE	5	3	2	3
31-GREEDY	0	4	3	3
31-GREEDY	1	5	0	0
31-GREEDY	2	4	0	0
31-GREEDY	3	3	0	0
31-GREEDY	4	3	0	0
31-GREEDY	5	3	0	0
31-DIVERSE	0	4	3	3
31-DIVERSE	1	0	0	0
31-DIVERSE	2	3	1	1
31-DIVERSE	3	2	2	2
31-DIVERSE	4	2	2	2
31-DIVERSE	5	4	2	2

S7 Subsequent Haiku loop replication and post hoc repair

This replication is additional evidence collected after the eight original arms. Its 600 planned slots are excluded from the 1,255-slot inventory in Table 7. The original four lineages and the 20 new lineages are analyzed separately with the same frozen replication code. The artifact’s `replication/` directory supplies the protocol, serving-configuration record, prompts, source, complete ledgers, and regenerated reports.

Design and serving configuration. Five fresh lineages per cell-regime pair cross $N = 13, 31$ with GREEDY and DIVERSE. Each has five unconditioned seeds and five subsequent generations of five proposals: 30 planned slots per lineage. A fresh Claude Code `general-purpose` subagent, requested as `haiku`, serves each slot. The runtime returns `claude-haiku-4-5-20251001`; this identifies the reported model, not independently attested weights. The original MU and L ledgers record no returned model identifier. Proposals do not share conversation history, but each mutation prompt includes its selected parent’s packing and score. Lineage, not proposal, is the replication unit. No proposal from another lineage is supplied.

The bare prompt, wrapper, score-bucket archive, and deterministic parent cycling are copied from the original stepper. Proposals valid at 10^{-6} enter the cumulative archive at geometric tolerance 10^{-6} ; GREEDY retains one score bucket and DIVERSE up to five. The archive rule is intentionally unchanged even when a retained parent fails the 10^{-9} gate. The wrapper requests no tools, and recorded tool-use counts are zero. Tool access was not technically disabled. Inherited instruction and memory attachments, whose delivered hashes are logged, make this an agent runtime rather than a bare API call. Every proposer received three such files (`replication/INSTRUMENT.md` section 4): the operator’s global agent instructions (15,456 characters; no circle-packing content), one unrelated browser-tool skill file (5,360 characters), and an auto-memory index (2,434 characters) whose entries name a research-paper operations memory, an “agent-as-proposer

harness” and an anonymized project, with no hypothesis, formula, prediction, outcome count or packing. These files are the operator’s private, identity-bearing configuration and are withheld from the artifact; that they carry no task content is an author assertion, checkable only against the logged digests, and the seed proposals, which form P1’s denominator, are the proposals most exposed to any priming. The original full system context, weights, decoding parameters, and thinking budget cannot be reconstructed. Runtime versions changed during collection from 2.1.260 to 2.1.266; the original run used 2.1.246 and a different spawn depth. All 260 slots on 2.1.266 carry runtime attachments that none of the 340 slots on 2.1.260 carry (types `advisor_tool`, `environment`, `model` and `prompt_snapshot`, the last up to 208,657 characters); the ledgers log only type and length. In 8 of the 10 $N = 31$ lineages, the seeds ran on 2.1.260 and every descendant on 2.1.266. The ten $N = 13$ lineages ran entirely on 2.1.260, and 31-DIVERSE r4 and r5 entirely on 2.1.266. Post hoc, the anchor-match rate is lower in descendants in all 12 of these single-version lineages (Table S7). A 225-character `remote_session_change` attachment also appears on generations 0 and 1 of all ten $N = 13$ lineages, but not on their generations 2 to 5, and on every 2.1.266 slot; the ledgers log only its type and length. So the ten $N = 13$ lineages share one runtime version but not one context. In 31-DIVERSE r4 and r5, the logged lengths of three attachments also change after generation 1, so no single-version lineage has a constant logged context. These tallies come from the `runtime_version` and `attachments` fields of `replication/ledgers/*/responses.jsonl`. By author inspection of two 2.1.266 transcripts, `prompt_snapshot` repeats the runtime’s own subagent system prompt and tool list, and a keyword scan found no circle-packing content.

Registration and deviations. The protocol and analysis were frozen in local commits `52ac2c1` and `fa70e73` before experimental sampling according to the local chronology. This is not an independently verified public preregistration; the review export of `replication/` carries no git history, so that chronology rests on `results/REGISTRATION.txt` and the commit hashes it names. The second commit corrected a pre-run modal-bucket implementation to use generation 0. Original committed files remain unchanged; redacted review copies have their own hash manifest. The initial concurrency plan exceeded the harness limit and was reduced to 20 concurrent subagents. Undelivered calls were retried under the fixed two-attempt policy, and delayed batches were completed without changing frozen prompts or parents. A delivered empty answer remained a terminal failure. These events and the runtime version change are disclosed in the exported serving-configuration and deviation records.

The frozen P2 prose says “best valid score after generation 0 versus best valid score over the whole lineage; difference,” without specifying the subtraction clearly. The committed code unambiguously computes best-of-run minus best-seed. We report that executed definition and disclose the wording ambiguity, without changing the analysis.

Outcomes fixed before collection. P1 is the fraction of valid, genuinely conditioned descendants with $|S - A_N| < 0.002$, minus the corresponding fraction among valid generation-0 proposals; empty denominators are missing. P2 is $\max_{g=0,\dots,5} S - \max_{g=0} S$, among proposals valid at 10^{-6} . P3 is whether any raw proposal is valid at 10^{-9} and exceeds F_N by more than 10^{-6} . The descriptive `rival` flag instead means $|S - F_N| < 0.002$; `above_secondary` means $S > F_N + 10^{-9}$ among proposals valid at 10^{-6} . Neither is P3. Scoreability requires three valid seeds and at least one valid proposal in three conditioned generations. The hypotheses predict declining anchor-match rate in most scoreable lineages (H1), small best-score changes and no P3 clearance (H2), and imperfect, cell-dependent validity (H3). “Small” and “materially below 1” have no numerical thresholds. We report the observed magnitudes rather than inventing pass/fail thresholds.

Registered raw outcomes. Every planned slot is terminal. There are 599 text-bearing deliveries and one empty delivery; 595 slots have parsed coordinates, including 12 with the wrong circle count. Four further text-bearing deliveries fail coordinate parsing. Wrong counts, parse failures, and the empty delivery remain failures. Of 600 slots, 357 are valid at 10^{-6} and 295 at 10^{-9} . All 20 lineages have negative P1; 19 meet the scoreability floor. The remaining lineage, 31-DIVERSE-r2, has only two valid seeds and is retained descriptively. Thus H1’s direction occurs in every scoreable lineage. Treating each of the 20 replication lineages as one independent observation (fresh conversations, no cross-lineage sharing), a post hoc two-sided sign test on the lineage-level differences (20 of 20 negative) gives $p = 2^{-19} \approx 1.9 \times 10^{-6}$ (3.8×10^{-6} over the 19 scoreable lineages). Its null is that the signs of the nonzero lineage differences are independent and equiprobable, conditional on the fixed design and on the lineages observed; it is not a test that seed and

Table S7: Every new lineage, with 30 planned slots each. Seed and descendant columns give anchor matches / proposals valid at 10^{-6} . v_6/v_9 gives geometric validity counts; P2 uses best-of-run minus best-seed scores among proposals valid at 10^{-6} . P3 is zero for every row. The asterisk marks the lineage below the seed-validity floor; its ratios remain descriptive.

Lineage	v_6/v_9	Seed	Descendant	P1	P2
13-D-r1	20/20	2/5	1/15	-0.3333	0.1279
13-D-r2	17/13	3/4	0/13	-0.7500	0.1511
13-D-r3	16/14	2/4	0/12	-0.5000	0.0000
13-D-r4	19/19	3/5	3/14	-0.3857	0.0000
13-D-r5	13/12	3/4	0/9	-0.7500	0.1450
13-G-r1	14/7	4/4	1/10	-0.9000	0.1511
13-G-r2	16/15	2/4	0/12	-0.5000	0.1415
13-G-r3	19/9	2/5	0/14	-0.4000	0.1355
13-G-r4	15/4	3/4	0/11	-0.7500	0.1095
13-G-r5	16/16	3/3	2/13	-0.8462	0.1453
31-D-r1	23/21	3/5	4/18	-0.3778	0.1652
31-D-r2*	11/11	2/2	2/9	-0.7778	0.1652
31-D-r3	24/18	5/5	1/19	-0.9474	0.1925
31-D-r4	22/17	3/4	2/18	-0.6389	0.0000
31-D-r5	20/15	4/4	0/16	-1.0000	0.1652
31-G-r1	20/19	4/5	1/15	-0.7333	0.1652
31-G-r2	19/16	4/5	1/14	-0.7286	0.1652
31-G-r3	18/18	3/5	1/13	-0.5231	0.1548
31-G-r4	19/15	3/4	0/15	-0.7500	0.1299
31-G-r5	16/16	1/3	0/13	-0.3333	0.0000

descendant emission probabilities are equal. It is the only inference on this direction, and a direction check against a weak null: it tests only the direction of the change under the mutation prompt, and the four original lineages are not pooled: they used a different runtime version, spawn depth and inherited context, and they motivated the replication. P3 is false in all 20 new and all four original lineages under the frozen replication criterion. The same-code original P1 differences are -0.4000 , -0.1875 , -0.5833 , and -0.7273 for 13-GREEDY, 13-DIVERSE, 31-GREEDY, and 31-DIVERSE. These repeated directions add evidence about the tested runtime; they do not identify a causal archive-width effect.

P2 ranges from zero to 0.1925, with the exact lineage values in Table S7. Validity at 10^{-6} is 165/300 at $N = 13$ and 192/300 at $N = 31$. These quantify H2’s uncalibrated “small” and H3’s cell contrast; neither adjective defines a registered statistical test.

Anchor-matching proposals and coordinate conformance (post hoc). P1 is a score share: a proposal matches the anchor when $|S - A_N| < 2 \times 10^{-3}$, whatever its geometry. Applying the S6 classifier, at coordinate tolerances 10^{-6} and 10^{-3} , to the 357 replication proposals valid at 10^{-6} (Table S8) shows that most anchor-matching proposals are not the k^* grid template. At $N = 13$, 7 of 27 anchor-matching seeds and 2 of 7 anchor-matching descendants conform at both tolerances; at $N = 31$, 14 of 32 seeds conform at 10^{-6} (15 at 10^{-3}) and 6 of 12 descendants at both. Every conformance witness is the k^* truncated grid, so these counts equal a direct test of “all N circles on distinct k^* grid centres with radius $1/(2k^*)$.” Of the 48 anchor-matching proposals that fail at 10^{-3} , 46 keep every radius at $1/(2k^*)$: they leave the family through the centres. Fifteen are hexagonal-row packings, rows at pitch $\sqrt{3}r$ with alternate rows offset by r : at $N = 31$, six seeds (31-D-r3 g0 s1, 31-D-r3 g0 s2, 31-D-r4 g0 s5, 31-D-r5 g0 s4, 31-G-r2 g0 s4, 31-G-r3 g0 s1) and two descendants (31-D-r1 g3 s4, 31-G-r2 g1 s1) place 31 circles of radius $1/12$ in six rows at the anchor value A_{31} ; at $N = 13$, seven seeds do the same at radius $1/8$. The r3 construction of Table S9 is this row rule at radius $1/11$ (six rows of five) with one corner circle; its recorded parent is 31-D-r3 g0 s4, the square-grid anchor of Figure 1a, not a hexagonal seed. Of the remaining 33 failures, 25 differ from the k^* template in exactly one

circle, as in the MU example of S6. The run is `review_audits/check_replication_conformance.py`; its report lists every anchor-matching proposal with verdicts, witnesses, radius profile and row structure.

Table S8: Coordinate conformance of anchor-matching replication proposals (post hoc). The anchor columns count score matches ($|S - A_N| < 2 \times 10^{-3}$) among proposals valid at 10^{-6} ; Q_ϵ counts those conforming to the k^* grid at coordinate tolerance ϵ . Hex counts anchor-matching seeds that are hexagonal-row packings of radius $1/(2k^*)$. Source: `review_audits/replication_conformance.json`.

Lineage	Seed anchor	$Q_{10^{-6}}$	$Q_{10^{-3}}$	Desc. anchor	$Q_{10^{-6}}$	$Q_{10^{-3}}$	Hex
13-D-r1	2/5	0	0	1/15	0	0	0
13-D-r2	3/4	1	1	0/13	0	0	1
13-D-r3	2/4	1	1	0/12	0	0	1
13-D-r4	3/5	1	1	3/14	0	0	0
13-D-r5	3/4	0	0	0/9	0	0	3
13-G-r1	4/4	2	2	1/10	1	1	0
13-G-r2	2/4	0	0	0/12	0	0	1
13-G-r3	2/5	1	1	0/14	0	0	1
13-G-r4	3/4	0	0	0/11	0	0	0
13-G-r5	3/3	1	1	2/13	1	1	0
31-D-r1	3/5	2	3	4/18	2	2	0
31-D-r2	2/2	1	1	2/9	2	2	0
31-D-r3	5/5	3	3	1/19	1	1	2
31-D-r4	3/4	0	0	2/18	0	0	1
31-D-r5	4/4	2	2	0/16	0	0	1
31-G-r1	4/5	1	1	1/15	1	1	0
31-G-r2	4/5	3	3	1/14	0	0	1
31-G-r3	3/5	1	1	1/13	0	0	1
31-G-r4	3/4	1	1	0/15	0	0	0
31-G-r5	1/3	0	0	0/13	0	0	0
$N = 13$ pooled	27/42	7	7	7/123	2	2	7
$N = 31$ pooled	32/42	14	15	12/150	6	6	6

Anchor-construction share by lineage and generation (post hoc). Among all valid proposals, 21 of 84 seeds and 8 of 273 descendants are the anchor construction at coordinate tolerance 10^{-6} (22 and 8 at 10^{-3}). The share is lower in descendants than in seeds in 14 lineages, zero in both in 6 and higher in none. Among the 19 scoreable lineages, the split is 13, 6 and 0. By generation, the anchor construction accounts for 4 of 49, 1 of 69, 1 of 58, 1 of 49 and 1 of 48 valid descendants in generations 1 to 5. In the four original lineages, the same classifier with the S6 witnesses finds 8 of 17 valid seeds and 2 of 49 valid descendants. The counts are identical at 10^{-6} and 10^{-3} ; Table 5 gives them by lineage. The share is lower in descendants in all four, although the count rises from 1 to 2 in 13-DIVERSE. In original generations 1 to 5, 1 of 10, 1 of 10, 0 of 10, 0 of 7 and 0 of 12 valid descendants are the anchor construction. Only 8 of the 11 anchor-matching original seeds and 2 of the 11 anchor-matching original descendants are the anchor construction. The script `review_audits/check_copy_rate.py` recomputes these counts with the frozen classifier and writes `copy_rate.json`. It checks the original witnesses row by row against `artifact/revision_membership.json` and `artifact/revision_sensitivity.json` (66 valid rows each), and the replication counts lineage by lineage against `replication_conformance.json`.

Historical archive arithmetic. The executed stepper and the frozen analysis (`runner.py`, line 105; `analysis.py`, line 47) both call Python’s built-in `sum`. The executed ledger scores equal the ordinary left-to-right sum on all 357 valid slots. The exported analysis scores instead equal `math.fsum` on all 357 and differ from the left-to-right sum on 192, by at most 2.22×10^{-15} : built-in `sum` has used compensated summation since CPython 3.12, and CPython 3.13.1 regenerates the packaged export byte for byte whereas CPython 3.11.15 does not. The interpreter that produced the export is not recorded, so this is a reproduction, not an execution record. Replacing the executed sum with `fsum` would move four proposals between 0.002-wide archive buckets and would select different parents for two later slots. We preserve the parents actually shown

to the proposer and replay all 500 conditioned requests exactly under the historical arithmetic. No P3 verdict changes. The independent audit computes both explicitly, a left fold and `fsum`, rather than rewriting the executed trajectory.

Fifty-four raw proposals valid at 10^{-6} in eleven new lineages meet the looser descriptive secondary threshold ($S > F_N + 10^{-9}$), compared with none in the four original lineages. Only three of the 54, a generation-4 r3 proposal and two of its generation-5 children, remain above $F_N + 10^{-6}$ after the repair; the other 51 exceed F_N by at most 6.0×10^{-6} (14 of the 54 exceed $F_N + 10^{-6}$ as emitted: the r3 trio and eleven within 6.0×10^{-6}) and fall below it under reductions $1 - \alpha$ of at most 7.1×10^{-6} , so the gap between the 10^{-6} and 10^{-9} counts is tolerance, not construction. That difference must not be confused with the zero P3 count. In particular, above-family proposals can enter the executed archive at 10^{-6} while failing 10^{-9} . Dropping those parents retrospectively would change subsequent prompts; a 10^{-9} -gate counterfactual loop was not run.

Repair analysis, chosen after observing the proposals. A raw-proposal failure at the 10^{-9} gate does not establish that its construction has no feasible improvement over F_N . We therefore apply one deterministic diagnostic to every correct-count, finite, positive-radius packing with centers strictly inside the unit square. Let $b_i = \min(x_i, 1 - x_i, y_i, 1 - y_i)$ and $d_{ij} = \|c_i - c_j\|_2$. Keep all centers fixed and replace every radius by αr_i , where

$$\alpha = (1 - 10^{-12}) \min \left\{ 1, \min_i \frac{b_i}{r_i}, \min_{i < j} \frac{d_{ij}}{r_i + r_j} \right\}.$$

A nonpositive factor cannot yield an eligible repair. This operation only decreases radii, does not add or remove circles, and supplies no new model proposals. In exact arithmetic a positive factor makes the repaired circles contained and non-overlapping; the implementation additionally checks the serialized repaired coordinates. Applying the same rule to the full eligible corpus prevents selection of only favorable repairs. It is a post hoc feasibility diagnostic, not an extra search generation or a registered success. Its outputs never enter the recorded archives or replace raw ledger rows.

Table S9: All above-family outcomes of the post hoc radius-only repair, from two 31-DIVERSE lineages. Each raw proposal fails validity at 10^{-9} ; every repaired packing retains 31 circles. The fractional radius reduction is $1 - \alpha$. $F_{31} = 2.748528137424$. Full-precision coordinates and verification are in the artifact.

Lineage / gen. / slot	Raw score	Repaired score	Excess	$1 - \alpha$
r3 / 4 / 2	2.772727273	2.772727147	0.024199009	4.55×10^{-8}
r3 / 5 / 1	2.775793388	2.775793261	0.027265124	4.55×10^{-8}
r3 / 5 / 4	2.775795455	2.775795328	0.027267191	4.55×10^{-8}
r5 / 5 / 2	2.751500000	2.750833902	0.002305765	2.42×10^{-4}

The diagnostic yields positive repair factors for 582 of the 600 slots. Four repaired packings clear F_N , from two $N = 31$ DIVERSE lineages. Three are a generation-4 proposal in r3 and two of its generation-5 children, and must not be counted as three independent discoveries. Their common radius reduction is approximately 4.55×10^{-8} of each radius. The fourth, r5 generation 5 slot 2, requires a fractional reduction of about 2.42×10^{-4} and fails even the raw 10^{-6} gate. Counting it, the repair yields four above-family packings: the three r3 survivors of the 54 and one proposal outside the 54. Uniform scaling yields exactly four clearances in this corpus; they establish at least four repairable proposals under the broader class of per-circle radius reductions, of which uniform scaling is a special case, so a per-circle repair could add above-family cases but not remove these. The count is specific to uniform scaling and the operator was chosen after the proposals were seen, and the same operator yields none among the 625 positive-factor original outputs (§7). The r3 staggered grid differs from the specified square-grid family; the retained score advantage after repair shows why a raw validity failure cannot be promoted into a family ceiling. The raw proposals are neither valid escapes nor a mere tolerance artifact: the raw coordinates fail 10^{-9} , while the repaired constructions exceed F_{31} . The corpus-wide check also finds the r5 case. Both facts are reported with their distinct analysis status.

The exact $N = 31$ construction (post hoc). Up to two slips, the r3 generation-4 slot-2 proposal is an exact construction with coordinates in $\mathbb{Q}(\sqrt{3})$. It has 30 circles of radius $1/11$ in six rows of five. Row $i = 0, \dots, 5$, counted from the bottom, sits at height $(1 + i\sqrt{3})/11$, with centres at $x = (2j + 1)/11$ in even

rows and $x = (2j + 2)/11$ in odd rows, $j = 0, \dots, 4$. A corner circle of radius $1/22$ centred at $(21/22, 1/20)$ completes it. In exact arithmetic this packing is feasible at zero tolerance, with 69 tangent pairs and 12 wall contacts. Its sum of radii, $61/22 = 2.7727\dots$, exceeds $F_{31} = V(5, 6) = (19 + 6\sqrt{2})/10 = 2.7485\dots$ by $48/55 - \frac{3}{5}\sqrt{2} \approx 0.0242$. The sign follows from $48/55 > 0$ and $(48/55)^2 = 2304/3025 > 18/25 = 2(3/5)^2$. The sum exceeds the anchor value $A_{31} = 31/12$ by $25/132$. The emitted coordinates carry two slips in row heights: $+9.5 \times 10^{-9}$ in the third row from the bottom and -9.1×10^{-12} in the top row. Every other emitted coordinate lies within 5.5×10^{-17} of its exact value. The slips are arithmetic errors, not rounding. Each row height is written with 17 decimals. The two slips are 4.2×10^7 and 4.0×10^4 times the rounding bound, half a unit in the 17th decimal plus 2^{-52} . Read as exact rationals, the emitted coordinates make nine circle pairs between the third and fourth rows overlap by 8.3×10^{-9} . These are the nine pairwise overlaps of §5, and they are why the proposal fails the 10^{-9} gate. The smaller slip makes nine pairs between the fifth and sixth rows overlap by 7.9×10^{-12} , below the gate. All three replication proposals valid at 10^{-6} that exceed F_{31} by more than 10^{-3} come from this lineage: this proposal and two generation-5 descendants. Both have it as parent and repeat its 30 row circles as floating-point values, slips included; only the corner circle changes. Four of their six height strings are written with fewer decimals (`replication/ledgers/31-diverse-r3/responses.jsonl`). No packing record is claimed: the best known sum at $N = 31$ is 2.889+ (Friedman, n.d.). The exact-arithmetic check is `review_audits/check_hex_exact.py`, which writes `hex_exact.json`.

Original corpus under the same repair (post hoc). The rule above was also applied to all 1,255 original planned slots of Table 7. Every parsed output with exactly N finite positive-radius triples enters the diagnostic regardless of validity: 651 outputs, of which 625 receive a positive factor; 26 contain two coincident centers and have none. No repaired packing scores above F_N , and none clears the 10^{-6} margin: zero repaired clearances in the original arms against four in the replication. By stratum, 295 of 321 such outputs invalid at 10^{-6} , all 37 valid at 10^{-6} but not 10^{-9} , and all 293 valid at 10^{-9} receive a positive factor; the largest repaired excess per stratum is -8.1×10^{-6} (L, $N = 13$), -3.7×10^{-8} (MU, $N = 21$) and -1.8×10^{-12} (exact family packings at $N = 13$, reduced only by the 10^{-12} safety factor). The three MU outputs above the score margin at 10^{-6} require $1 - \alpha$ between 1.4×10^{-6} and 1.8×10^{-6} and end 1.4×10^{-6} to 2.4×10^{-6} below F_N ; the one positive raw excess among outputs valid at 10^{-9} , 5.29×10^{-10} at GM3 $N = 13$, becomes -4.7×10^{-10} . The check is `review_audits/check_repaired_clearance.py`, which imports the audit’s repair function and writes `repaired_clearance.json`.

Copying (post hoc). For every valid descendant, `review_audits/check_copy_rate.py` compares the proposal with the parent shown in its prompt. A score copy returns the parent’s score within 10^{-6} ; an improvement or a lower score differs from it by more than 10^{-6} . A coordinate copy has a one-to-one circle matching in which every coordinate and radius lies within 10^{-6} of the parent’s. Scores are left-fold sums, the executed arithmetic. Replication parents are the coordinates recorded in each request. The original ledgers keep only the displayed parent score. Original parents are therefore rebuilt with the registered archive rule, and the rebuilt score matches the recorded one in 100 of 100 rows. In 47 original rows an earlier proposal shows the same score with other coordinates; counting a match to any of them still gives 13 coordinate copies. Table S10 gives the counts by generation. Of the 118 replication improvements, 103 are at least 2×10^{-3} ; at coordinate tolerance 10^{-3} , 111 descendants are coordinate copies.

By parent class in the replication, parents that match the anchor have 49 valid descendants: 38 improve, 9 return the score and 2 score lower. Of these 49, 11 stay anchor matches, the 9 copies and 2 improvements inside the matching window. Parents above the anchor value have 202 valid descendants: 60 improve, 129 return the score and 13 score lower. Only 2 of them are anchor matches, both by scoring lower than the parent. Parents below the anchor value have 22: 20 improve and 2 return the score, and 6 are anchor matches, all by improving. In 10 lineages, the first proposal above the anchor value descends from an anchor-matching parent; in 2, from a below-anchor parent; in 8, it is a seed. In the four original lineages, anchor-matching parents have 13 valid descendants: 6 improve, 7 return the score and none score lower, and 7 stay anchor matches. Parents above the anchor value have 32, of which 2 are anchor matches. The first proposal above the anchor value is an anchor-parent descendant in 3 original lineages and a seed in 1. An instruction-free control was not run, so these counts do not separate copying from a judgment that the parent is already good.

Table S10: Copying in the loops by generation (post hoc). Counts are among valid descendants; parentheses give percentages of each row’s valid descendants (integer, rounded half up; original per-generation rows give counts only). Score copy: within 10^{-6} of the parent’s score. Coordinate copy: every coordinate and radius within 10^{-6} of the parent’s. Improve and Lower: above or below the parent’s score by more than 10^{-6} . Source: `review_audits/copy_rate.json`.

Corpus	Generation	Valid	Score copy	Coordinate copy	Improve	Lower
Replication	1	49	16 (33%)	11 (22%)	31 (63%)	2 (4%)
	2	69	35 (51%)	25 (36%)	30 (43%)	4 (6%)
	3	58	32 (55%)	17 (29%)	23 (40%)	3 (5%)
	4	49	32 (65%)	22 (45%)	16 (33%)	1 (2%)
	5	48	25 (52%)	22 (46%)	18 (38%)	5 (10%)
	1–5	273	140 (51%)	97 (36%)	118 (43%)	15 (5%)
Original L	1	10	4	2	6	0
	2	10	7	5	2	1
	3	10	4	3	5	1
	4	7	3	1	1	3
	5	12	5	2	6	1
	1–5	49	23 (47%)	13 (27%)	20 (41%)	6 (12%)

References

- Anonymous authors. Capability or optimizer? What a zero-shot LLM proposer contributes before evolutionary program search begins, on circle packing, 2026. Unpublished companion manuscript. An anonymized copy is available to the action editor on request.
- Erich Friedman. Packing center: circles in squares, maximizing the sum of radii. <https://erich-friedman.github.io/packing/cirRsqu/>, n.d. Accessed 2026.