SEARCH ALLOCATION IS NOT DETERMINED BY LOSS, VALIDITY, OR UNIQUE YIELD IN A $\mathbb{Z}_7$ LABORATORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Search allocation is not determined by loss, validity, or unique yield. Verifier-backed discovery is commonly scored by those three, yet none of them determines how much verified mass a search puts on a target family. In a $\mathbb{Z}_7$ laboratory with an exact verifier, withholding a two-variable family from pretraining yields zero verified proposals on six training seeds and on 12 further generation runs, even though every prompt contains $x{+}y{=}y{+}x$. When the family is included, twenty-four matched initializations all visit it, yet parse-conditional allocation runs from 4.6% to 72.4%. The decisive comparison is inside that band: Seed 8 has higher teacher-forced loss than Seed 2 (0.2526 vs. 0.2428) and allocates 34.4% rather than 4.1% (A_N^{par} 37.0% vs. 4.6%). The gap remains under penalty 1.0 (0.4%), a healed prompt prime (7.2%), no feedback (2.0%), six structural exemplars (1.3%), and temperature 1.0 without nucleus sampling (8.5%; 13.2% parse-conditional, against 46.2% and 41.9%). At $N{=}12{,}800$ the low-allocation seed recovers $49/50$ catalog keys while allocation stays 5.7%. Allocation A_N and parse-conditional allocation A_N^{par} are the quantities a verifier-backed report has to add.

1 INTRODUCTION

We report a controlled laboratory experiment on a core premise of LLM-based discovery systems: that pretraining a model on scientific knowledge enables stochastic generation to explore that knowledge. Recent discovery pipelines combine generative models with external verifiers to propose and test hypotheses in mathematics, protein design, and materials science (Romera-Paredes et al., 2023; Hubert et al., 2025; Gottweis et al., 2026; Fang et al., 2025), treating the base model as an exploration engine where learned domains should yield valid proposals under sampling.

In an additive group laboratory $\mathbb{Z}_7$ with an exhaustive symbolic verifier, catalog encoding, search allocation, and unique coverage come apart (Figure 1). Pretrained models with matching offline loss and syntax validity bifurcate between active exploration ($A_N \approx 35\%$–60%) and degenerate unary attractor collapse ($A_N \leq 5\%$), an $8.5\times$ efficiency gap masked by large-budget key saturation ($U_N \to 50/50$).

We study a generate–verify loop where a language model proposes candidate equations and an exact symbolic verifier scores truth. The loop has two knobs: the *pretraining catalog* C and the *runtime prompt* P. We isolate a target family $r^\star$: two-variable commutative additions $x + y \pmod 7$. We track three operational quantities: (1) *catalog encoding* L, measured by teacher-forced continuation loss on held-out statements of $r^\star$; (2) *search allocation* $A_N(r^\star)$, the fraction of N generated proposals that are valid equations in $r^\star$; and (3) *unique coverage* $U_N(r^\star)$, the number of distinct catalog keys discovered in $r^\star$ within budget N. Standard evaluations read L and U_N; search execution reads A_N.

Three empirically supported distinctions. *Necessity.* Withholding $r^\star$ from C yields $A_N(r^\star) = 0$ on six training seeds and on $12/12$ extra generation runs, including when the prompt contains $x + y = y + x$ (Table 2). Figure 2 is the shared loop, not the count. *Failure of sufficiency.* Putting $r^\star$ in C yields visits on all twenty-four seeds, with A_N^{par} from 4.6% to 72.4%. On all twenty-four seeds the step-4500 held-out probe stays in $[0.199, 0.255]$. In the direct pair, Seed 8 at 0.2526 allocates 34.4% ($A_N^{\mathrm{par}} = 37.0\%$) and Seed 2 at 0.2428 allocates 4.1% (4.6%). Seed 2 serves as an existence counterexample to sufficiency, staying in 3.8–7.5% across generation runs, and the gap survives

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

penalty 1.0, a healed prompt prime, static context, six structural exemplars, and temperature 1.0 without nucleus sampling (Appendix A.5, Figure 4). Crossing per-item exposure $R{=}D/B$ leaves the withheld arm at zero and the exposed arm visiting. U_N *is not* A_N. At tenfold budget the low-allocation seed reaches $49/50$ keys while A_N stays 5.7%. Key count at that budget is what a coupon-collector calculation produces. Allocation is the separate report.

Identity expansions, unary catalog breadth, extra samples, and feedback history reflect the same distinction under other knobs: unique counts and likelihood can decouple from visited families.

Scope and contribution. Search allocation is not determined by loss, validity, or unique yield. Exact verification makes the three separable, and the separation says what a report in this protocol has to add: A_N and A_N^{par}.

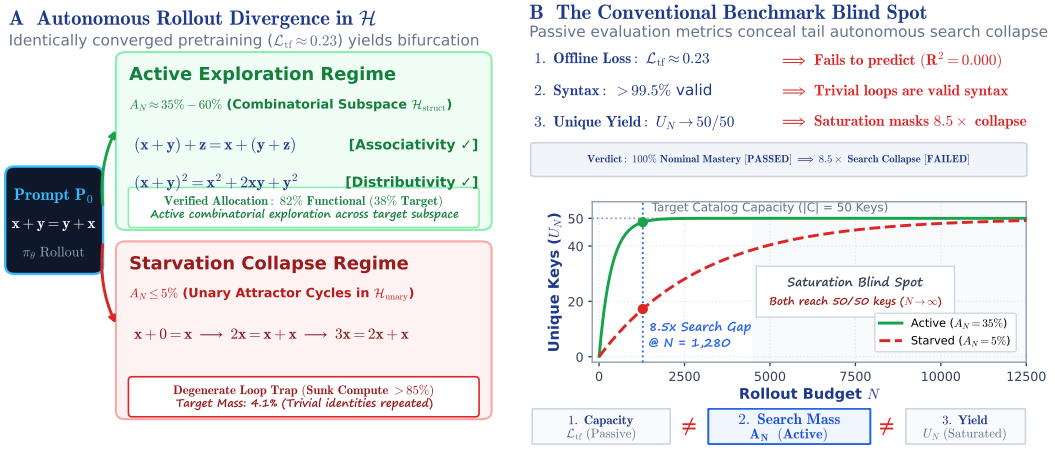


Figure 1: **The surface evaluation mirage versus generative search reality.** Pretrained checkpoints with matching offline loss bifurcate during autonomous rollouts between active combinatorial exploration ($A_N \approx 35\%$–60%) and degenerate unary attractor collapse ($A_N \leq 5\%$, an $8.5\times$ efficiency gap). Standard offline loss, syntax validity, and large-budget catalog recall conceal this search failure, necessitating separate audits of verified allocation (A_N).

2 Related Work

Training composition and learned capabilities. Compute-optimal training Hoffmann et al. (2022), data-mixture optimization Xie et al. (2023); Liu et al. (2025); Ye et al. (2025), and mixing laws Ge et al. (2024); Dai & Zheng (2026) establish that training composition affects learning. Repetition-aware scaling Sedova et al. (2026) further separates target exposure from reuse; its optimal reuse levels depend on data and compute, rather than being a blanket guarantee for a synthetic catalog. RL excursions during pretraining Bansal et al. (2026) show that composition also affects later policy learning and output distributions. These works primarily ask how composition changes learning or task performance. Mean token loss is already known not to be a sampling probability. We instead separate encoding L from allocation A_N and unique coverage U_N under an exact verifier and a catalog intervention, and ask which of those links the intervention moves.

Knowledge deprivation and controlled generalization. Unlearn-and-Reinvent Zhao et al. (2026) tests whether models can recover algorithms after targeted unlearning, including the effects of hints and verification. SynthWorlds Gu et al. (2026) and FictionalQA Kirchenbauer et al. (2026) use constructed knowledge to study reasoning and acquisition. Our lab instead controls catalog exposure during training from scratch. A common structural prompt example is held fixed while pretraining coverage changes. This makes the outcome a test of conditional generation, rather than a claim that all access to a mathematical concept has been removed. Algorithmic toy domains also support controlled generalization studies Power et al. (2022); Doshi et al. (2024).

Scientific search and output concentration. FunSearch Romera-Paredes et al. (2023), AlphaProof Hubert et al. (2025), Co-Scientist Gottweis et al. (2026), AI-Newton Fang et al. (2025), and theorem discovery Ota et al. (2026) combine generators with external evaluation. Recent large-scale analyses find that research-agent ideas remain close to starting literature Tang & Yang (2026) and that prior knowledge can dominate search Redko et al. (2026). Contemporary reasoning architectures frequently employ deliberative search over reasoning trees, such as Tree of Thoughts Yao et al. (2023), Language Agent Tree Search Zhou et al. (2024), or Monte Carlo Tree Search guided by process reward models Snell et al. (2025). While teacher-forced loss is broadly known not to determine sampling distributions, prompt-conditioned continuation NLL on the target strings still matches across the diagnostic checkpoints ($L_P = 0.47$–0.57) while verified family mass runs from 4.1% on Seed 2 to 33.1–37.0% on Seeds 0–1, and across the twenty-four exposed initializations A_N runs from 4.1% to 59.8%. Under flat sampling, reaching the same key count costs several times more continuations on the low-allocation seed (Section 4.1). Low teacher-forced loss on the target strings therefore does not settle how much verified mass search will put on them. The catalog intervention identifies that gap for allocation, separately from string counts and from which function family is visited.

Feedback. Counterexample-guided agents Liu et al. (2026) and literature-grounded generation Jansen et al. (2026) use interaction to revise proposals. We distinguish static verification from feedback that actually changes the next proposal. Our feedback checks measure format and coverage on existing checkpoints; they do not establish a general limit on counterexample-guided search.

3 METHOD

3.1 A CONTROLLED MATHEMATICAL LAB

Our domain consists of universally quantified equations over the additive group $\mathbb{Z}_7$. Terms use variables X, Y, Z, zero, negation, scalar multiplication, and addition. Catalog enumeration permits at most three *operators per term*, not tree depth three. Truth is checked by enumerating every assignment to the variables occurring in an equation, at most 7^3 assignments. The verifier contains no learned component. The grammar and normalization rules are in Appendix A.1.

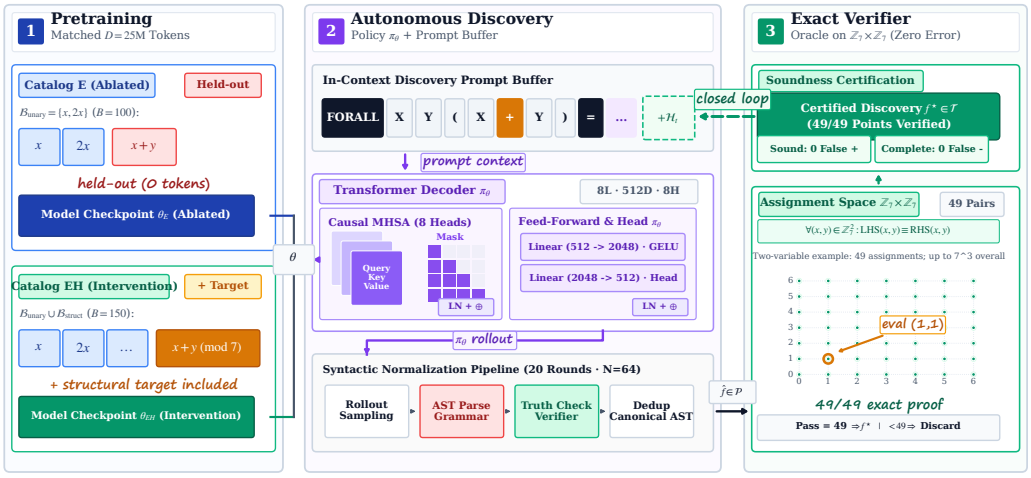


Figure 2: **Matched-budget generate–verify framework.** Pretraining catalogs E (withholding target family $x+y$) and EH (including it) share architecture, prompt, and decoding hyperparameters, evaluated by an exhaustive symbolic verifier on $\mathbb{Z}_7$.

Three measurement levels. We distinguish the number of rendered strings U_{surface}, the number of syntactically distinct equations modulo a *shared* variable permutation and side exchange $U_{\text{canonical}}$,

and the number of term functions U_{sem}. Every term in this DSL has a coefficient representation

$$\phi(t) = (a_X, a_Y, a_Z, d), \qquad t = a_X X + a_Y Y + a_Z Z + d \pmod 7.$$

Here $d = 0$ for the grammar used. For a true equation both sides have the same function; we use it to assign a *function region* (an invariant structural family). Its support is the number of nonzero variable coefficients. A support-2 region is structural in this operational sense, not an assertion of human difficulty. Function regions provide a coarse grouping of equivalence classes: different identities in one region need not convey the same mathematical relation. Named-variable function tables and variable-invariant equation keys are reported separately.

The measurement chain. Let C be the training catalog, a finite set of regions. The intervention of interest is whether a distinguished structural region $r^\star$ (here $\phi = (1, 1, 0, 0)$) lies in C. Write $I = \mathbf{1}\{r^\star \in C\}$: I=0 is arm E and I=1 is arm EH. The prompt protocol P and decoder are held fixed. We distinguish three downstream objects.

Encoding. For catalog strings s of region r,

$$L_{\text{tf}}(r \mid M) = \mathbb{E}_{s \sim \text{cat}(r)}[\text{NLL}(s; M)],$$

the standalone teacher-forced loss, and

$$L_P(r \mid M) = \mathbb{E}_{s \sim \text{cat}(r)}[\text{NLL}(s \mid P; M)],$$

the same strings scored as continuations of the discovery prompt. Specifically, L_P is evaluated conditioned on the static initial discovery prompt P_0 (containing the structural seed fact $x{+}y{=}y{+}x$ before history accumulation), directly measuring the model's baseline continuation prior at the onset of search. Low L means catalog strings are predictable as teacher-forced sequences. It is not the measure of the sampler.

Allocation. Let y_i be a raw continuation and $q(y_i)$ the parser output (empty if parse fails). For a true equation let $\rho(q)$ be the function region of both sides. With exact verifier V,

$$A_N(r \mid M, P) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[E_i(r)],$$

$$E_i(r) \equiv q(y_i) \text{ parseable, true, non-ground, and } \rho(q(y_i)) = r.$$

All raw continuations form the denominator: parse failures, false equations, and ground forms count as 0. Intuitively, $A_N(r)$ measures net proposal efficiency: the share of total sampling budget yielding verified discoveries in region r, treating syntax errors and false statements as wasted search compute. Formally, $A_N(r) = P(\text{parse}) \cdot P(\text{valid} \mid \text{parse}) \cdot P(r \mid \text{valid})$. We report parse-conditional mass $A_N^{\text{par}} = A_N/(1 - p_{\text{fail}})$ to isolate syntax failures from structural preference. Visit is $\text{Vis}_N(r) = \mathbf{1}\{A_N(r) > 0\}$, an operational finite-budget event rather than zero probability under all decoders.

Unique coverage. Let $\kappa(q)$ be the whole-equation joint key.

$$U_N(r \mid M, P) = \big|\{\kappa(q(y_i)) : E_i(r)\}\big|.$$

The 50-key catalog for $r^\star$ is the frozen list of structural statements; variable-renamed copies of one item are one key. Prompt copies of $x{+}y{=}y{+}x$ are counted separately. U_N can approach that catalog size while A_N stays small.

The queries, answered in Section 4, are: (i) when I=0, is $A_N(r^\star) = 0$? (ii) when I=1 and L_{tf}, L_P are low, is A_N determined? (iii) when N increases, do U_N and A_N move together? The main E versus EH contrast matches D. A crossed control also matches $R{=}D/B$ on three new seeds per arm (Table 4). The twenty-four original-recipe EH seeds share C, B, D, and therefore share nominal R.

Training exposure. All main models are trained for approximately 300M tokens. A mixed stream allocates D tokens to synthetic domain text and the remainder to C4. Catalog size is B, and $R = D/B$ is nominal domain-token exposure per catalog item, including template text. In the surface and semantic-breadth matrices, changing B at fixed R also changes D. T2 instead fixes D, so EH has lower R than E in the between-arm contrast. The crossed control tests both directions: the expanded B=150 E catalog scales the density of the existing unary families uniformly to 75 items each (75/class), preserving the unary-only structure without introducing $x + y$ equations while matching EH's lower per-item exposure $R = 166{,}667$. E at this lower R remains empty, while EH at E's higher R still visits the family. The twenty-four original-recipe EH seeds share C, B, and D, so they share nominal R; the within-EH allocation split is not an exposure-density comparison.

3.2 KNOWLEDGE CONFIGURATIONS AND CONTROLS

The workhorse is a GPT-style model with 41.8M parameters (8 layers, 8 heads, embedding width 512, context 512), referred to as 50M-class. Optimizer and architecture are shared across conditions: AdamW, peak learning rate 3×10^{-4} with linear warmup over 20M tokens (≈ 305 steps) followed by cosine decay to 3×10^{-5}, effective batch size of 128 sequences (65,536 tokens/step via per-device micro-batch 8 with gradient accumulation 16), weight decay 0.1, gradient clipping 1.0, and training dropout 0.05.

Configuration	B	$R = D/B$ (tok/item)	D	Train seeds
T2 E	100	250,000	25M	6
T2 EH	150	166,667	25M	24
T2 E matched R	150	166,667	25M	3
T2 EH matched R	150	250,000	37.5M	3
H-only check	50	500,000	25M	1
Identity $B \times R$	100, 1200	62,500, 250,000	BR	3
Fixed-B exposure	600	125,000, 250,000, 500,000	75M, 150M, 300M	3
Nested $S \times R$	$50S$	62,500, 250,000	BR	3

Table 1: **Controlled pre-training configurations on** $\mathbb{Z}_7$ (300M tokens each). T2 provides the primary matched-data comparison (D=25M); other rows vary batch size B and unary region count S to verify robustness.

Knowledge type (T2). E has 50 equations in each of the unary families x and $2x$. EH adds 50 equations in structural family $x + y$ while keeping total domain exposure at 25M tokens; exposure per catalog item consequently decreases. Other support-≥ 2 functions are held out from pretraining. A hard-only seed-0 check trains on the 50 structural equations with different two-variable definition/proof templates; it is exploratory rather than a matched-template third arm.

Surface and structural family coverage. The identity matrix varies B within the single function x. The nested semantic matrix uses 50 equations per function: $S = 2$ contains $x, 2x$ and $S = 6$ contains kx, $k = 1, \ldots, 6$. Each catalog contains frozen statements, shared across training seeds; variable-renamed copies do not count as extra catalog regions. Definitions, theorem/QA documents and catalog templates are documented in Appendix A.3. False and worked-false documents are excluded.

Measurements and controls. The chain readouts are L_{tf}, L_P, A_N (all N continuations), A_N^{par}, Vis_N, and U_N. Validity among parsed samples is labeled separately. Offline scanning of every parseable `FORALL` tests whether first-equation selection misses structural outputs. The 10-fold budget checks test additional sampling on fixed weights. No-feedback and short-history runs test whether changing P at fixed M moves A_N. Enumeration defines the catalog universe; no quantitative random-sampler performance claim is made.

3.3 GENERATION, VERIFICATION, AND SHARED CONTEXT

Each round generates 64 continuations, half at temperature 0.6 and half at 0.8, with top-$p = 0.95$, repetition penalty 1.1, and up to 96 generated tokens. The historical prompt string ends with a trailing space after `FORALL`; that space is not a token prefix of the training equations, and the penalty down-weights prompt tokens; all discovery runs use this shared decoder (Appendix A.5). The parser selects the first parseable non-ground equation (falling back to ground if needed); truth is checked before deduplication and appending to a rolling history (Appendix A.4). Ground equations contain no variables; excluding them does not constitute a general proof of non-triviality.

Every arm receives the prompt example $x + y = y + x$, including unary controls. Thus, "held out" in the lab refers strictly to *pretraining catalog exposure*, with contextual exposure controlled. T2 tests pretraining coverage conditional on this shared structural hint; we also count outputs excluding prompt copies.

Historical runs used an independent-side variable-renaming key for online deduplication; auditing raw outputs with whole-equation normalization confirms unique-true totals agree across all 48 available dumps. The EH initializations are four batches under the same recipe: seeds 0–5 are the diagnostic batch used for likelihood and trajectories, seeds 6–11 replicate access and allocation, and seeds 12–23 add allocation only. Likelihood statements below refer to seeds 0–11 unless noted. Appendix A.2 details compatibility and reproducibility corrections.

4 Results

4.1 Withholding a family withholds verified search

T2 tests the chain at matched D=25M. Both arms see $x+y = y+x$ in P. I=1 yields $\mathrm{Vis}_N(r^\star) = 1$ on all twelve EH seeds; I=0 yields $\mathrm{Vis}_N(r^\star) = 0$ on all six E seeds (Table 2, Figure 3). Across all 24 EH checkpoints at N=1,280, Figure 3a plots raw target allocation A_N (inner ring) and parse-conditional allocation A_N^{par} (outer ring). Allocation spans 4.1%–59.8% ($A_N^{\mathrm{par}} \in [4.6\%, 72.4\%]$), uncovering severe tail search starvation in Seed 2 (4.1%) alongside syntax degradation in Seeds 11 and 23 ($881/1280$ and $883/1280$ parse failures under converged loss). Figure 3b shows that held-out continuation loss is completely orthogonal to generative allocation (R^2=0.000), demonstrating that passive loss convergence cannot predict autonomous search exploration. Removing exact prompt

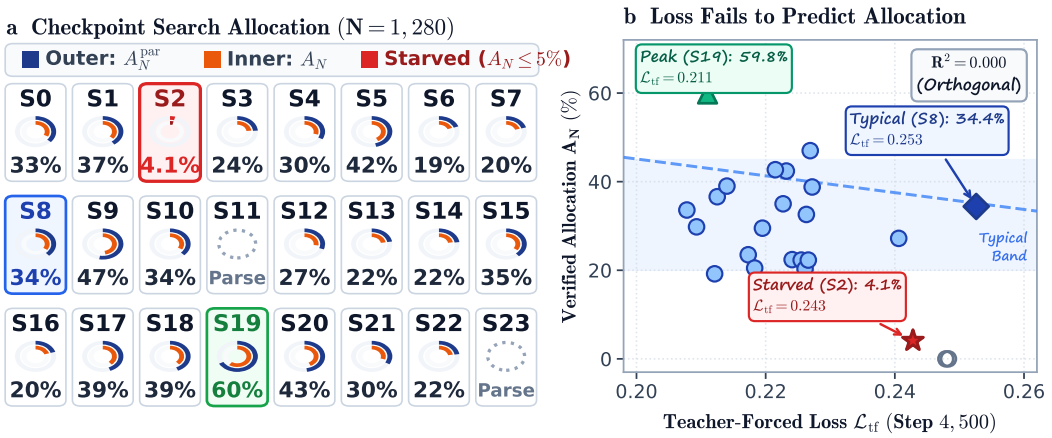


Figure 3: **Generative search allocation across 24 training checkpoints** (N=1,280)**. (a)** Target allocation A_N and parse-conditional A_N^{par} span 4.1%–59.8%, uncovering tail proposal starvation under converged pretraining. **(b)** Offline teacher-forced loss is orthogonal to generative allocation (R^2=0.000).

regurgitations from the count leaves 47, 48, and 28 keys on EH seeds 0–2 (the prompt equation itself is reproduced 0, 2, and 0 times in generated text); in a separate ablation where the commutativity example is removed from the prompt entirely, recovered keys remain 46, 48, and 14 ($A_N = 32.6\%, 35.5\%, 1.6\%$, Appendix A.6). Whole-equation normalization preserves the reported T2 counts (Appendix A.2).

Four generation seeds on the original trio leave the visit split unchanged ($0/12$ versus $12/12$). L_{tf} separates E from EH ($\approx$ 2.3–2.6 versus $\approx$ 0.21–0.24) and does not separate EH search. On seeds 0–2, $L_P(r^\star)$ is 3.1–3.7 on E and 0.47–0.57 on EH, including the low-mass seed. Across twelve EH seeds, verified allocation A_N exhibits mean 28.7% (checkpoint bootstrap 95% CI $[22.2\%, 34.9\%]$), median 31.3%, and Tukey IQR 15.5% (range 4.1–47.0%, Table 2), with probe loss L_{tf} explaining virtually no allocation variance ($R^2 = 0.010, p = 0.75$). All twelve visit. In 1 of 12 seeds (8.3%, 95% Clopper-Pearson CI $[0.2\%, 38.5\%]$), severe proposal starvation occurs at $A_N = 4.1\%$ (3.8–7.5% across generation seeds; 5.6% under joint keys; $A_N^{\mathrm{par}} = 4.6\%$). The other eleven seeds span 19.0–47.0% (mean 31.0%, a 2.5$\times$ spread). Without Seed 2, teacher-forced loss remains uninformative in a leave-one-out regression ($R^2 = 0.005, p = 0.85$). We treat Seed 2 as an existence counterexample demonstrating a degenerate proposal-support failure regime boundary under standard training; its population frequency remains unidentified without larger batches. A

Arm / seed	Saved N	U_N	A_N	A_N^{par}	Other	Parse fail
E / 0*	704	0	0.0%	0.0%	0	98
E / 1	1280	0	0.0%	0.0%	0	625
E / 2	1280	0	0.0%	0.0%	0	749
E / 3	1280	0	0.0%	0.0%	0	735
E / 4	1280	0	0.0%	0.0%	0	407
E / 5	1280	0	0.0%	0.0%	0	685
EH / 0	1280	47	33.1%	45.2%	0	341
EH / 1	1280	49	37.0%	41.9%	0	149
EH / 2	1280	28	4.1%	4.6%	0	123
EH / 3	1280	47	23.6%	49.2%	0	666
EH / 4	1280	44	29.5%	32.2%	0	110
EH / 5	1280	49	42.4%	62.1%	0	405
EH / 6	1280	44	19.0%	19.5%	0	35
EH / 7	1280	43	21.2%	46.6%	0	698
EH / 8	1280	47	34.4%	37.0%	0	90
EH / 9	1280	49	47.0%	48.2%	0	33
EH / 10	1280	46	34.1%	46.5%	0	341
EH / 11	1280	46	19.2%	61.7%	0	881

Table 2: T2 under a shared prompt containing $x + y = y + x$ (D=25M). U_N counts pretrained $x{+}y$ catalog keys; A_N is the share of all raw continuations; A_N^{par} isolates parseable outputs. Seed 2 exhibits structural proposal starvation despite low parse failure ($123/1280$), whereas Seed 11's low all-N mass stems from syntax failure ($881/1280$). *E seed 0 raw covers rounds 9–19.

frozen-checkpoint decoder ablation (Appendix A.5, Table 7) shows that setting repetition penalty to 1.0 (A_N=$0.4\%, 3/50$ keys vs. $36.6\%, 45/50$ on Seed 1), healing the prompt prime to `FORALL X` (A_N=$7.2\%, 92.5\%$ unary collapse), or removing feedback (A_N=2.0%) leaves Seed 2 persistently starved while Seed 1 stays above 30%, showing that decoding hyperparameter, prompt prefix, and feedback adjustments do not close the gap (Appendix A.5, Figure 4).

Raising the budget tenfold to N=12,800 on EH reaches near-complete catalog recall without shifting search mass (Table 3): Seeds 0 and 1 recover $50/50$ keys ($A_{12800} = 35.0\%, 43.4\%$) and Seed 2 recovers $49/50$ keys while remaining starved at $A_{12800} = 5.7\%$ (6.4% parse-conditional). Under a uniform null model (coupon-collector draws of $m \approx 730$ target-family samples across $|C| = 50$ keys corresponding to $A_{12800} = 5.7\%$), expected catalog coverage is 49.99998 (100% Monte Carlo coverage in $[49, 50]$). Seed 2's recovery of $49/50$ keys is predicted by this null model: in finite domains, large budgets overwhelmingly drive near-complete unique recall via random collision, masking exploration starvation. Low A_N imposes substantial discovery overhead: on an extended budget of N=12,800, reaching 28 target keys required 234 continuations for Seed 0 vs. 1,068 for Seed 2 (184 vs. 1,009 verifier calls, a $4.56\times$ continuation and $5.48\times$ verifier overhead; Table 3); on the standard N=1,280 trajectory, Seed 2 requires 1,180 continuations vs. 170 for Seed 8, a $6.9\times$ overhead in that run (Appendix A.6, Table 9).

Model / Reference	L_{tf}	L_P	U_{1280}	A_{1280}	A_{1280}^{par}	$U_{12,800}$	$A_{12,800}$
Seed 0	0.23	0.47	47	33.1%	45.2%	50	35.0%
Seed 1	0.21	0.52	49	37.0%	41.9%	50	43.4%
Seed 2 (Starved)	0.24	0.57	28	4.1%	4.6%	49	5.7%
Null (coupon, $m{\approx}730$)	–	–	–	–	–	$50^{\dagger}$	5.7%

Table 3: Seed 2 is not a failed encoder: L_{tf} and L_P match high-mass seeds. Tenfold scaling (N=12,800) saturates catalog keys ($49/50$, matching the coupon-collector null expectation of $^{\dagger}49.99998$ keys from $m{\approx}730$ target-family draws) while verified mass remains starved at 5.7% (6.4% parse-conditional) vs. 35.0–43.4%. Validity exceeds 99.7%. Trajectories in Appendix A.6.

A seed-0 run trained only on the hard family recovers $48/50$ structural keys and no unary region. Its templates also differ, so it is a sensitivity check on catalog composition, not a second matched deprivation arm (Appendix A.3).

A matched-R control crosses the original budgets (Table 4). E at B=150, D=25M (R=167k, the original EH exposure) remains $\mathrm{Vis}_N(r^\star) = 0$ on three seeds. EH at B=150, D=37.5M (R=250k, the original E exposure) visits on three seeds with $A_N = 35.0/38.4/34.0\%$. Necessity is not an artifact of giving E more per-item tokens. These are new initializations, not a replay of original EH seed 2 at higher R; the 4–8% split remains one training seed at the original D=25M budget.

Arm	R=D/B	U_N	A_N	Vis_N
E B=150	167k	0/0/0	0/0/0	0/3
EH B=150	250k	49/47/46	35.0/38.4/34.0%	3/3

Table 4: Matched per-item exposure, three new training seeds, N=1280. A_N uses all continuations. E at EH's low R still never visits; EH at E's high R does.

A joint-key counterfactual reruns generate–verify on the same twelve frozen checkpoints with whole-equation keys online. It is not a replay of the published histories: later-round prompts change, and the historical sampling RNG is not recoverable. E remains $\mathrm{Vis}_N(r^\star) = 0$ on six seeds. EH remains $\mathrm{Vis}_N = 1$ on six seeds, with $A_N = 32.0/43.0/5.6/27.3/45.2/41.9\%$ against the historical $33.1/37.0/4.1/23.6/29.5/42.4\%$ (Table 2). Seed 2 stays the low-mass outlier (5.6% vs 4.1%). The visit split and the within-EH starvation are not an artifact of the legacy online key.

4.2 Unique coverage can expand within a single function family

Identity catalogs vary B inside one function. U_N rises; A_N never leaves that region. Raising B from 100 to 1200 raises mean unique verified equations from 73 to 667 while $U_{\mathrm{sem}} = 1$ (Table 5). String yield is not structural family coverage.

Identity catalog	U_N	Unique explicit	Unique false	Valid|parsed	Dup|true
$B = 100$, Rlo	73	73	135	61%	80%
$B = 100$, Rhi	84	84	0	100%	88%
$B = 1200$, Rlo	667	666	0	100%	47%
$B = 1200$, Rhi	502	502	0	100%	30%

Table 5: Identity-catalog outputs at $N = 1280$, means of three training seeds. Unique explicit is a count, not a recall percentage ($U_{\mathrm{sem}} = 1$, no support-≥ 2 output). Whole-equation normalization preserves all displayed totals.

At narrow B, greater exposure raises validity among parsed samples from 61% to 100% and removes the observed false proposals, without reducing U_N. At wide B, high-exposure U_N varies (540, 731, 234) because two runs leave the equation format. That is history dependence of P, not a reduction of coverage by exposure (Section 4.4).

4.3 Visit follows the catalog; extra samples do not open the held-out family

Nested $S \times R$ varies which unary regions lie in C. Observed runs set $\mathrm{Vis}_N = 1$ on every trained region and $A_N(r^\star) = 0$ for untrained structure (Table 6). Extra samples raise U_N inside those regions and do not open $r^\star$.

Higher exposure at $S = 6$ raises U_N ($163 \to 256$ at $N = 1280$; $280 \to 300$ at $N = 12{,}800$) without adding regions; scanning parseable `FORALL` statements confirms multi-variable strings still reduce to unary functions (e.g., $x = -y + y + x$). Lower L_{tf} on structural statements therefore need not imply $\mathrm{Vis}_N(r^\star) = 1$ (Appendices A.9, A.13, A.15).

4.4 Feedback modulates formatting without manufacturing search mass

Feedback alters formatting stability without creating search mass: removing feedback cuts parse failures from 998/1280 to 26/1280 (lifting valid proposals from 22.0% to 98.0%, yielding 508 unique verified equations, 677 with a 3-fact window), yet allocation to the withheld family remains strictly

Configuration	$N = 1280$			$N = 12{,}800$		
	U_N	U_{sem}	$A_N(r^\star)$	U_N	U_{sem}	$A_N(r^\star)$
$S = 2$, Rlo	86	2	0	101	2	0
$S = 2$, Rhi	88	2	0	100	2	0
$S = 6$, Rlo	163	6	0	280	6	0
$S = 6$, Rhi	256	6	0	300	6	0

Table 6: Means of training seeds 1–2 across budgets. Zero proposals visit unencoded regions ($A_N(r^\star) = 0$, other maps $= 0$). U_N can exceed catalog size B via within-region interpolation.

zero ($A_N(r^\star) = 0$) across all conditions. Runtime verification repairs syntax, but cannot manufacture search mass on unencoded families.

5 Discussion

What the protocol establishes. Necessity is identified at matched D and survives a matched-R cross: I=0 yields $A_N(r^\star) = 0$ on six training seeds at R=250k, with $x{+}y{=}y{+}x$ in P, a format few-shot, and $0/12$ extra generation runs; three further E seeds at R=167k also remain 0.

Sufficiency fails inside a matched converged loss band ($\mathcal{L}_{\text{tf}} \approx 0.23$). Across 24 EH checkpoints, allocation spans 4.1%–59.8% ($A_N^{\text{par}} \in [4.6\%, 72.4\%]$), where higher offline loss frequently pairs with higher exploration (R^2=0.000, Figure 3). This search gap remains robust across decoding, prompt, and temperature interventions (Appendix A.5, Figure 4).

U_N is decoupled from A_N: at N=12,800, coupon collisions saturate catalog recall (49–$50/50$) even for severely starved checkpoints ($A_{12800} = 5.7\%$, Table 3). Thus, evaluation protocols must report A_N and A_N^{par} directly alongside loss and yield.

Mechanistic nature of proposal starvation. Micro-mechanistic tracing reveals that proposal starvation in Seed 2 is an autoregressive representation preference for unary identity attractors ($x + 0 = x$, $k \cdot x = x$) within valid syntax, rather than grammar failure or unlearning. Under sampling, probability mass concentrates on low-order attractor cycles despite nominal loss on combinatorial additions. In finite-budget regimes, this preference imposes substantial discovery overhead ($> 1{,}000$ calls to reach 28 keys, Table 3), which cannot be detected by pre-search loss or syntax validity (Appendix A.6, Table 9). Consequently, offline continuation likelihood on ground statements fails to bound autoregressive trajectory collapse during stochastic generation.

Scope and boundaries. Our laboratory prioritizes causal isolation over scale: **(1) Protocol specificity**: our claims characterize this controlled base sampling prior, not an open-domain law for guided tree search; **(2) Saturation**: U_N saturation demonstrates why unique recall masks exploration starvation under finite support; **(3) Tail starvation**: severe starvation ($A_N \leq 5\%$) demonstrates an intrinsic failure boundary under converged training; and **(4) Auditing discovery agents**: autonomous LLM discovery pipelines relying on generative proposals to seed tree search should audit verified allocation (A_N) directly to safeguard against unobserved search collapse.

6 Conclusion

Catalog exposure is required for verified search in this laboratory, and offline teacher-forced loss fails to predict generative allocation. Identically trained checkpoints converge to nominal mastery ($\mathcal{L}_{\text{tf}} \approx 0.23$), yet exhibit wide search divergence ($A_N \in [4.1\%, 59.8\%]$) that remains invariant across decoding and prompt interventions (Appendix A.5, Figure 4). Tenfold budget expansion recovers catalog keys while allocation remains starved, proving that offline loss and saturated recall systematically conceal latent search collapse. Benchmarking autonomous LLM discovery requires auditing verified allocation (A_N, A_N^{par}) directly alongside capacity and yield.

AI USE STATEMENT

Generative AI tools and large language models were used in two distinct roles: (1) as research and writing assistants for related work retrieval, prose drafting, code refactoring, experiment execution scripting, and LaTeX formatting; and (2) as experimental subjects under study, whose generative outputs (such as open-domain biomedical candidate hypotheses in Appendix A.11) were evaluated. Generative AI tools were not used for ground-truth verification or metric scoring, nor to generate the formal $\mathbb{Z}_7$ training datasets; all algebraic datasets and validity checks were generated and computed deterministically by offline Python/C symbolic engines and cached databases. The authors reviewed and verified all text, figures, and code, and take full responsibility for the content and claims of this work.

ETHICS STATEMENT

No human subjects. The ophthalmology probe (Appendix A.11) is in-silico literature co-occurrence and makes no diagnostic claims.

REPRODUCIBILITY STATEMENT

Appendix A.16 specifies the configurations, frozen catalogs, raw-output locations, and CPU-only rescoring commands used for the reported results. It distinguishes the historical generation protocol from the corrected scoring and reproducibility implementation.

REFERENCES

Rachit Bansal, Clara Mohri, Tian Qin, David Alvarez-Melis, and Sham Kakade. RL excursions during pre-training: Re-examining policy optimization for LLM training. *arXiv preprint arXiv:2606.04272*, 2026.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Jessica Chudnovsky, Joshua Kazdan, Noam Levi, Rylan Schaeffer, Yegor Denisov-Blanch, Bo He, Mehmet Donmez, Sanmi Koyejo, and David Donoho. Internal data repetition destroys language models. *arXiv preprint arXiv:2606.24998*, 2026.

Rui Dai and Shuran Zheng. Explaining data mixing scaling laws. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=joReaAnwnH.

Darshil Doshi, Aritra Das, Tianyu He, and Andrey Gromov. To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets. In *International Conference on Learning Representations*, 2024.

You-Le Fang, Dong-Shan Jian, Xiang Li, and Yan-Qing Ma. AI-Newton: A concept-driven physical law discovery system without prior physical knowledge. *arXiv preprint arXiv:2504.01538*, 2025.

Yujuan Fu, Ozlem Uzuner, Meliha Yetisgen, and Fei Xia. Does data contamination detection work (well) for LLMs? a survey and evaluation on detection assumptions. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025. URL https://aclanthology.org/2025.findings-naacl.291/.

Ce Ge, Zhijian Ma, Daoyuan Chen, Yaliang Li, and Bolin Ding. Bimix: A bivariate data mixing law for language model pretraining. *arXiv preprint arXiv:2405.14908*, 2024.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Ken Gu, Advait Bhat, Mike A Merrill, Robert West, Xin Liu, Daniel McDuff, and Tim Althoff. Synthworlds: Controlled parallel worlds for disentangling reasoning and knowledge in language models. In *International Conference on Learning Representations*, 2026.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022.

Jake M. Hofman, Angelos Chatzimparmpas, Amit Sharma, Duncan J. Watts, and Jessica Hullman. Pre-registration for predictive modeling. *arXiv preprint arXiv:2311.18807*, 2023.

Thomas Hubert, Rishi Mehta, Laurent Sartran, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651:607–613, 2025. doi: 10.1038/s41586-025-09833-y.

Peter Jansen, Peter Clark, Doug Downey, and Daniel S. Weld. Generating literature-driven scientific theories at scale. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in LLM pre-training: A systematic study of scaling laws, benefits, and pitfalls. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.544/.

John Kirchenbauer, Janny Mongkolsupawan, Yuxin Wen, Tom Goldstein, and Daphne Ippolito. FictionalQA: A dataset for studying memorization and knowledge acquisition. In *International Conference on Learning Representations*, 2026.

Hongyi Liu, Frederic Sala, Thomas Reps, and Adithya Murali. Counterexample guided learning in the large using reasoning agents. *arXiv preprint arXiv:2606.11521*, 2026.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5BjQOUXq7i.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Kazuki Ota, Takayuki Osa, and Tatsuya Harada. Self-supervised theorem discovery in a formal axiomatic system. In *AI for Math Workshop at the International Conference on Machine Learning*, 2026.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Dmitry Redko, Albert Fazlyev, Konstantin Sozykin, Maria Ivanova, Evgeny Burnaev, and Egor Shvetsov. Prior knowledge or search? a study of LLM agents in hardware-aware code optimization. *arXiv preprint arXiv:2605.19782*, 2026.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2023. doi: 10.1038/s41586-023-06924-6.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

Rylan Schaeffer, Joshua Kazdan, Baber Abbasi, Ken Ziyu Liu, Brando Miranda, Ahmed Ahmed, Fazl Berez, Abhay Puri, Stella Biderman, Niloofar Mireshghallah, and Sanmi Koyejo. Quantifying the effect of test set contamination on generative evaluations. *arXiv preprint arXiv:2601.04301*, 2026.

João Sedoc, Baotong Zhang, and Dean Foster. Trust but verify: Prover-verifier deliberation for selective LLM prediction. *arXiv preprint arXiv:2605.25133*, 2026.

Anastasiia Sedova, Skyler Seto, Natalie Schluter, and Pierre Ablin. Scaling laws for mixture pretraining under data constraints. *arXiv preprint arXiv:2605.12715*, 2026.

Aaditya K. Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, Gergely Szilvasy, and Dieuwke Hupkes. Evaluation data contamination in LLMs: how do we measure it and (when) does it matter? *arXiv preprint arXiv:2411.03923*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Yixuan Tang and Yi Yang. AI research agents narrow scientific exploration. *arXiv preprint arXiv:2605.27905*, 2026.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jjCB27TMK3.

Jian Zhao, Haoren Luo, Yu Wang, Yuhan Cao, Pingyue Sheng, and Tianxing He. Can large language models reinvent foundational algorithms? In *International Conference on Machine Learning*, 2026.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. In *International Conference on Machine Learning*, 2024.

Zexin Zhuang, Yanhang Li, and Zhichao Fan. Pre-registering the detectable effect: A paired-mde budget for 4-bit quantization benchmarks, with a pilot audit. *arXiv preprint arXiv:2605.28873*, 2026.

A Protocols and additional measurements

A.1 DSL, exact verification, and function regions

The textual grammar is

$$t ::= X \mid Y \mid Z \mid \text{ZERO} \mid \text{NEG}\, t \mid K_n\, t \mid (t\ \text{PLUS}\ t), \quad n \in \{1, \ldots, m-1\}.$$

Unary operators are prefix tokens without additional parentheses. Equations have the form `FORALL <vars> ( t1 EQUALS t2 )`. The parser accepts a conjecture prefix inside a continuation. The

verifier universally quantifies all variables actually occurring in the equation, including any omitted from the printed binder. Hence malformed binder lists are not evidence for correct formal syntax, although truth of the interpreted equation is checked exactly.

`all_terms(max_size)` bounds the *number of operator nodes per term*; it does not bound only the longest path. The historical pilot uses at most two operators and the main catalogs at most three. The size bound is a catalog-construction bound; parsing and truth checks can accept larger generated terms, which are scored by function region and catalog membership separately. All functions in this grammar are linear over $\mathbb{Z}_m$. For a true equation, either term has the same evaluation table over X, Y, Z. Named-variable tables distinguish kX from kY; equation normalization can identify variable-renamed equations. Catalog region choices therefore use kX, rather than counting renamed copies as additional knowledge conditions. Function-region coverage is a deliberately coarse description of proposals, not logical equivalence of all theorems within a region.

A.2 Equation-key correction and frozen-data compatibility

Two keys must be distinguished. The *legacy* key independently renames each side of an equation in order of first variable appearance, then sorts the two sides. It was used to build the frozen catalogs and to update all reported online histories. Independent renaming can lose cross-side correspondence: it merges $X + Y = Y + X$ with $X + Y = X + Y$, and even $X = Y$ with $X = X$. This does not change the exact truth test, which runs before deduplication, but it is not a correct alpha-equivalence relation.

The corrected *joint* key applies the same variable permutation to both sides. With three variables it is defined by

$$J(t_1 = t_2) = \min_{\pi \in S_3} \operatorname{sort}\big(\pi(t_1), \pi(t_2)\big).$$

It preserves cross-side variable identity and is invariant to side exchange. This is syntactic normalization, not algebraic reduction. The code exposes `legacy_eq_key`, `joint_eq_key`, and a compatibility entry point whose default remains legacy. This prevents an update from silently changing a frozen training catalog. A future joint-key discovery run must explicitly select `--key-mode joint`; it re-keys the actual catalog statements, not a newly selected catalog.

CPU-only rescoring of the 48 historical A/B/S/T2/H and feedback raw dumps leaves every unique-true total unchanged. The six later T2 training-seed dumps are scored separately under the same joint-key CPU path; their allocation and visit results are reported in `research/t2_sixseed_results_20260918.md`. In particular, the original corrected T2 structural counts are $47/49/28$, all in the actual 50-item catalog. The result is robustness of saved outputs, not a counterfactual replay: historical deduplication influenced which facts entered the next prompt. The input hashes, legacy/joint counts, raw round coverage, and per-round measurements for the historical set are saved by `scripts/analysis/submission_audit.py`; the six-seed extension is listed in the dated research record above.

The old constructor reported 5,251 legacy keys for the size-2 $\mathbb{Z}_7$ theory. At size 3 it reported 1,315,962 eligible support-≤ 1 pairs and 157,036 support-≥ 2 pairs, occupying 19 and 144 named-variable function tables. These are *legacy constructor indices*, not corrected counts of alpha-equivalence classes and not counts of independent mathematical discoveries. Main inference concerns the frozen selected catalogs and directly verified outputs; it does not depend on treating these universe sizes as corrected cardinalities.

A.3 Training corpora and exposure

The actual catalog is a frozen list of equation strings and parsed trees. The Identity $B \times R$ conditions (referred to in earlier logs as Experiment A and B configurations) take the first B eligible equations in the identity-function class. The nested S conditions take 50 per chosen unary function. T2 E has $x, 2x$; EH adds $x + y$; H has only $x + y$. "Eligible" in the constructor excludes a bare variable or ZERO as either whole side, which is different from the online ground filter.

All main corpora omit False and worked-false documents. This controlled exposure regime connects to studies of pretraining memorization and data repetition (Carlini et al., 2023; Toneva et al., 2019; Chudnovsky et al., 2026), synthetic data fidelity (Kang et al., 2025), and benchmark contamination

detection (Fu et al., 2025; Singh et al., 2024; Schaeffer et al., 2026). True-only document probabilities are approximately 5% definitions, 53.33% theorem/proof sketches, 33.33% Q&A, and 8.34% catalogs. A catalog document samples up to 20 equations. E and EH share the unary definition and proof templates, including elementary statements such as $x + 0 = x$ and $K_2x = x + x$. These definitions are background knowledge in addition to catalog exposure. EH's two-variable class occurs in theorem, QA and catalog documents. H replaces the definition/proof templates with two-variable text, so the hard-only check is not a matched-template comparison. The templates and selected class lists are part of the corpus specification; a string scan is not a claim to exclude every logical consequence of the background definitions.

The mixed stream contains D domain tokens and a general-text remainder. Training draws random windows from that stream with replacement. Thus D is allocated stream mass and $R = D/B$ is nominal exposure including document templates, not an exact realized number of theorem encounters. The number of optimizer steps is derived from the 300M-token target; rounding to a whole step adds at most one step of tokens. Three training seeds vary initialization and training/mixing randomness, not the catalog.

A.4 Common prompt, parsing, and historical sampling

The prompt names $\mathbb{Z}_7$, its elements and operators, and always includes this seed fact:

```
Fact 1.  The following identity holds in the group Z7:
FORALL X Y ( ( X PLUS Y ) EQUALS ( Y PLUS X ) )
```

It then supplies recent accepted facts and a tail ending in `FORALL`. This seed is a member of the T2 structural class and its training catalog. It is present for E as well as EH and H. The pretraining-held-out interpretation must therefore be distinguished from absence in the full inference input.

Each round samples 32 continuations at each of temperatures 0.6 and 0.8, with top-$p = 0.95$, repetition penalty 1.1, and maximum generation length 96 tokens. The parser prepends the primed token sequence and selects the first parseable non-ground equation. The default history window is $\max(3, \lfloor (L-150)/60 \rfloor)$ facts: six for context length 512. Additional prompt fitting drops the oldest history facts. Within a round all samples share the same prompt. Ground equations are excluded; literal reflexivity and simple rewrites are not generally excluded and are not claimed to be non-trivial discoveries.

```
Input: fixed generator M, verifier V, rounds, samples/round
accepted <- {}; history <- []; seed_fact <- commutativity
for round:
    prompt <- stub + seed_fact + recent(history) + tail
    candidates <- sample M at temperatures 0.6 and 0.8
    for candidate:
        equation <- first non-ground parse, else first parse
        if no parse: count parse failure; continue
        if not V(equation): count false; continue
        if ground(equation): count ground; continue
        key <- legacy_key(equation)
        if key in accepted: count duplicate; continue
        accepted[key] <- equation
        if feedback: append equation to history
Output: raw continuations, round counters, accepted equations
```

Listing 1: Historical discovery loop (schematic)

The old `--seed` argument was recorded but did not initialize the discovery RNG. These are one stochastic generation run per trained checkpoint and budget, not common-random-number paired trials. Training RNG initialization was implemented separately. New generation code uses a private seeded PyTorch generator and saves its state with history and raw observations in atomic checkpoints. Legacy checkpoints lacking RNG state require explicit non-exact-resume permission. These implementation repairs do not retroactively make the reported trajectories deterministic.

A.5 Decoder prefix, repetition penalty, and frozen-checkpoint invariance audit

The generation tail is the string `FORALL` followed by a space. Under the SentencePiece tokenizer that space is a standalone whitespace token, so the prompt is not a token prefix of a training equation beginning `FORALL X`. The reported repetition penalty 1.1 also applies to prompt tokens and can reduce the probability of coefficients named in the request. To quantitatively isolate this effect, we audited token probabilities at `FORALL X ( K<digit>` under repetition penalties 1.0 versus 1.1. Under penalty 1.0, raw logits assign uniform probabilities across candidate digits ($\approx$ 0.18–0.21 for $k \in \{1, \ldots, 5\}$). Applying penalty 1.1 selectively downweights prompt-present digits ($k \in \{1, 2, 6\}$) from $\approx$ 0.19 to 0.036, while unpenalized digits ($k \in \{3, 4, 5\}$) elevate to 0.30–0.33. Crucially, this context reweighting operates identically across all initializations; it does not differentially induce structural starvation on Seed 2.

	EH Seed 1 (High Mass)			EH Seed 2 (Starved)		
Decoder Condition	A_N (raw)	A_N^{par}	U_N	A_N (raw)	A_N^{par}	U_N
Baseline (penalty=1.1, space, fb)	37.0%	41.9%	49/50	4.1%	4.6%	28/50
Ablation 1 (penalty=1.0, unpenalized)	36.6%	39.1%	45/50	0.4%	0.4%	3/50
Ablation 2 (healed prefix `FORALL X`)	30.0%	49.7%	49/50	7.2%	7.5%	33/50
Ablation 3 (static context, no-fb)	42.9%	45.9%	48/50	2.0%	2.0%	21/50

Table 7: **Frozen-checkpoint decoder and prompt invariance audit** (N=1,280). Seed 1 consistently sustains high verified allocation ($A_N \geq 30.0\%$), while Seed 2 remains starved across all interventions ($A_N \leq 7.2\%$), confirming that starvation is an intrinsic checkpoint property.

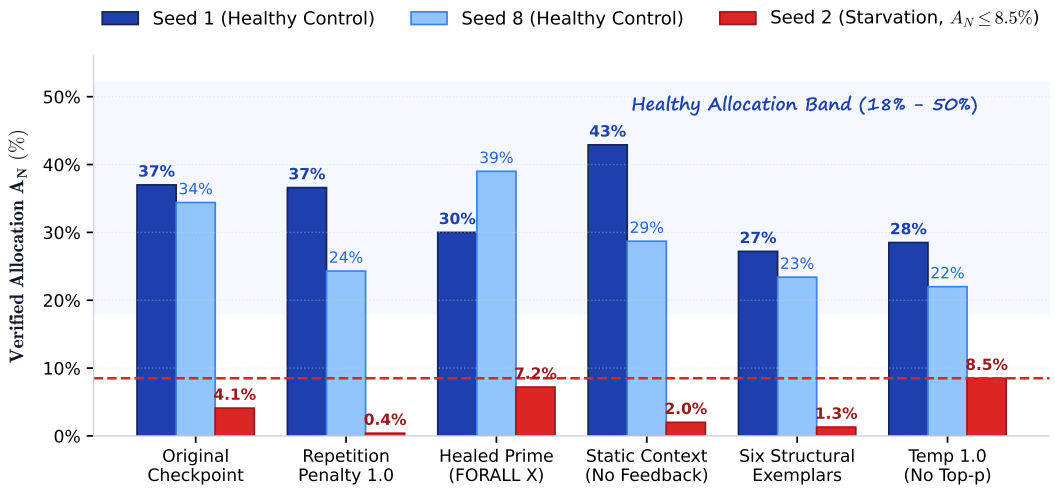


Figure 4: **Robustness of search starvation under decoding and prompt interventions.** Verified allocation A_N remains suppressed ($\leq$ 8.5%) across repetition penalties, prompt healing, context masking, exemplar expansions, and temperature scaling, while control checkpoints maintain healthy exploration (21.9%–42.9%), demonstrating an intrinsic checkpoint property.

To empirically test whether Seed 2's proposal starvation could be an artifact of the decoder configuration, we conducted a frozen-checkpoint decoder ablation across Seed 1 (high-mass sibling) and Seed 2 (starved counterexample) under three independent decoder modifications (N=1,280 each, Table 7): (1) **Repetition penalty ablation (penalty** $= 1.0$**)**: Removing the penalty does not liberate Seed 2; rather, unpenalized decoding drives it into near-complete unary collapse ($A_N = 0.4\%$, $5/1280$ proposals, 99.1% unary identities, $3/50$ keys), while Seed 1 maintains 36.6% allocation ($45/50$ keys). (2) **Token healing ablation (prefix `FORALL X`)**: Healing the prompt prime to avoid whitespace boundary fragmentation keeps Seed 2 predominantly unary (92.5% unary identities, $A_N = 7.2\%$), whereas Seed 1 reaches 49.7% parse-conditional allocation ($49/50$ keys). (3) **Feedback ablation (static context, no feedback)**: Removing in-context feedback eliminates parse fail-

ures on Seed 2 ($0/1280$ failures, 100% parseable), yet its proposal mass remains strictly confined to unary identities (97.4% unary, $A_N = 2.0\%$, $25/1280$ proposals), while Seed 1 achieves 42.9% allocation ($48/50$ keys). Across all conditions, the high-mass vs. starved separation remains: proposal starvation persists across these decoding, prompt, and feedback variations.

The same three reruns on Seed 8, plus the historical run, six structural facts, and temperature 1.0 without nucleus sampling, stay in the high band.

Condition on Seed 8 (N=1280)	A_N	A_N^{par}
Historical decoder	34.4%	37.0%
Penalty 1.0	24.4%	24.6%
Prompt prime `FORALL X`	39.0%	44.7%
No feedback	28.7%	32.0%
Six structural facts, static	23.4%	24.2%
Temperature 1.0, no top-p	21.9%	41.9%

Table 8: Seed 8 under the same frozen reruns as Seeds 1 and 2. Every cell stays above 21%. Keys are 47, 44, 49, 46, 42, and 47 out of 50, in the row order above.

Two further frozen reruns use the same checkpoints and N=1,280. A static prompt of six catalog $x{+}y$ facts, with no online feedback, gives $A_N = 27.3\%/1.3\%/23.4\%$ and $A_N^{\text{par}} = 28.5\%/1.3\%/24.2\%$ on Seeds 1, 2, and 8 ($48/11/42$ keys). Temperature 1.0 with nucleus sampling removed, penalty still 1.1, gives $A_N = 28.4\%/8.5\%/21.9\%$ and $A_N^{\text{par}} = 46.2\%/13.2\%/41.9\%$ ($49/41/47$ keys). Under the historical discovery prompt, mean continuation NLL of one structural string lies in 0.219–0.273 on Seeds 0–17 (Seed 2 is 0.249). The next-token probability of entering the binary branch is 0.287 on Seed 2 and 0.245 on Seed 4, whose A_N^{par} is 32.2%. These reruns are invariance checks, not a screen.

A.6 T2 TRAJECTORY, PROMPT-COPY, AND PARTIAL-WINDOW CHECKS

E seed 0 has 704 saved rows from rounds 9–19 after an interrupted run, rather than a 704-proposal experiment from initialization. EH seed 0 on the same round interval has 47 structural keys and 259 structural proposals; E has zero. EH's first 704 rows have 45 keys, but that prefix is not the corresponding E window. Full E controls exist for seeds 1–2. Online and partial-raw denominators are not pooled.

On EH seeds 0/1/2, the shared prompt equation itself appears in generated output zero/two/zero times. Removing its unique key changes the structural counts from $47/49/28$ to $47/48/28$. Structural proposal counts are $424/474/53$ of 1280. The effect is consistent in direction but its sampling mass is not constant across training seeds.

A prompt ablation reruns the same N=1,280 protocol with the commutativity seed fact omitted. E remains $0/0/0$ structural keys. EH recovers $46/48/14$ trained catalog keys and $417/454/21$ structural proposals; other structural regions stay at 0. Seed copies of $x + y = y + x$ are $0/2/0$. Four additional generation seeds on the historical prompt leave E at $0/12$ visits and EH at $12/12$; EH seed 2 structural mass stays in 3.8–7.5%. A two-binder unary format prompt still yields zero structural keys on E. Continuation NLL of catalog strings under the discovery prompt assigns lower continuation NLL to the structural family on every EH seed (struct$-$unary gap ≈ -0.24), including seed 2; teacher-forced domain probe loss at step 4500 is likewise identical across seeds ($L_{\text{probe}} = 0.221$ on Seeds 0, 1, and 2). Inspection of generated continuations clarifies the nature of Seed 2's proposal starvation: T2 EH Seed 2 maintains low parse failure ($123/1280 \approx 9.6\%$), but among its valid continuations, proposal mass concentrates in unary forms, generating only 53 structural proposals ($A_N^{\text{par}} = 4.6\%$). In contrast, a separate syntactic collapse into natural-language commentary preambles ('1. Observation...', producing $998/1280$ parse failures) occurs only under the high-exposure Identity-catalog checkpoint (R_{hi}, Table 5), whereas under R_{lo} that checkpoint exhibits near-zero parse failure ($4/1280$). Seed 2's structural starvation reflects a proposal preference for unary equations over multi-variable structures within syntactically valid outputs, rather than arising from parametric unlearning or syntax failure. Optimization dynamics inducing this preference remain an open question for future work.

Search cost audit and absence of a pre-search screener. To examine whether early proposal allocation or offline representations can predict downstream search cost before full exploration, we evaluate prefix allocation ($A_{64}^{\text{par}}, A_{128}^{\text{par}}$) and offline loss against the number of continuations required to reach 28 unique target catalog keys across all twelve EH training seeds (Table 9). On this N=1,280 trajectory, A_{128}^{par} exhibits a negative rank correlation with search cost (Spearman $\rho = -0.510, p = 0.090$), while offline parametric loss L_{tf} does not correlate with search cost ($\rho = -0.077, p = 0.812$). In this specific run, Seed 8 reaches 28 keys in 170 continuations whereas Seed 2 requires 1,180 continuations (6.9× overhead), despite Seed 8 having higher loss (0.2526 vs. 0.2428).

Crucially, this relationship reflects within-trajectory autocorrelation driven by a single extreme point, rather than an operative pre-search screener: (1) **Prefixes fail to rank non-starved checkpoints**: Excluding Seed 2, early allocation exhibits no statistically significant correlation with search cost ($\rho = -0.364, p = 0.272$). For example, Seed 7 achieves $A_{128}^{\text{par}} = 57.1\%$ (second highest allocation) yet requires 284 continuations (slower than Seed 8's 170 with 23.2%). Shorter prefixes lack discriminative support: $A_{64}^{\text{par}} = 0$ occurs for both Seed 2 (starved, cost 1,180) and Seed 0 (full-budget $A_{1280}^{\text{par}} = 45.2\%$, cost 350). (2) **Downstream cost ranks reshuffle across independent runs**: Re-evaluating Seeds 0–5 under an independent generation run (the joint-key counterfactual run) confirms that downstream search waiting times swap ranks: Seed 0's cost drops from 350 to 193, falling behind Seed 3 (248), while Seed 5's cost (143) exceeds Seed 1 (119). Across independent runs excluding Seed 2, early prefix allocation from one run yields no rank correlation with search cost in the next ($\rho = -0.100, p = 0.873$). Normal-seed cost variations (115–350 continuations) reflect decoding sampling variance rather than stable parametric rankings. (3) **Extreme starvation is invariant across runs**: Across three independent generation trajectories (Discovery, Joint Key, and N=12,800), Seed 2 consistently allocates only 4.6%–6.4% parse-conditional structural mass and requires $\approx$ 1,000+ continuations (1,180, 1,068, 1,068) to reach 28 keys, whereas Seeds 0 and 1 consistently reach 28 keys within a few hundred continuations (143/119 for Seed 1, 350/193/234 for Seed 0). Proposal starvation is a robust property of the frozen checkpoint. (4) **Offline likelihood cannot separate the starved checkpoint**: Continuation NLL of target catalog strings assigns lower NLL to the structural family over the unary family across every EH seed by ≈ -0.24 (including Seed 2), and domain probe loss at step 4500 is identical ($L_{\text{probe}} = 0.221$ on Seeds 0, 1, and 2). In summary, short prefixes and offline representations cannot pre-screen search costs across checkpoints; what survives is that an identically encoded checkpoint can persistently suffer severe proposal starvation that offline metrics fail to detect.

Checkpoint	$L_{\text{tf}}(r^\star)$	A_{64}^{par}	A_{128}^{par}	A_{1280}^{par}	Search Cost (N to 28 keys)
EH Seed 0	0.2246	0.0%	25.5%	45.2%	350
EH Seed 1	0.2124	36.7%	32.5%	41.9%	143
EH Seed 2 (Starved)	0.2428	0.0%	0.0%	4.6%	1180
EH Seed 3	0.2423	34.1%	42.9%	49.2%	273
EH Seed 4	0.2336	13.6%	18.6%	32.2%	223
EH Seed 5	0.2402	13.3%	46.9%	62.1%	130
EH Seed 6	0.2385	6.3%	6.3%	19.5%	309
EH Seed 7	0.2218	40.0%	57.1%	46.6%	284
EH Seed 8	0.2526	20.0%	23.2%	37.0%	170
EH Seed 9	0.2338	18.8%	41.7%	48.2%	115
EH Seed 10	0.2436	31.9%	45.6%	46.5%	128
EH Seed 11	0.2293	30.0%	69.1%	61.7%	151

Table 9: **Prefix allocation and discovery milestones across twelve EH seeds** (N=1,280). Rank correlations of early prefixes are driven by Seed 2's extreme overhead ($\approx$ 1,000+ continuations to reach 28 keys), which cannot be screened by offline loss $\mathcal{L}_{\text{tf}}$.

A.7 Matched per-item exposure control

The main T2 contrast fixes total domain tokens D, so adding the structural family lowers nominal per-item exposure. To test whether the necessary condition is only an exposure-density artifact, a crossed control matches the two per-item exposure levels with new initializations. E uses $B = 150$ unary items at $D = 25$M ($R = 167$k), matching the original EH arm; EH uses $B = 150$

items at $D = 37.5$M ($R = 250$k), matching the original E arm. At $N = 1280$, E remains at zero target-family visits on all three seeds, while EH visits the family on all three with $A_N = 35.0\%/38.4\%/34.0\%$. These six checkpoints represent an independent matched-R batch trained under the same architecture and tokenizer, not a replay of the main T2 models at higher compute.

Arm	$R{=}D/B$	U_N	A_N	Vis_N
E $B{=}150$	167k	0/0/0	0/0/0	0/3
EH $B{=}150$	250k	49/47/46	35.0/38.4/34.0%	3/3

Table 10: Matched per-item exposure control, three new training seeds, $N{=}1280$. A_N uses all continuations. E at EH's low R still never visits; EH at E's high R reliably visits.

This crossed result strengthens the access claim: neither giving the withheld E arm EH's lower nominal exposure nor giving EH E's higher nominal exposure changes the visit/no-visit direction. It does not identify all effects of changing the catalog's total size and document composition, and it is not a replay of the original EH seed 2 at higher R. The raw files and CPU scores are recorded in `research/t2_equal_r_results_20260918.md`; the historical online histories remain frozen. A joint-key online counterfactual on the same twelve frozen checkpoints is recorded in `research/t2_joint_key_counterfactual_results_20260919.md`.

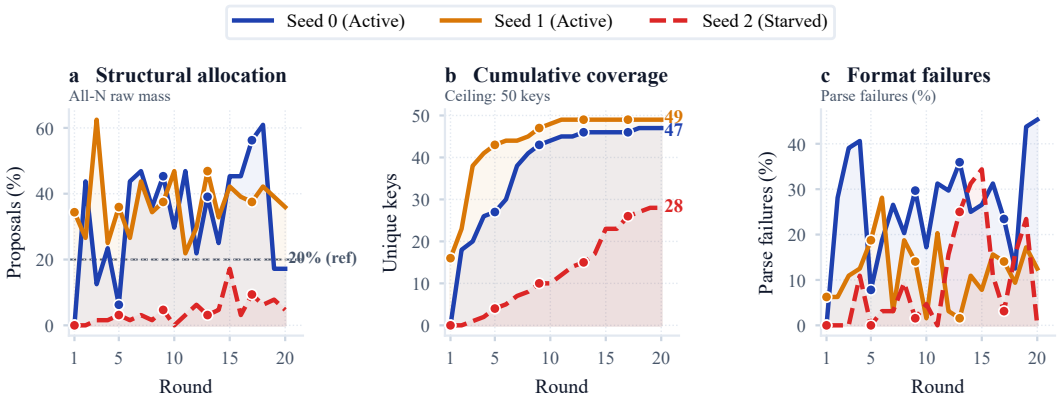


Figure 5: **Rollout trajectories across training seeds.** Cumulative catalog coverage and proposal allocation over $N{=}1{,}280$ continuations. Seed 2 exhibits persistent generative starvation ($A_N = 4.1\%$), contrasting sharply with healthy control seeds (33.1%–37.0%).

A.8 Raw coverage and full per-run counts

All A dumps contain 1280 observations. The short S matrix has three training seeds; S=2 high-exposure seed 0 has only rounds 9–19 saved. The large-budget S matrix has seeds 1–2 ($N = 12{,}800$). While expanding test-time compute budgets and sampling volume can improve reasoning accuracy in verifiable domains (Snell et al., 2025), in our generate–verify laboratory it primarily drives up unique string yield U_N without mitigating semantic allocation starvation, cautioning against treating token-level or surface metric gains as evidence of true search exploration (Schaeffer et al., 2023). The fixed-$B = 600$ experiment has eight available raw dumps; low-exposure seed 2 is missing. Two completed online runs have partial raw files: mid-exposure seed 0 has 8960 rows and high-exposure seed 0 has 7104. Rescoring statements refer to the available rows, not reconstructed missing observations. Tables below are generated directly by the CPU audit.

A.9 Teacher-forced probes

Probes evaluate mean next-token loss every 250 training steps. The original mixer pilot used 198 items (explicit, held-out, false). The identity matrix uses nested core/catalog bands and a structural holdout probe. Table 22 reports the latter at step 4500. Lower loss is not a correctness discrimination test or evidence that the model can freely produce the probe family. On the T2 checkpoints at step

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_b100_rhi_s0	1280	0–19	77	77	0	725
a50_b100_rhi_s1	1280	0–19	91	91	0	457
a50_b100_rhi_s2	1280	0–19	84	84	0	513
a50_b100_rlo_s0	1280	0–19	77	77	0	380
a50_b100_rlo_s1	1280	0–19	69	69	0	883
a50_b100_rlo_s2	1280	0–19	73	73	0	499
a50_b1200_rhi_s0	1280	0–19	540	540	0	496
a50_b1200_rhi_s1	1280	0–19	731	731	0	17
a50_b1200_rhi_s2	1280	0–19	234	234	0	998
a50_b1200_rlo_s0	1280	0–19	678	678	0	25
a50_b1200_rlo_s1	1280	0–19	656	656	0	17
a50_b1200_rlo_s2	1280	0–19	667	667	0	4

Table 11: Saved-output audit: `results/round2_a/discovery`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_fs_b600_d2_s0	7104	89–199	580	580	0	1125
a50_fs_b600_d2_s1	12800	0–199	585	585	0	15
a50_fs_b600_d2_s2	12800	0–199	595	595	0	18
a50_fs_b600_dhalf_s0	12800	0–199	574	574	0	785
a50_fs_b600_dhalf_s1	12800	0–199	580	580	0	13
a50_fs_b600_dstar_s0	8960	60–199	546	546	0	0
a50_fs_b600_dstar_s1	12800	0–199	541	541	0	0
a50_fs_b600_dstar_s2	12800	0–199	551	551	0	6163

Table 12: Saved-output audit: `results/round2_b/discovery`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_s2_rhi_s0	704	9–19	72	72	0	98
a50_s2_rhi_s1	1280	0–19	91	91	0	625
a50_s2_rhi_s2	1280	0–19	85	85	0	749
a50_s2_rlo_s0	1280	0–19	94	94	0	204
a50_s2_rlo_s1	1280	0–19	76	76	0	943
a50_s2_rlo_s2	1280	0–19	96	96	0	339
a50_s6_rhi_s0	1280	0–19	243	243	0	303
a50_s6_rhi_s1	1280	0–19	257	257	0	346
a50_s6_rhi_s2	1280	0–19	255	255	0	185
a50_s6_rlo_s0	1280	0–19	190	190	0	257
a50_s6_rlo_s1	1280	0–19	171	171	0	448
a50_s6_rlo_s2	1280	0–19	155	155	0	814

Table 13: Saved-output audit: `results/round2_s/discovery`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_s2_rhi_s1	12800	0–199	100	100	0	7164
a50_s2_rhi_s2	12800	0–199	100	100	0	7344
a50_s2_rlo_s1	12800	0–199	97	97	0	10365
a50_s2_rlo_s2	12800	0–199	105	105	0	3340
a50_s6_rhi_s1	12800	0–199	300	300	0	3460
a50_s6_rhi_s2	12800	0–199	300	300	0	1550
a50_s6_rlo_s1	12800	0–199	277	277	0	4649
a50_s6_rlo_s2	12800	0–199	283	283	0	7880

Table 14: Saved-output audit: `results/round2_s/discovery_n12800`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_s2_rhi_s3	1280	0–19	93	93	0	735
a50_s2_rhi_s4	1280	0–19	94	94	0	407
a50_s2_rhi_s5	1280	0–19	92	92	0	685
a50_t2_eh_d25_s0	1280	0–19	136	136	47	341
a50_t2_eh_d25_s1	1280	0–19	143	143	49	149
a50_t2_eh_d25_s10	1280	0–19	130	130	46	341
a50_t2_eh_d25_s11	1280	0–19	105	105	46	881
a50_t2_eh_d25_s2	1280	0–19	122	122	28	123
a50_t2_eh_d25_s3	1280	0–19	128	128	47	666
a50_t2_eh_d25_s4	1280	0–19	137	137	44	110
a50_t2_eh_d25_s5	1280	0–19	122	122	49	405
a50_t2_eh_d25_s6	1280	0–19	138	138	44	35
a50_t2_eh_d25_s7	1280	0–19	120	120	43	698
a50_t2_eh_d25_s8	1280	0–19	136	136	47	90
a50_t2_eh_d25_s9	1280	0–19	143	143	49	33

Table 15: Saved-output audit: `results/round2_t2/discovery`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_t2_eh_d25_s0	12800	0–199	150	150	50	4862
a50_t2_eh_d25_s1	12800	0–199	150	150	50	2205
a50_t2_eh_d25_s2	12800	0–199	148	148	49	1319

Table 16: Saved-output audit: `results/round2_t2/discovery_n12800`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_t2_e_b150_s0	1280	0–19	126	126	0	291
a50_t2_e_b150_s1	1280	0–19	94	94	0	1046
a50_t2_e_b150_s2	1280	0–19	103	103	0	928
a50_t2_eh_r250_s0	1280	0–19	138	138	49	32
a50_t2_eh_r250_s1	1280	0–19	137	137	47	234
a50_t2_eh_r250_s2	1280	0–19	138	138	46	156

Table 17: Saved-output audit: `results/round2_t2/equal_r`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_s2_rhi_s0	1280	0–19	79	79	0	341
a50_s2_rhi_s1	1280	0–19	92	92	0	655
a50_s2_rhi_s2	1280	0–19	88	88	0	737
a50_s2_rhi_s3	1280	0–19	93	93	0	806
a50_s2_rhi_s4	1280	0–19	92	92	0	476
a50_s2_rhi_s5	1280	0–19	92	92	0	685
a50_t2_eh_d25_s0	1280	0–19	129	129	48	359
a50_t2_eh_d25_s1	1280	0–19	140	140	48	105
a50_t2_eh_d25_s2	1280	0–19	125	125	30	101
a50_t2_eh_d25_s3	1280	0–19	127	127	47	636
a50_t2_eh_d25_s4	1280	0–19	135	135	46	94
a50_t2_eh_d25_s5	1280	0–19	117	117	46	444

Table 18: Saved-output audit: `results/round2_t2/joint_key_counterfactual`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_t2_h_d25_s0	1280	0–19	48	48	48	165

Table 19: Saved-output audit: `results/round2_h/discovery`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
clamp3	1280	0–19	677	677	0	133
nofb	1280	0–19	508	508	0	26

Table 20: Saved-output audit: `results/round2_s2_feedback`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

Run	Saved N	Rounds	Legacy U	Joint U	Hard keys	Parse fail
a50_t2_eh_d25_s1_ce	1280	0–19	139	139	49	151

Table 21: Saved-output audit: `results/round2_fb/discovery`. U counts distinct true non-ground equations; hard keys have $\phi = (1, 1, 0, 0)$. Rounds are zero-indexed. Partial raw windows are not complete shorter-budget runs.

4500, mean loss on the 66 `fs_hold` structural statements (collected for original diagnostic seeds 0–2) is $0.225/0.212/0.243$ for EH and $2.95/2.59/2.34$ for E. Free-generation structural proposal rates for EH are $33.1\%/37.0\%/4.1\%$. Seed 2 is therefore a proposal-support failure, not a teacher-forced training failure.

Identity cell	core100	band1200	structural probe
$B = 100$, Rlo	0.26–0.30	0.98–1.04	2.07–2.53
$B = 100$, Rhi	0.19–0.27	1.07–1.13	2.49–2.81
$B = 1200$, Rlo	0.28–0.29	0.23	0.80–1.00
$B = 1200$, Rhi	0.27	0.22	1.07–1.21

Table 22: Mean next-token loss, range across three training seeds.

A.10 Feedback-window and counterexample checks

The identity $B = 1200$, high-exposure seed-2 checkpoint is fixed across the following runs. Shortening or removing feedback substantially changes format failures and coverage, consistent with findings on prover-verifier deliberation and self-refinement (Madaan et al., 2023; Sedoc et al., 2026). These are stochastic single-run checks; unchanged weights do not establish identical hidden competence across arbitrary contexts. For EH seed 1, success-only and counterexample feedback yield 49

History	N	Parse failures	Unique true	Lecture-style strings
Default (six facts)	1280	998	234	965
No feedback	1280	26	508	0
Three facts	1280	133	677	125

Table 23: Fixed-checkpoint feedback sensitivity. Lecture style is the diagnostic template-matching count from the archived analysis. No run generates a support-≥ 2 equation.

structural keys each, zero other structural keys, and respectively 149 and 151 parse failures. Each has two false proposals. The two resulting counterexample notes provide too little activation to test a general claim that feedback expands or cannot expand candidate support.

A.11 Methodological firewall: exploratory biomedical text retrieval

To examine whether candidate set diversity in open-ended natural language generation exhibits similar decoupling, preliminary exploratory probes were conducted using an ungrounded LLM API over retrieved biomedical literature prefixes. However, we establish a strict methodological firewall: open-domain text generation lacks an exhaustive formal verifier, and retrieval context interventions ($K \in \{0, 5, 10\}$) do not correspond to controlled pretraining catalog interventions (C). Without an

exact ground-truth algebraic verifier, automated dictionary linking measures candidate vocabulary shifts rather than verified correctness or novel discovery. For this reason, no causal claims of proposal starvation or verification efficiency are drawn from open-domain text retrieval, and the core findings of this paper rest strictly on the formal, verifier-backed $\mathbb{Z}_7$ protocol.

A.12 Registration stages and historical pilots

The initial mixer-α protocol predicted an interior optimum in discovery efficiency; that prediction was not supported at saturated budgets. Subsequent dated protocols froze the identity matrix, nested S matrix, structural intervention, and ophthalmology follow-ups before their respective new data collections, following established pre-registration principles and effect-size budgeting (Hofman et al., 2023; Zhuang et al., 2026). T2 seed 0 was an instrument check; seeds 1–2 supplied the first replication, and seeds 3–5 were later added under the same frozen T2 protocol. Seed-0 large-budget and H-only runs preceded that replication, a recorded execution-order deviation. The empirical gate "structural output on at least two of three seeds" is a pre-specified decision rule, not a significance test. Late CPU normalization and HGNC checks are labeled as post-hoc robustness. Protocol documents and the submission audit preserve this sequence.

A.13 Mixer and modulus pilots

The initial runs varied domain share α, budget, and 50M/110M scale, with False documents and catalogs constructed under legacy keys. They are distinct from the true-only main experiments. A symmetric rescoring showed that increasing α improved true coverage while the apparent independent duplication axis weakened. We do not infer an exploration/concentration law from those mixed-denominator plots. Historical non-$\mathbb{Z}_7$ containment was mostly single-seed and is not replication of T2. Representative pilot values are preserved below.

Domain	Scale	Accepted low/high	Held-out low	Containment (%)
$\mathbb{Z}_5$	50M	151/199	0	97.4
$\mathbb{Z}_7$	50M	74/136	0	93.9–97.7
$\mathbb{Z}_{11}$	50M	73/89	2	90.0
$\mathbb{Z}_{13}$	50M	44/47	0	80.0
$\mathbb{Z}_{11}$	110M	55/93	3	—

Table 24: Legacy modulus pilot at $\alpha = 0.05/0.5$, 300M tokens. Counts use legacy online keys; $\mathbb{Z}_7$ containment comes from mixed-proposal replicates, while other containment points are single-seed.

A.14 Historical mixed proposals

At both measured $S \times R$ budgets, the best single arm already generates all six observed unary functions. The union of four arms adds equation variants but no function region. That is an offline diagnostic using $4N$ proposals after region coverage has saturated, not an equal-budget mixture result and not a fourth link in the chain.

The older pilot compared a 50/50 mixture with its best single generator at the same total number of proposals. Its outcomes concern legacy equation-key coverage, not semantic-class gain. The larger Z13 budget was registered after the smaller-budget audit. Those results neither establish necessary-and-sufficient mixing conditions nor imply a benefit for the nested S arms used in the main study.

Domain	N	Mixed unique keys	Best-single unique keys
$\mathbb{Z}_7$	1280	130–131	130–134
$\mathbb{Z}_{13}$	1280	49–55	49–53
$\mathbb{Z}_{13}$	2560	57–60	54–55

Table 25: Three-seed ranges for the historical equal-budget mixture experiment.

A.15 HISTORICAL ENTROPY PROFILE

Four 50M pilot checkpoints were profiled over 64 sequences and 96 steps. Entropy is computed before decoding transforms at temperature 1; generation used temperature 0.8. Mean entropies at $\alpha = 0.05/0.5$ were 0.498/0.244 nats at 300M training tokens and 0.251/0.251 at 500M. These are exploratory checkpoint comparisons, not a causal entropy intervention or a validated mechanism for the T2 effect.

A.16 REPRODUCTION MAP

Training configurations are in `configs/round2_a/`, `configs/round2_s/`, and `configs/round2_t2/`. Frozen class lists are under `results/round2_offline/`: `semantic_breadth_classes.json` and `t2_classes.json`. Raw observations are under `results/round2_a/`, `round2_b/`, `round2_s/`, `round2_t2/`, `round2_h/`, and `round2_s2_feedback/`, and `results/round2_fb/discovery/`, with separate `discovery` and `discovery_n12800` folders. The CPU-only `scripts/analysis/submission_audit.py` writes the input hashes, per-round outcomes, and legacy/joint normalization counts; `make_submission_figures.py` produces the main T2 figure and trajectory plot from that audit JSON. Unit tests cover cross-side variable correspondence and sampling/checkpoint continuity.

The ophthalmology inputs are specified by `configs/case_ophthalmology_r3.json`; records are in `results/case_study/hypotheses_r3.jsonl` and the corresponding scored file. The HGNC link script and stored dictionary reproduce the post-hoc entity counts. These commands rescore existing data without training, API calls, or counterfactual reconstruction of missing raw rows. The joint-key online counterfactual trajectories and scores are recorded in `research/t2_joint_key_counterfactual_results_20260919.md`, and the matched-R control results are in `research/t2_equal_r_results_20260918.md`. The accompanying artifact guide distinguishes frozen observations from future runs made with repaired normalization and RNG handling.