

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

Abstract

Does the answer an LLM proposer gives most often zero-shot persist inside a search loop? For an agent runtime (a Claude Code subagent requested as Haiku) on circle packing, an AlphaEvolve showcase task, it mostly does not. Without a parent, the subagent’s valid outputs at $N = 13$ and 31 most often match the score of one grid layout, the anchor construction. For single calls that show a parent and ask for a higher sum of radii, we registered that the anchor would persist. Instead, shown that construction as a parent, most valid outputs move above it; shown a parent scoring the family maximum (the best related-grid score), all 26 valid outputs keep or improve its score. Only a parent far below the anchor value often yields the anchor value. Inside a minimal archive-and-mutate loop, the registered anchor-match rate is lower in descendants than in parentless seeds in all 19 replicate lineages with enough valid outputs to score. Post hoc, most valid descendants keep or improve their parent’s score; few reproduce the anchor construction. Validity differs by serving configuration: through a pinned API with thinking off, no directly emitted packing was valid in 105 calls at seven circle counts, unlike most parentless agent-runtime outputs.

1 Introduction

Evolutionary search systems show a language model earlier solutions and keep the best of its proposals (Romera-Paredes et al., 2024; Novikov et al., 2025).¹ Packing N circles in a unit square to maximize the sum of their radii is a showcase task for AlphaEvolve (Novikov et al., 2025). It is also an example task of the open frameworks OpenEvolve and ShinkaEvolve (Sharma, 2025; Lange et al., 2026). Zero-shot tests show the proposer no earlier solution; a loop always shows one. Unselected mutation chains often revisit earlier forms (Gurkan et al., 2026). We ask whether the answer a proposer gives most often zero-shot persists inside the loop. For an agent runtime (a Claude Code subagent requested as Haiku) on circle packing, it mostly does not. Inside a loop, the proposer’s valid outputs mostly keep or improve the parent shown (post hoc) rather than return to the proposer’s zero-shot anchor, the grid construction whose score it returns most often without a parent (§2). Validity differs by serving configuration, and the validity tolerance decides which outputs count. Our predictions were committed to local git history before sampling, without a verifiable public timestamp (§3). Every miss is reported, including our predictions that the anchor would persist under an anchor or maximizing parent. The headline registered loop test is the seed-versus-descendant anchor-match rate; construction and copy counts are post hoc.

Terms. A $k \times k$ grid places circles of radius $1/(2k)$ at its sites. A truncated grid fills only $N < k^2$ sites and scores $T(k, N) = N/(2k)$. A grid with m fillers adds circles of radius $(\sqrt{2} - 1)/(2k)$ at interior vertices and scores $V(k, m) = k/2 + m(\sqrt{2} - 1)/(2k)$. These grids form the recipe family. The *anchor construction* is the member built on the $k^* \times k^*$ grid, $k^* = \text{round}(\sqrt{N})$; its score is the *anchor value* A_N . The *family maximum* F_N is the best score of any member with exactly N circles. At $N = 31$, $A_{31} = T(6, 31) \approx 2.583$ and $F_{31} = V(5, 6) \approx 2.749$ (Figure 1a,b). The best known packing scores 2.889+ (Friedman, n.d.), so F_N

¹Claude models assisted the prose and implementation; the authors approved designs and registrations before sampling. An OpenAI coding assistant also helped with editing, analysis, and verification. The authors remain responsible for scientific claims and final submission approval.

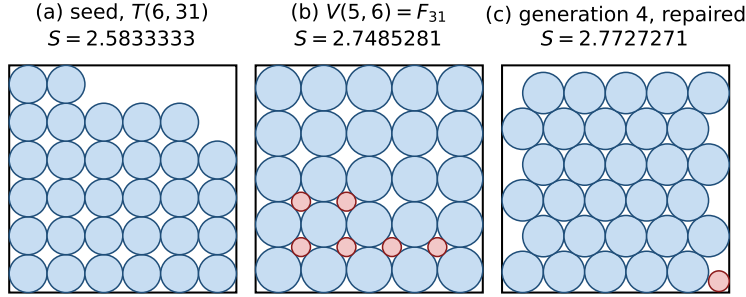


Figure 1: *Three $N = 31$ packings from one replication lineage (§5), for illustration.* (a) A seed: the anchor construction $T(6, 31)$, a 6×6 grid of radius $1/12$ missing five sites. (b) Generation 1: the maximizing construction $V(5, 6)$, a 5×5 grid of radius $1/10$ with six fillers (red). (c) Generation 4, a child of (a): the model’s construction, 30 circles of radius $1/11$ in six offset rows plus a corner circle of radius $1/22$ (red). Post hoc, its exact sum, $61/22$, exceeds F_{31} by 0.0242 . An arithmetic slip of 9.5×10^{-9} in one emitted row height makes it invalid at tolerance 10^{-9} . Shown after post hoc radius-only repair (§5), which scores 2.7727271 .

bounds the family, not the problem. An output is an *anchor match* when its score lies within 2×10^{-3} of A_N . This tests only the score, so we also test whether coordinates and radii *conform* to the anchor construction at coordinate tolerance 10^{-6} . Without a parent, through this runtime, the model’s valid outputs at $N = 13$ and 31 most often (a plurality at each cell) score the anchor value. Of 84 valid parentless seeds in our loop replication, 59 do, though only 21 are the anchor construction (post hoc). A companion study (Anonymous authors, 2026) describes this construction in closed form.

Findings. The loop and single-parent experiments ran through this agent runtime. Its version and inherited context differ between the original loop and the replication, and within the replication (§3; Supplement S7). The registered pinned-API re-draws of both experiments were not run, so neither is shown for a bare API call. Table 1 summarizes the registered arms reported here and the loop replication.

The loop leaves the anchor value and construction. Each lineage of our minimal archive-and-mutate loop, at $N = 13$ and 31 , starts with parentless seeds; five generations of descendants each see one archived parent and a request to modify it for a higher sum of radii. The registered anchor-match rate falls from seeds to descendants in all 19 replicate lineages with enough valid outputs to score, in the other one, and in the 4 original ones, which inherited no operator files (§3). Post hoc, in the replication, 21 of 84 valid seeds but 8 of 273 valid descendants are the anchor construction, and 140 of the 273 return their parent’s score. Also post hoc, given a parent that matches the anchor, 38 of 49 improve on it; given a parent above the anchor value, only 2 of 202 are anchor matches. Of the 19 valid descendants that match the anchor, 11 have an anchor-matching parent; with no instruction-free control, this describes the loop but does not explain it.

One parent at a time. Shown a single maximizing parent, one scoring the family maximum, all 26 valid outputs (of 45 calls) keep or improve its score (registered); post hoc, 25 return it within 10^{-6} . Shown an anchor parent, 17 of 28 move above it. A poor parent, a diagonal chain far below the anchor, leaves 20 of 34 valid outputs at the anchor value, 12 of 14 at $N = 31$ but 1 of 7 at 13.

Two serving configurations differ sharply in validity. Through a pinned API (an OpenRouter model alias) with thinking off, the model returned no valid packing in 105 direct-emission calls at seven circle counts. Most outputs parse but overlap (§6). On the same day, the same $N = 13$ prompt returned 6 valid of 6 through the agent runtime (post hoc). With thinking on, that API returned 12 valid of 14 completed calls at $N = 13$ (registered).

In one lineage, the validity tolerance decides what counts as better. In exact arithmetic (post hoc), one lineage’s hexagonal-row construction (Figure 1c) is feasible and 0.0242 above F_{31} . Two emitted row heights carry arithmetic slips, far larger than rounding, of 9.5×10^{-9} and 9.1×10^{-12} . The larger makes the 10^{-9} gate reject it, though the loop’s 10^{-6} archive kept it. It is not a packing record.

Table 1: Registered experiments (arms) and the loop replication. Met or missed: whether the data reach a registered threshold. Rates count outputs valid at 10^{-6} . Supplement S1 lists each arm’s prediction codes.

Arm	Setting	Outcome
L, replication	Loop, agent runtime; 20 lineages	Anchor-match rate lower in descendants in 19/19 scoreable lineages and in the unscoreable one. Anchor construction at coordinate tolerance 10^{-6} (post hoc): 21/84 seeds, 8/273 descendants.
L	Same loop; 4 lineages	Anchor matches: 11/17 seeds, 11/49 descendants (predicted fall per cell, met descriptively). Anchor construction (post hoc): 8/17 seeds, 2/49 descendants. Modal-seed and archive-width predictions missed.
MU	One parent, agent runtime	Maximizing parent: 26/26 keep or improve its score, so 0/26 match the anchor (registered anchoring-dissolved rule met, entailed by keep-or-improve); predicted anchor return and grid-order change missed. Anchor parent: 7/28 are anchor matches ($\geq 50\%$ predicted, missed); 17/28 move above it. Off-family parent: 20/34 match the anchor; the modal output uses the anchor’s grid order (met at $N = 21$ and 31; missed at 13 on a tie).
P, P-D	Pinned API, no parent	P: 0/105 direct-emission calls valid at seven circle counts and 4/75 at five held-out ones, so its direct-emission predictions are unscoreable (post hoc reading); same-day control through the agent runtime 6/6 at $N = 13$ (post hoc); no valid program clears F_N (met; its opposite missed). P-D at $N = 13$: thinking on 12 of 14 completed calls valid, off 0 of 13 (met).
GM, GM2, GM3, V	Other models, no parent (§6; Supplement S4)	GM: 20/140 valid (original parser), underpowered. GM2: 0/140 parseable. GM3: pooled anchor-match rate met; per-cell anchor-mode and two-radius-signature predictions missed. V: 10 of 13 aliases give too few valid outputs to score; 2 of the other 3 meet the anchor-majority prediction.

Contribution and scope. For this runtime and task, the observations suggest testing a proposer under the loop’s own parents, serving configuration and validity tolerance, and reporting conformance apart from score. We did not compare zero-shot and in-loop rankings of proposers. The loop and single-parent results cover one agent runtime and one task. Other-model experiments are summarized in §6 and detailed in Supplement S4. The loop evolves packings, not programs; we do not claim that it replicates FunSearch.

2 The template and its family maximum

Task. Maximize the sum of radii of N non-overlapping circles inside the unit square. Deterministic geometric checks, not a human or LLM judge, assign scores. The bare prompt requests direct emission, a Python list of $[\mathbf{x}, \mathbf{y}, \mathbf{r}]$ triples, and forbids executing code. The program channel requests a program returning a packing. Supplement S3 gives the bare prompt and hash; the artifact contains the mutation and program contracts.

Recipe family. Truncated grids $T(k, N)$ and grids with fillers $V(k, m)$ (§1) form the family; up to $(k - 1)^2$ interior vertices may carry fillers. A template specifies one such construction; Table 2 collects core terms.

Family maximum. The *family maximum* F_N (“family argmax” in the registrations) is a scalar: the family’s best score with exactly N circles. A *maximizing construction* attains it:

$$F_N = \max_{1 \leq k \leq \lceil \sqrt{N} \rceil} \begin{cases} N/(2k), & N < k^2, \\ k/2 + (N - k^2)(\sqrt{2} - 1)/(2k), & 0 \leq N - k^2 \leq (k - 1)^2. \end{cases} \quad (1)$$

Orders satisfying neither condition lack filler capacity and are excluded. Larger truncation orders score lower and cannot improve this maximum.

Anchor value. The *anchor construction* is the template at grid order $k^*(N) = \text{round}(\sqrt{N})$; its score is the *anchor value* A_N . When $N \geq k^{*2}$, the k^* grid extends with $N - k^{*2}$ fillers and the anchor value is $A_N = V(k^*, N - k^{*2})$. When $N < k^{*2}$, it truncates and $A_N = T(k^*, N)$; this branch covers $N \in [k^{*2} - k^* + 1, k^{*2} - 1]$. A tested cell is *discriminating* when $A_N \neq F_N$. Only these cells separate matching the anchor from matching the family maximum. At the tested discriminating cells, the maximizing construction is the lower-order grid with fillers. That is not so at every N : Supplement S5 has a counterexample at $N = 60$ and the full check.

Worked cells. At $N = 13$, $k^* = 4$ and the anchor truncates: $A_{13} = T(4, 13) = 13/8 = 1.625$. The family maximum extends the 3×3 grid with four fillers: $F_{13} = V(3, 4) = 1.5 + 4(\sqrt{2} - 1)/6 = 1.7761424$. At $N = 31$, $k^* = 6$, $A_{31} = T(6, 31) = 31/12 = 2.5833333$ and $F_{31} = V(5, 6) = 2.5 + 6(\sqrt{2} - 1)/10 = 2.7485281$ (Figure 1a,b). Of the seven square cells, where arm P requests direct emission without a parent (§6), $N = 13, 21, 31, 43$ are discriminating. At $N = 17, 35, 37$ the two values agree, including the truncated anchor at 35.

The premise in this paper’s data. The premise rests on our own no-parent outputs, all at the bare prompt, chiefly the loops’ generation 0. Through the agent runtime (§3), the original loop’s valid seeds (proposals without a parent) match the anchor in 4 of 9 at $N = 13$ and 7 of 8 at $N = 31$. In the loop replication they match it in 27 of 42 and 32 of 42. Through the pinned API with thinking on, the anchor is the plurality at $N = 13$, 5 of 12 (§6). No arm here samples the bare prompt at $N = 21$ without a parent through the agent runtime, and the pinned API directly emits no valid packing there. Post hoc, other layouts, including hexagonal rows, also reach the anchor value; the loop analysis (§5) tracks score and construction.

Published bounds. Friedman’s Packing Center (Friedman, n.d.) lists best-known sums of radii of 1.829+ at $N = 13$ and 2.889+ at $N = 31$, both found by D. W. Cantrell in 2011. So F_{13} and F_{31} lie at least 0.052 and 0.140 below the best known packings. Clearing F_N means leaving the recipe family, not improving the state of the art.

Table 2: Core terms.

Term	Operational meaning
Anchor / anchor match	Anchor value A_N ; an output is an anchor match (“on-prediction” in the registrations) when its sum S satisfies $ S - A_N < 2 \times 10^{-3}$. This does not certify geometry, and “anchor” names no cognitive mechanism.
Anchor construction	Template at grid order k^* that scores A_N . At a truncating cell, any N equal circles of radius $1/(2k^*)$ also score A_N .
Family maximum	Scalar F_N in Eq. 1; distinguished from its maximizing construction.
Conformance	Coordinates and radii lie within coordinate tolerance ϵ of a construction. Anchor-construction conformance, to the anchor construction, is reported beside the registered anchor-match rate; it was chosen post hoc. Family conformance, to any feasible family construction, is the Conform column of Table 7; at a discriminating cell a maximizing construction conforms to the family, not to the anchor construction.
Cell / discriminating	Tested circle count N ; discriminating when $A_N \neq F_N$.
Valid at τ / clearance	Correct count, finite coordinates, positive radii, containment and non-overlap within geometric tolerance τ . Clearance: valid at 10^{-9} and excess $S - F_N$ above the margin, 10^{-6} .
Serving configuration	Serving path together with request settings and inherited context. A pinned request does not pin model weights.
k -inheritance	Dominant-radius grid-order estimate equals the parent’s; a radius-based summary, not a coordinate-conformance test.

An independent re-score, whose parser was revised once after a first run, agreed on 129 of 135 cell counts (Supplement S1). These include arm P’s three program cells, where it only compared the frozen report’s sums with F_N .

3 Registration, serving configurations, and evaluation

Registration. Each arm has a committed design, prompt digest, and registered predictions whose decision rules, usually numeric thresholds, the data meet or miss; the loop replication registered directions without thresholds (Supplement S7). A falsifier is a registered condition against the arm’s hypothesis; it fires when met. Codes such as P-MU1 or L1 name predictions in tables. Supplement S1 maps the artifact’s original files, amendments, and raw outputs to results; S2 lists deviations. Earlier observations motivated some registrations: arm P-D followed arm P’s validity collapse, L’s predicted fall followed MU’s result, and V added aliases after its original run began. These are sequentially registered extensions, not an entirely outcome-independent study. Original results and post hoc analyses are kept separate. A registered threshold crossing is not itself a calibrated significance test.

Registrations are git commits, not public timestamps; commit ancestry fixes their recorded order. The MU, P and P-D registration files assert a remote push that the anonymized package cannot show, and the MU, L and V ledgers (per-call records) lack row timestamps (Supplement S1). For those arms, registration before sampling rests on our records. The loop replication (§5) was committed only locally before sampling. OpenTimestamps stamped the artifact manifest hash and source archive on 2026-09-16; a stamp proves existence by its attesting block, not sampling order.

Serving configurations. MU, L and the loop replication ran through an agent runtime: a Claude Code subagent requested as Haiku. Only the replication records the returned model identifier (Table 3); the MU and L ledgers record none. MU and L can be re-scored from their ledgers but not re-drawn; registered pinned-API re-draws of both were not run (Supplement S2). Runtime version and inherited context vary across and within runs. The original L run used version 2.1.246, and the replication changed from 2.1.260 to 2.1.266 during collection; MU’s version is unrecorded (Supplement S7). Outside the hashed prompt, the runtime sends tool and skill listings, which the original L transcripts record, and a subagent system prompt they do not show. MU’s context is unrecorded and presumed the same. Those transcripts carried no instruction files, but every replication proposer inherited three operator files: global agent instructions, an unrelated browser-tool skill and an auto-memory index. They are withheld from the artifact as identity-bearing; their digests are logged. By author assertion, none contains a hypothesis, formula, outcome count or packing; the memory index names a research-paper operations memory and an “agent-as-proposer harness” (Supplement S7). The version change added attachments, logged only by type and length, to 260 of the replication’s 600 slots. In 8 of the 10 lineages at $N = 31$, the change falls between seeds and descendants. Post hoc, the anchor-match rate still falls in all 12 lineages run on one version (ten at $N = 13$, two at $N = 31$; Supplement S7). In all 12, logged attachments change after generation 1: the version is fixed, the context is not. Runtime transcripts show extended thinking on, and no replication slot records a tool use. MU’s wrapper asked for no tool use, but its ledger has no tool-use field. That context and the decoding settings cannot be reconstructed from the artifact, and the alias-to-weights binding is unattested. Results from MU, L and the replication therefore describe the agent runtime, not a bare API call.

P uses OpenRouter’s `claude-haiku-4.5` chat-completions route at temperature 1.0 with no system prompt; P-D modifies its reasoning and system-message settings. Table 3 lists each model with its identifier. Other-model serving paths and floors are in Supplement S4. Runtime and pinned-API counts are not pooled.

Local evaluator. A parsed packing must contain exactly N numeric triples with finite entries and $r_i > 0$. At geometric tolerance τ , containment requires $x_i - r_i, y_i - r_i \geq -\tau$ and $x_i + r_i, y_i + r_i \leq 1 + \tau$; pairwise validity requires $\|c_i - c_j\|_2 + \tau \geq r_i + r_j$. The score is $S = \sum_i r_i$. We report $\tau = 10^{-6}$ (the original primary tolerance) and 10^{-9} . These are numerical checks, not exact-arithmetic certificates. A *slot* is one planned call of a registered design. Parsing failures, invalid geometry and refusals are retained and count against slot validity; P-D also reports completed-response counts.

Anchor-match rates (value matching) use the matching window (Table 2) among outputs valid at 10^{-6} . Clearance asks whether an output improves on the family maximum; original clearance predicates differed across arms, and the supplement preserves each. The *harmonized post hoc audit* applies the clearance rule of Table 2 to every arm and separately counts any positive excess ($S - F_N > 0$) and excess above the margin among outputs valid at 10^{-6} .

Table 3: Models, identifiers and serving configurations; no identifier attests model weights.

Model	Identifier	Serving configuration (arms)
Claude Haiku (4.5 where recorded)	Agent runtime: requested as Haiku; the replication runtime returns <code>claude-haiku-4-5-20251001</code> . Pinned API: <code>anthropic/claude-haiku-4.5</code>	Agent runtime, a Claude Code subagent (MU, L, loop replication); pinned API through OpenRouter (P, P-D)
Gemini 2.5 Flash-Lite	<code>gemini-2.5-flash-lite</code>	Vendor’s first-party API; 85 of 140 rows routed through OpenRouter by amendment (GM)
Gemma, 26B mixture of experts	<code>gemma-4-26b-a4b-it</code>	Vendor’s first-party API; 41 of 140 GM3 rows routed through OpenRouter by amendment (GM2, GM3)
Thirteen aliases, several vendors	Free-tier aliases, listed in Supplement S4	Two free-tier serving paths (V)

Cells and floors. A floor is the fewest valid outputs a registered verdict needs; a cell, alias or lineage that reaches it is scoreable. Held-out cells 50, 58, 62, 65, 75, excluded from the companion manuscript’s fit (Supplement S5), are five further cells that arm P tests without a parent. P’s code cells, where it requests programs, and MU’s conditioned cells are 13, 21, 31; L uses 13, 31. P requires five valid outputs per cell; P-D’s registered thresholds count valid slots of 15 (§6). L requires, per lineage, three valid seed proposals and a valid proposal in three of five conditioned generations; all four original lineages pass (Supplement S7 gives the replication’s scoreability). Floors govern registered verdicts, not the post hoc audit. Empty denominators receive no imputed rate.

Data ownership and uncertainty. This paper is the primary report of the eight original arms and the later loop replication; the companion manuscript cites the original findings as support. Confidence intervals in independent-call summaries are nominal Wilson binomial intervals, conditional on the sampled configuration and validity. Pooled MU intervals ignore differences between cells, and no such interval quantifies variation across adaptive loop replications. Supplement S1 retains the pooled loop calculations with this limitation, and §5 reports the trajectories descriptively.

4 One parent in the prompt: arm MU

Arm MU gives the proposer one parent and one instruction, with no iteration, through the agent runtime (Table 4). Given the maximizing parent, all 26 valid outputs keep or improve its score, and none is an anchor match. Given the anchor parent, 7 of 28 are anchor matches and 17 move above it. Given a far-below off-family parent, 20 of 34 land at the anchor value: 12 of 14 at $N = 31$, 7 of 13 at 21 and 1 of 7 at 13.

Design. Each call is the bare prompt, an existing packing with its score, and the instruction “Propose a modification of this packing that increases the sum of radii.” An output keeps or improves on its parent when it scores no more than the 2×10^{-3} matching window below it. The *anchor parent* is built at the anchor value $T(k^*, N)$. The *maximizing parent* sits at the family maximum $V(k^* - 1, m)$, provably better than the anchor at these cells. As given, it fails the 10^{-9} gate at all three cells, through overlaps below 10^{-6} . The *off-family parent* is a diagonal-chain packing scoring far below either. Each parent runs at $N = 13, 21$ and 31 with $n = 15$ calls per cell and parent, 135 calls in all. MU’s ledger lists the parent conditions largely in sequence rather than interleaved, without timestamps.

The maximizing parent meets the anchoring-dissolved rule. The arm tests whether unconditioned anchoring transfers to parent-conditioned calls. The registered decision rule calls anchoring dissolved when, under this parent, the pooled anchor-match rate is at most 20% and the keep-or-improve rate at least 50%; the falsifier fires under the same condition. Both conditions are met: 0 of 26 valid outputs are anchor matches, and 26 of 26 keep or improve. The falsifier (F-MU1) therefore fires, counting against transfer, and the prediction of at least 35% anchor matches (P-MU2) is missed. The maximizing parent outscores the

Table 4: Arm MU by parent condition; arm L’s generation 0 (§5), a separate run that sends the bare prompt, with no parent or request, through the same dispatch wrapper (Supplement S3), is the no-parent reference, not a matched control. Valid: at 10^{-6} , of 45 calls per MU parent. Brackets: 95% Wilson intervals on the preceding rate. The threshold column gives registered thresholds only; verdicts sit with the anchor-match counts. Codes name registered predictions (Supplement S1).

Parent	Valid	Anchor match (95% Wilson); verdict	Registered threshold	Score against parent	k inherited
None (arm L generation 0)	17	4/9 at $N = 13$; 7/8 at $N = 31$; no row at $N = 21$; unconditioned reference	—	—	—
Maximizing parent (family maximum)	26	0/26 (0%, [0–13%]); dissolved rule met	dissolved if $\leq 20\%$ (with keep-or-improve $\geq 50\%$)	26/26 keep or improve (100%, [87–100%]); 25 equal within 10^{-6} , 1 above by 1.2×10^{-6}	26/26
Anchor parent (the anchor construction)	28	7/28 (25%, [13–43%]), missed; per cell 4/12, 1/5 and 2/11 at $N = 13, 21$ and 31	$\geq 50\%$ (P-MU1)	17/28 above the anchor, 4/28 below	22/28
Off-family parent (diagonal chain, sum ≈ 0.6)	34	20/34 match the anchor (59%, [42–74%]); per cell 1/7 [3–51%], 7/13 [29–77%] and 12/14 [60–96%] at $N = 13, 21$ and 31; proxy met in two of three cells	modal output on the k^* grid (P-MU4)	—	—

anchor $T(k^*, N)$ by 0.151, 0.159, and 0.165 at $N = 13, 21$ and 31; each gap is over seventy times the same 2×10^{-3} window. So 0 of 26 follows from keep-or-improve alone.

Post hoc, the outputs sit much closer to the parent than keep-or-improve requires. Of the 26, 25 return the parent’s score within 10^{-6} , and one exceeds it by 1.2×10^{-6} . In coordinates, 18 reproduce the parent within 10^{-6} everywhere. These include the output 1.2×10^{-6} above it, whose excess comes from sub-tolerance radius perturbations. One output mirrors the parent, and 7 relocate fillers but keep the radius multiset and the parent’s sum to seven decimals. All 26 inherit the parent’s grid order, so the prediction that k -inheritance falls below 50% is missed (P-MU3). Without a matched instruction-free control, these outputs do not separate recognizing an optimum from echoing the input.

The anchor parent is not self-stable. The registered prediction was an anchor-match rate of at least 50%; the observed rate is 7 of 28 (missed). Instead, 17 of 28 move above the anchor and 4 fall below it. Of the 28, 22 inherit the parent’s grid order, 79% (60–90%). The 6 that change it all move above the anchor, and they include 3 that reach the family maximum value.

Off-family parent: met in two of three cells under the registered proxy. The registration predicted that the modal valid output under this parent sits on the k^* -grid. MU’s frozen scorer scores it per cell by dominant-radius order k^* within the most frequent score bucket. This proxy is met at $N = 21, 31$ and missed at $N = 13$. Seven of its 20 anchor matches (of 34 valid outputs) fail the post hoc family-conformance test (Supplement S6). Post hoc, we read the anchor value as a fallback from a far-below parent, not a ceiling, since anchor and maximizing parents are mostly kept or improved.

5 Inside a search loop: arm L

Both loops in this section ran through the agent runtime. The registered measure is the anchor-match rate. It is lower in descendants than in seeds in each of arm L’s four original lineages (11 of 17 valid seeds, 11 of 49 valid descendants) and in all 19 scoreable replication lineages. We also report anchor-construction conformance (post hoc), because a score match does not identify the construction. Its share falls in each original lineage (Table 5). In the replication, 21 of 84 valid seeds but 8 of 273 valid descendants are the anchor construction at coordinate tolerance 10^{-6} (22 and 8 at 10^{-3}).

Design. Two discriminating cells, $N = 13, 31$, are crossed with GREEDY (best valid proposal retained) and DIVERSE (up to five best distinct-valued proposals retained). Each lineage has a fresh unconditioned generation 0 and five conditioned generations of population 5. Two runtime rejections, excluded and counted by the registration, are invalid slots here. The mutation prompt is MU’s, with the parent’s coordinates and score. Parent selection cycles deterministically through the archive. The loop has no islands, crossover, program representation, or evaluator feedback beyond the parent’s score.

Table 5: The four original lineages. Validity is at 10^{-6} ; best scores to seven decimals. Among valid proposals, Constr. counts the anchor construction at coordinate tolerance 10^{-6} (post hoc; identical at 10^{-3}) and Match counts anchor matches. Best (first gen.): best score and the generation in which it first appears. Fractions are descriptive; descendants are not independent replications.

Lineage	Seed valid	Constr.	Match	Desc. valid	Constr.	Match	Best (first gen.)	F_N
13-GREEDY	5/5	1/5	2/5	4/25	0/4	0/4	1.7761424 (0)	1.7761424
13-DIVERSE	4/5	1/4	2/4	16/25	2/16	5/16	1.7761424 (5)	1.7761424
31-GREEDY	4/5	3/4	3/4	18/25	0/18	3/18	2.6104167 (5)	2.7485281
31-DIVERSE	4/5	3/4	4/4	11/25	0/11	3/11	2.7485281 (2)	2.7485281

Registered outcomes. The prediction that the modal valid seed matches the anchor in both cells, which the committed scorer checks per lineage, is missed (L1; Supplement S2). In 13-DIVERSE, two valid seeds match the anchor and two share another score, a tie the scorer counts against the prediction. The rate prediction pools both archive regimes over generations 1–5 and requires a lower anchor-match rate than at generation 0 in both cells. It is met descriptively (the rule compares rates, with no significance test): 4/9 falls to 5/20 at $N = 13$, and 7/8 to 6/29 at $N = 31$. The falsifier, a non-decrease in both cells, does not fire (Supplement S1). The archive-regime prediction, that GREEDY’s descendant anchor-match rate exceeds DIVERSE’s in at least one cell and is lower in neither, is missed in both comparisons: $0/4 < 5/16$ at 13 and $3/18 < 3/11$ at 31. No direction was registered for best-of-run relative to F_N . With one lineage per cell–regime pair, these data show only that the anchor was not restored, not an archive-width effect or a population clearance rate. Validity also falls, from 17/20 seed slots to 49/100 descendant slots. Descendants share parents.

Subsequent replication. A replication in 20 fresh lineages, five per cell–regime pair, supports the original direction. Each proposal had a fresh conversation but inherited the files of §3; selected parents still induce dependence within lineages. Of the 100 seeds and 500 descendants, 84 and 273 are valid at 10^{-6} . By score, the anchor-match rate is lower in descendants than in seeds in all 19 scoreable lineages and in the 20th (Table 6). Supplement S7’s post hoc sign test on this direction checks only a weak null. Since most descendants see parents above the anchor value (post hoc), keeping or improving parents alone would lower the pooled rate; the copy baseline below is more informative. Over all planned slots, anchor matches fall from 59 of 100 seeds to 19 of 500 descendants. Seeds and descendants differ in more than the parent: descendants add the request, and runtime version or context changes within lineages (§3). Post hoc, shown a parent that matches the anchor, 11 of 49 valid replication descendants and 7 of 28 valid MU outputs match it. Without a parent, 59 of 84 valid seeds do. The anchor-construction share is lower in descendants than in seeds in 14 of the 20 lineages, zero in both in 6 and higher in none (Supplement S7 splits the 19 scoreable ones). Only 21 of 59 anchor-matching seeds and 8 of 19 anchor-matching descendants conform.

Copying (post hoc). Had each of the 273 valid replication descendants copied its parent, 49 would be anchor matches; 19 are. Of the 273 valid descendants, 140 (51%) return their parent’s score within 10^{-6} , 118 (43%) improve on it (score more than 10^{-6} above it) and 15 (5%) score lower, so 258 of 273 keep or improve at 10^{-6} , a stricter test than the 2×10^{-3} window of §4. Of the 118, 103 improve by at least 2×10^{-3} . In 97 (36%), every coordinate repeats the parent’s. Given a parent that matches the anchor, 38 of 49 valid descendants improve on it, 9 return its score and 2 score lower. Eleven remain anchor matches: the 9 that return its score and 2 improvements inside the matching window. Given a parent that scores above the anchor value, 2 of 202 are anchor matches, both by scoring lower than that parent. The other 22 have parents below the anchor value; 6 of them are anchor matches, all by improving. The score-copy share rises from 33% in generation 1 to between 51% and 65% in generations 2 to 5. These counts are an accounting, not a causal decomposition, and an instruction-free control was not run. Their direction is consistent with the single-parent pattern of §4. k -inheritance was not computed for loop descendants.

No raw clearance, and the exact construction. No raw proposal clears F_N in any of the 20 lineages. Among proposals valid at 10^{-6} , 14 raw proposals exceed F_N by more than 10^{-6} , the margin. None of them is valid at 10^{-9} . Eleven exceed it by at most 6.0×10^{-6} ; the other three, all from lineage 31-DIVERSE r3, by 0.024 to 0.027. Their shared construction fails the 10^{-9} gate only through nine pairwise overlaps of about 8.27×10^{-9} between two adjacent rows. It lies outside the recipe family. In exact arithmetic (post hoc), the construction behind the generation-4 proposal (Figure 1c) is feasible at zero tolerance, and its sum, $61/22$, exceeds F_{31} by 0.0242. The proposal writes two row heights with arithmetic slips of 9.5×10^{-9} and 9.1×10^{-12} , not rounding of 17-decimal values; the larger causes the nine overlaps. The construction comes from one lineage and is not a packing record.

A deterministic, post hoc, radius-only repair, applied to every eligible proposal in the corpus, makes four proposals from two $N = 31$ DIVERSE lineages valid above $F_{31} = 2.7485281$. It keeps every center fixed and scales all radii by the largest uniform factor $\alpha \leq 1$ that removes every overlap and boundary violation, less a 10^{-12} safety offset (Supplement S7). Three are in r3, a generation-4 proposal and two of its generation-5 children, with repaired scores 2.7727271, 2.7757933, and 2.7757953 and $1 - \alpha = 4.55 \times 10^{-8}$ each. Each stays at least 0.024 above F_{31} . The fourth, in r5, has repaired score 2.7508339 and $1 - \alpha = 2.42 \times 10^{-4}$. The best repaired score remains 0.113 below the best known 2.889+ (Friedman, n.d.). Repaired proposals never enter registered counts or the executed archives; Supplement S7 lists every deviation and repair.

Table 6: Original loops and subsequent replication, kept separate. ΔA is the descendant-minus-seed difference in anchor-match rate among proposals valid at 10^{-6} . Raw clear. counts lineages with a raw proposal that clears F_N (Table 2), a registered replication outcome (Supplement S7). The last column counts post hoc repaired clearances, with contributing lineages in parentheses; it is not a registered outcome.

Corpus	Slots	Scoreable / all	$\Delta A < 0$ / all	Raw clear.	Repaired
Original L	120	4/4	4/4	0/4	0 (0)
Replication	600	19/20	20/20	0/20	4 (2)

6 Serving configurations and other models: arms P, P-D, GM–GM3 and V

For one requested model, bare prompt and day, validity at $N = 13$ differs sharply by serving configuration.

Arm P. The pinned API, with thinking off, receives byte-identical registered prompts, 15 calls in each of seven square, five held-out, and three code cells; six credit refusals, none at a square cell, count as invalid. No square-cell output is valid at either tolerance (0/105). Most fail by overlap (median maximum overlap 0.10 at $N = 13$; taxonomy in Supplement S5). Held-out validity is 4/75, all at $N = 50$ with sum 2.5.

Two post hoc diagnostics at $N = 13$ leave the cause open. The same bare prompt, sent the same day through the agent runtime with its dispatch wrapper, gives six valid of six, four of them anchor matches. Through the pinned API, temperatures 0.0, 0.5, and 1.0 give zero valid outputs of six per setting. The two configurations differ in system context, wrapper and decoding at once.

Arm P’s registered predictions. Direct emission has two anchor-mode predictions and a falsifier; programs have a no-clearance prediction and the opposite clearance prediction. No direct-emission cell reaches the five-valid floor. We therefore read both anchor-mode predictions as unscorable and the falsifier’s denominator as empty. This post hoc scoreability correction is not the original verdict: the frozen scorer recorded the second anchor-mode prediction as met (Supplement S5). The original program report, replayed post hoc (Supplement S6), gives 12/45 valid, with only $N = 21$ reaching the floor. No valid program clears: the no-clearance prediction is met, its opposite missed.

Arm P-D. After P, a new registration fixes two alternating $N = 13$ conditions on the pinned API. Thinking-on enables extended thinking with the routing layer’s default effort and no system message. Thinking-off disables it and adds the system message “You are a careful assistant. Think before you answer.” Both share the model alias, bare prompt, temperature, and output limit. The first prediction requires at least eight valid thinking-on slots and at most three valid thinking-off slots, out of 15 each. Competing predictions reverse those thresholds or cover other patterns. A secondary prediction, with no verdict weight, expects the anchor $T(4, 13)$ to be modal among valid outputs in whichever condition reaches eight valid slots.

Of 14 completed thinking-on responses, 12 are valid at 10^{-6} (11 at 10^{-9}); of 13 thinking-off, none. An amendment closes the run there after refusals for lack of account credit (Supplement S5). The first prediction is met as registered, and missing completions cannot reverse it: thinking off could reach at most two valid slots. The anchor $A_{13} = 1.625$ is the thinking-on plurality, 5/12, meeting the secondary prediction.

The registered label “the reasoning budget is the cause” names a prediction, not a causal effect. The two conditions change two features together, and arm P’s earlier thinking-off, no-system run cannot exclude interactions or time-dependent serving changes.

Other models. Four more registered arms ask, without a parent, whether the anchor transfers to other models (Table 3). Supplement S4 gives their designs, verdicts, amendments and Figure S1. GM returns 20 valid outputs of 140 and only three of the five scoreable cells its verdict needs, so it is underpowered. GM2 gives Gemma a 4,096-token output budget, and none of 140 responses parse. GM3 reruns Gemma with 16,384 tokens. Its pooled anchor-match rate, 46 of 80, meets the registered 30% threshold. The per-cell anchor-mode and two-radius-signature predictions are missed. Post hoc, 35 of the 46 matches lie at $N = 17$ and 35, two of the three cells where the anchor equals the family maximum. At the three scoreable discriminating cells, the modal score is the family maximum. In arm V, ten of thirteen free-tier aliases fall below the floor. Two of the three scoreable aliases meet the anchor-majority prediction at point estimates whose intervals include 50%; the third has no anchor match among 15 valid outputs. No output of these arms clears the family maximum (Table 7).

7 Post hoc numerical and coordinate audit

Leaving the family’s geometry is not the same as exceeding F_N . In the original corpus, 116 outputs valid at 10^{-9} miss the family’s coordinates even at 10^{-3} , yet all score strictly below F_N (Table 7, column D_9).

We replay the frozen ledgers and separately check coordinates against the family of §2. The audit changes neither registrations nor verdicts. It covers all 1,255 original planned slots, including aliases below the validity floor; the 600-slot replication is reported separately (§5, Supplement S7). Supplement S6 gives definitions, per-arm and loop details, and replay checks.

Numerical test. No original output clears F_N . We sum emitted radii at full precision and recompute F_N over every admissible order. Three MU outputs valid at 10^{-6} exceed the margin; none remains valid at 10^{-9} (MU’s seven-decimal excess counts: Supplement S2). One GM3 output valid at 10^{-9} has a positive excess, 5.29×10^{-10} , below the margin.

Radius-only repair. No repaired packing exceeds F_N in any original arm or cell. The Supplement S7 rule is applied to every parsed original output with N finite positive-radius triples, at any validity: 651 outputs, 625 with a positive factor α . After repair, the three MU outputs above the margin end 1.4 to 2.4×10^{-6} below F_N . The GM3 excess becomes -4.7×10^{-10} .

Table 7: Post hoc audit of the eight original arms. v_6, v_9 : valid at geometric tolerance $10^{-6}, 10^{-9}$. C_9 : valid at 10^{-9} with excess above the 10^{-6} margin. Conform and Nonconform split v_6 at coordinate tolerance $\epsilon = 10^{-6}$. D_9 : valid at 10^{-9} yet nonconforming at $\epsilon = 10^{-3}$; all score below F_N . Slots include failures. Counts are an inventory across separate serving configurations, not a pooled transfer rate. Arms GM, GM2, GM3 and V are described in Supplement S4.

Arm	Slots	v_6	v_9	C_9	Conform	Nonconform	D_9
MU	135	88	63	0	47	41	39
L	120	66	63	0	23	43	36
P	225	16	15	0	6	10	9
P-D	30	12	11	0	3	9	9
GM	140	20	20	0	11	9	9
GM2	140	0	0	0	0	0	0
GM3	140	80	79	0	72	8	4
V	325	48	42	0	30	18	10
Inventory total	1,255	330	293	0	192	138	116

Coordinate test. Of the 330 outputs valid at 10^{-6} , 138 do not conform to the specified axis-aligned family. Conformance requires coordinates and radii within coordinate tolerance $\epsilon = 10^{-6}$ of a feasible construction and distinct sites; fillers require the full grid. Valid off-grid layouts and rounding can both break conformance (examples in Supplement S6). Relaxing the coordinate tolerance to 10^{-3} leaves 129 nonconforming.

8 Related work

Search systems. Language Model Crossover (Meyerson et al., 2024) recombines text genotypes through few-shot prompting. ELM (Lehman et al., 2022), EvoPrompt (Guo et al., 2024), EoH (Liu et al., 2024), and LLaMEA (van Stein & Bäck, 2025) evolve programs, prompts, or heuristics with LLM-generated variation. FunSearch (Romera-Paredes et al., 2024) and AlphaEvolve (Novikov et al., 2025) prompt with earlier solutions and keep evaluated programs in a database. OpenEvolve (Sharma, 2025), ShinkaEvolve, GigaEvo, and AdaEvolve (Lange et al., 2026; Khrulkov et al., 2025; Cemri et al., 2026) are similar frameworks; HELIX and ThetaEvolve add reinforcement learning (Su et al., 2026; Wang et al., 2025). OPRO (Yang et al., 2024) puts earlier solutions and their scores in the prompt of a language-model optimizer. ReEvo (Ye et al., 2024) evolves heuristics with the model’s reflections on earlier ones. Both show the model earlier solutions, the setting that arm MU isolates. Under matched resources, simple baselines match or exceed tested code-evolution pipelines (Gideoni et al., 2026), and off-the-shelf Xpress and SCIP solvers improve packing incumbents (Berthold et al., 2026). Both motivate separating proposal quality from the search procedure.

Trajectories. Evolutionary search helps on a heuristic-design benchmark (Zhang et al., 2024), and a 15-model, eight-task trajectory study ties stronger optimizers to frequent incremental improvements and localized search (Zhang et al., 2026). Without selection pressure, in 87% of LLM-driven mutation chains, over 93% of mutations revisit a previously seen structural form (Gurkan et al., 2026). Arm L instead conditions on retained high-scoring constructions and separates the anchor-match rate from the best score.

Anchoring and diversity. Our anchor value names no cognitive mechanism. Anchoring studies examine initial information and biased hints in LLM judgments (Huang et al., 2025; Lou & Sun, 2024), within broader cognitive-bias evaluations (Malberg et al., 2025). LLM research ideas can be narrower than human ones (Chen et al., 2026), and open-ended responses homogenize (Jiang et al., 2025). Bin-packing reassessments examine evolved heuristics’ interpretability and generalization (Herrmann & Pallez, 2025; Sim et al., 2025).

Reliability. Larger, more instructable models give more plausible errors (Zhou et al., 2024), accuracy collapses at high puzzle complexity (Shojaee et al., 2025), and verifier benchmarks test geometry and constructive mathematics (Kim et al., 2026; Balunović et al., 2025).

Registration. Related practice preregisters before an eligible model is released (Thomas et al., 2026), blinds deployment forecasts to outcomes (Williams et al., 2026), offers preregistration formats for agent experiments (Vaccaro, 2026), locks physics diagnostics (Zhang, 2026), and freezes an audit protocol (Jia, 2026).

9 Limitations and reproducibility

Arm limits. MU, L and the replication describe a Claude Code subagent requested as Haiku, not a bare API call. Its full system context cannot be reconstructed, and its version and inherited context vary across and within runs (§3). Registrations are local git commits with no verifiable public timestamp before sampling. No instruction-free control was run: every call with a parent also carries the request for a higher sum, so the shift from no-parent calls is not separated from that request. No program-channel loop or loop at other circle counts was run, no numerical-optimizer baseline is reported here, and MU tests three cells. Its fallback reading is post hoc and rests on one far-below parent kind, mainly on $N = 31$ (12 of 14; 1 of 7 at $N = 13$; no arm here makes parentless $N = 21$ calls in this runtime). It also rests on scores: 7 of the 20 anchor matches fail family conformance. The loop adds little support: post hoc, only 6 of 22 valid replication descendants of below-anchor parents are anchor matches. L’s replication adds five lineages per cell–regime pair, but within-lineage proposals remain dependent. P-D compares two serving configurations, not an isolated thinking intervention. Cross-vendor differences mix model variants and serving paths (Supplement S4).

Scope and selection. Each prompt type has one fixed wording, so paraphrase robustness is untested; larger budgets, tasks, models and prompts remain open. Concentration could also reflect typicality (Zhang et al., 2025), memorization (Carlini et al., 2021), benchmark contamination (Golchin & Surdeanu, 2024), format sensitivity (Sclar et al., 2024), alignment-induced diversity loss (Kirk et al., 2024), or arithmetic tractability, named in MU’s registration (Supplement S1). We do not separate them. A non-clearing score does not imply family conformance, and no finite run sets a universal bound.

Adaptation and deviations. P’s temperature/runtime controls were chosen post hoc; its scorer was committed after sampling. F-MU1 fired, but its registered rule restricting the title and abstract to calls without a parent was not followed, because our question concerns calls with a parent (Supplement S2). MU’s and the loops’ copy counts, the loops’ anchor-construction counts, MU/GM3/V clearance analyses, GM3’s discriminating split, harmonized audits and the exact-arithmetic check are post hoc. GM and GM3 have serving-path amendments, GM3 an excluded timing call and a duplicate-row repair (Supplement S2), and L a scorer correction. Supplement S2 gives original rules and outcomes, S7 the replication’s local registration, deviations, and repair analysis. Disclosure does not remove multiplicity or selection effects.

Reproduction. The artifact’s BUNDLE_README.md maps ledgers, registrations, amendments, and frozen reports; REPRODUCE.md documents the post hoc audit and builds. P and P-D requests are repeatable, but alias/provider drift may change their distribution. Redacted registration copies lose original byte hashes; manifests distinguish original and anonymized files.

10 Conclusion

In the tested agent runtime, with a prompt asking for a higher sum of radii, single calls mostly keep or improve the scores of anchor and maximizing parents. From a far-below parent, 20 of 34 valid outputs land at the anchor value: 12 of 14 at $N = 31$, 7 of 13 at 21 and 1 of 7 at 13. In the loop, the registered anchor-match rate falls from seeds to descendants in all four original and all 19 scoreable replication lineages. Post hoc, most descendants see parents above the anchor value and rarely return to it, consistent with the single-parent pattern. Without an instruction-free control, neither result separates the parent from the request relative to no-parent calls. At $N = 13$, same-day validity ranges from none to most outputs across serving configurations. For this runtime and task, we suggest testing a proposer under the loop’s own parents, serving configuration and evaluator tolerance, and reporting construction conformance apart from score.

References

Anonymous authors. Capability or optimizer? What a zero-shot LLM proposer contributes before evolution-

- ary program search begins, on circle packing, 2026. Unpublished companion manuscript. An anonymized copy is available to the action editor on request.
- Mislav Balunović, Jasper Dekoninck, Nikola Jovanović, Ivo Petrov, and Martin Vechev. MathConstruct: Challenging LLM reasoning with constructive proofs, 2025. URL <https://arxiv.org/abs/2502.10197>.
- Timo Berthold, Dominik Kamp, Gioni Mexi, Sebastian Pokutta, and Imre Polik. Out-of-the-box global optimization for packing problems: New models and improved solutions, 2026. URL <https://arxiv.org/abs/2605.04850>.
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650. USENIX Association, 2021. URL <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>.
- Mert Cemri, Shubham Agrawal, Akshat Gupta, Shu Liu, Audrey Cheng, Qiuyang Mang, Ashwin Naren, Lutfi Eren Erdogan, Koushik Sen, Matei Zaharia, Alex Dimakis, and Ion Stoica. AdaEvolve: Adaptive LLM driven zeroth-order optimization, 2026. URL <https://arxiv.org/abs/2602.20133>.
- Ziyu Chen, Yilun Zhao, and Arman Cohan. Measuring the gap between human and LLM research ideas, 2026. URL <https://arxiv.org/abs/2607.01233>.
- Erich Friedman. Packing center: circles in squares, maximizing the sum of radii. <https://erich-friedman.github.io/packing/cirRsqu/>, n.d. Accessed 2026.
- Yonatan Gideoni, Sebastian Risi, and Yarin Gal. Simple baselines are competitive with code evolution, 2026. URL <https://arxiv.org/abs/2602.16805>.
- Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2308.08493>. arXiv:2308.08493.
- Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2309.08532v2>. arXiv:2309.08532.
- Can Gurkan, Forrest Stonedahl, and Uri Wilensky. Mutation without variation: Convergence dynamics in LLM-driven program evolution, 2026. URL <https://arxiv.org/abs/2606.05408>.
- Julien Herrmann and Guillaume Pallez. An in-depth study of LLM contributions to the bin packing problem, 2025. URL <https://arxiv.org/abs/2510.27353>.
- Yiming Huang, Biquan Bie, Zuqiu Na, Weilin Ruan, Songxin Lei, Yutao Yue, and Xinlei He. Understanding the anchoring effect of LLM with synthetic data: Existence, mechanism, and potential mitigations, 2025. URL <https://arxiv.org/abs/2505.15392>.
- Haozhe Jia. HindsightBench: A black-box behavioral audit protocol for parametric hindsight in time-indexed LLM decision tasks, 2026. URL <https://arxiv.org/abs/2607.18867>.
- Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025. URL <https://arxiv.org/abs/2510.22954>.
- Valentin Khrulkov, Andrey Galichin, Denis Bashkirov, Dmitry Vinichenko, Oleg Travkin, Roman Alferov, Andrey Kuznetsov, and Ivan Oseledets. GigaEvo: An open source optimization framework powered by LLMs and evolution algorithms, 2025. URL <https://arxiv.org/abs/2511.17592>.

- Jinwoong Kim, Rui Yang, and Huishuai Zhang. GeoBuildBench: A benchmark for interactive and executable geometry construction from natural language, 2026. URL <https://arxiv.org/abs/2605.13167>.
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2310.06452>. arXiv:2310.06452.
- Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. ShinkaEvolve: Towards open-ended and sample-efficient program evolution. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://arxiv.org/abs/2509.19349>. arXiv:2509.19349.
- Joel Lehman, Jonathan Gordon, Shawn Jain, Kamal Ndousse, Cathy Yeh, and Kenneth O. Stanley. Evolution through large models, 2022. URL <https://arxiv.org/abs/2206.08896>.
- Fei Liu, Xialiang Tong, Mingxuan Yuan, Xi Lin, Fu Luo, Zhenkun Wang, Zhichao Lu, and Qingfu Zhang. Evolution of heuristics: Towards efficient automatic algorithm design using large language model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32201–32223. PMLR, 2024. URL <https://proceedings.mlr.press/v235/liu24bs.html>. arXiv:2401.02051.
- Jiaxu Lou and Yifan Sun. Anchoring bias in large language models: An experimental study, 2024. URL <https://arxiv.org/abs/2412.06593>.
- Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in LLMs, 2025. URL <https://arxiv.org/abs/2410.15413>.
- Elliot Meyerson, Mark J. Nelson, Herbie Bradley, Adam Gaier, Arash Moradi, Amy K. Hoover, and Joel Lehman. Language model crossover: Variation through few-shot prompting, 2024. URL <https://arxiv.org/abs/2302.12170>. Published in ACM Transactions on Evolutionary Learning and Optimization.
- Alexander Novikov, Ng  n V  , Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery, 2025. URL <https://arxiv.org/abs/2506.13131>.
- Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6. URL <https://doi.org/10.1038/s41586-023-06924-6>.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2310.11324>. arXiv:2310.11324.
- Asankhaya Sharma. OpenEvolve: an open-source evolutionary coding agent. GitHub repository, <https://github.com/algorithmicsuperintelligence/openevolve>, 2025. Official repository README; accessed 2026-09-13.
- Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL <https://arxiv.org/abs/2506.06941>.
- Kevin Sim, Quentin Renau, and Emma Hart. Beyond the hype: Benchmarking LLM-evolved heuristics for bin packing, 2025. URL <https://arxiv.org/abs/2501.11411>.

- Chang Su, Zhongkai Hao, Zhizhou Zhang, Zeyu Xia, Youjia Wu, Hang Su, and Jun Zhu. HELIX: Evolutionary reinforcement learning for open-ended scientific problem solving, 2026. URL <https://arxiv.org/abs/2603.07642>.
- Maria Thomas, Kristina Gligoric, and Nihar B. Shah. Mitigating LLM-based p-hacking by preregistering for the next LLM, 2026. URL <https://arxiv.org/abs/2606.27687>.
- Michelle Vaccaro. Preregistration for experiments with AI agents, 2026. URL <https://arxiv.org/abs/2606.11217>.
- Niki van Stein and Thomas Bäck. LLaMEA: A large language model evolutionary algorithm for automatically generating metaheuristics, 2025. URL <https://arxiv.org/abs/2405.20132>.
- Yiping Wang, Shao-Rong Su, Zhiyuan Zeng, Eva Xu, Liliang Ren, Xinyu Yang, Zeyi Huang, Xuehai He, Luyao Ma, Baolin Peng, Hao Cheng, Pengcheng He, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. ThetaEvolve: Test-time learning on open problems, 2025. URL <https://arxiv.org/abs/2511.23473>.
- Marcus Williams, Hannah Sheahan, Cameron Raymond, Tomek Korbak, Deng Pan, Peilin Yang, Leon Maksin, Ningyi Xie, Phillip Guo, Ian Kivlichan, and Micah Carroll. Predicting LLM safety before release by simulating deployment, 2026. URL <https://arxiv.org/abs/2607.07184>.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2309.03409>. arXiv:2309.03409.
- Haoran Ye, Jiarui Wang, Zhiguang Cao, Federico Berto, Chuanbo Hua, Haeyeon Kim, Jinkyoo Park, and Guojie Song. ReEvo: Large language models as hyper-heuristics with reflective evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL <https://arxiv.org/abs/2402.01145>. arXiv:2402.01145.
- Dong Zhang. Testing frontier large language models’ physics literacy in parallel physical worlds, 2026. URL <https://arxiv.org/abs/2607.00276>.
- Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity, 2025. URL <https://arxiv.org/abs/2510.01171>.
- Rui Zhang, Fei Liu, Xi Lin, Zhenkun Wang, Zhichao Lu, and Qingfu Zhang. Understanding the importance of evolutionary search in automated heuristic design with large language models, 2024. URL <https://arxiv.org/abs/2407.10873>.
- Xinhao Zhang, Xi Chen, François Portet, and Maxime Peyrard. What makes an LLM a good optimizer? A trajectory analysis of LLM-guided evolutionary search, 2026. URL <https://arxiv.org/abs/2604.19440>.
- Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032): 61–68, 2024. doi: 10.1038/s41586-024-07930-y. URL <https://doi.org/10.1038/s41586-024-07930-y>.