COVERWEAVE: COVERAGE PERCOLATION GOVERNS COMPOSITIONAL DEPTH EXTRAPOLATION UNDER VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) is now the standard recipe for training models on checkable reasoning tasks, and the capability it is meant to buy is composition: a policy trained on shallow derivations should solve deeper ones it has never seen. Which property of the training set decides whether this happens is unknown, because existing compositional benchmarks fix combination structure and data volume together and cannot separate them. We introduce COVERWEAVE, a procedurally generated symbolic environment with an exact checker and a hindsight data engine in which every legal rollout ends in a legal goal, so verified training examples are produced by symbolic execution at zero model cost. In this engine, pair coverage, the fraction of atom pairs that co-occur in training, is an exactly controlled variable at fixed trajectory count and depth profile. A percolation account predicts that depth extrapolation is governed by the giant component S of the atom co-occurrence graph, with success $\approx c\,S(\rho)^d$ at depth d. The prediction holds without a free parameter: on a coverage sweep, $S(\rho)^d$ matches measured success at three test depths within 2.3 points on average. Coverage dominates volume: one quarter of the data above threshold reaches 93.5% average solve rate, while four times the data below threshold reaches 21.6%, no better than the base amount. Binder cumulant crossings confirm that the realized co-occurrence graph has its structural threshold at the random-graph value $1/(N-1)$ for vocabulary sizes 12 to 48, with a finite-size exponent $\nu = 3.08$ close to the mean-field value 3. A threshold of the same shape appears with no weight updates when coverage is varied only in the prompt, and its location agrees across six frozen models to within 0.028. Training above threshold transfers to building-block-constrained retrosynthesis, raising restricted-catalogue route success from 6.3% to 21.9%.

1 INTRODUCTION

RLVR has become the default way to train models on tasks whose answers can be checked (Shao et al., 2024; DeepSeek-AI, 2025; Lambert et al., 2024). What it is meant to buy is composition: a model trained on shallow derivations should solve deeper ones built from the same parts. Whether it does is contested. One line of evidence finds that verifiable-reward training mainly sharpens the sampling distribution over solutions the base model could already produce (Chen et al., 2026), another that prolonged training expands what the model can solve (Liu et al., 2026), and aggregate benchmark movement cannot tell sharpening from composition. The underlying question is a data question: *what property of the training set decides whether a policy extrapolates in depth?* It has been hard to answer because every existing compositional benchmark (Lake & Baroni, 2018; Kim & Linzen, 2020; Keysers et al., 2020; Ruis et al., 2020; Li et al., 2023) fixes the combination structure and the data volume together.

We build an environment in which they separate. COVERWEAVE (Figure 1) is a procedurally generated symbolic world with an exact checker, and its data engine exploits a hindsight property: any legal action sequence ends in a state that is itself a legal goal, so the pair (terminal state, causal action subsequence) is a verified training example produced by symbolic execution alone (Andrychowicz et al., 2017). Because generation is free, the engine can afford rejection sampling, which we use to

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.

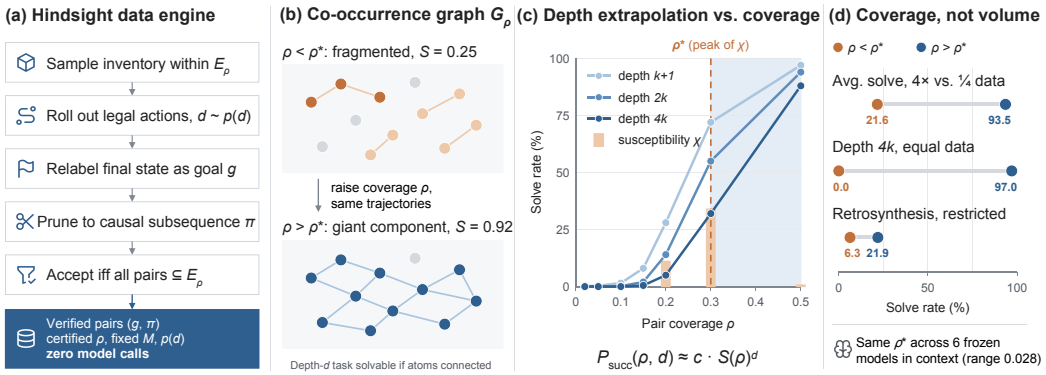

Figure 1: **CoverWeave at a glance.** (a) The hindsight engine samples inventories whose atoms induce pairs only in a target edge set E_ρ, rolls out legal actions, relabels the terminal state as the goal, prunes to the causal subsequence, and accepts only certified trajectories, with zero model calls. (b) Raising coverage at fixed trajectory count turns a fragmented co-occurrence graph into one with a giant component. (c) Depth extrapolation on the $N{=}32$ coverage sweep (Table 2); the dashed line marks the behavioral threshold ρ^*, the peak of the across-seed susceptibility χ. (d) Coverage, not volume: $4\times$ data below threshold versus $\frac{1}{4}$ data above it, equal-data depth-$4k$ success, and transfer to restricted-catalogue retrosynthesis.

make *pair coverage* ρ, the edge density of the graph whose edges are atom pairs that co-occur in training derivations, an exactly controlled variable at fixed trajectory count and fixed depth profile.

With coverage under control, a mechanism suggests itself. If a model learns the role of an atom by contrast with the atoms it is observed alongside, and role information propagates along observed pairs, then a depth-d derivation is resolvable only when its d atoms lie in a common connected component of the co-occurrence graph. This is a percolation account, and it predicts that success behaves as $c\,S(\rho)^d$, where S is the fraction of atoms in the giant component (Erdős & Rényi, 1960; Newman, 2010). Appendix Figure 2 presents a synthetic empirical case study demonstrating how fragmented versus connected support affects compositional rollouts. The account makes predictions that a smooth data-scaling story (Kaplan et al., 2020; Hoffmann et al., 2022) does not, and we test them with the finite-size-scaling machinery that separates a critical point from a crossover (Binder, 1981; Stauffer & Aharony, 1994); the metric-artifact critique of emergence (Schaeffer et al., 2023) makes this machinery necessary rather than optional.

Contributions. (i) A verified, zero-cost data engine in which coverage is an exactly controlled variable (Section 3). (ii) A percolation account with separately falsifiable predictions, distinguishing a *structural* threshold ρ_c of the co-occurrence graph from the *behavioral* threshold ρ^* at which extrapolation becomes measurable (Section 4). (iii) Evidence that coverage, not volume, governs depth extrapolation: $S(\rho)^d$ predicts measured success without fitting, $4\times$ subcritical data does nothing, and finite-size scaling confirms that the realized co-occurrence graph has the random-graph threshold and exponent that the analysis assumes (Section 6). (iv) The threshold appears in context with no weight updates, at a location that varies by only 0.028 across six frozen models, and supercritical pretraining transfers to a domain with no symbolic overlap (Sections 6.4 and 6.5). We report the places where the account needed refinement, above all that held-out atoms are solved far more often than it predicts, as prominently as the ones where it held.

2 Related Work

Compositional generalization. SCAN, COGS, CFQ, gSCAN and SLOG (Lake & Baroni, 2018; Kim & Linzen, 2020; Keysers et al., 2020; Ruis et al., 2020; Li et al., 2023) test whether models interpret novel combinations of seen primitives, Hupkes et al. (2020) decompose compositionality into testable parts, and LLM errors compound with composition depth (Dziri et al., 2023; Press et al., 2023). These benchmarks are static, exposed to contamination (Sainz et al., 2023), and cannot vary combination structure at fixed volume. Synthetic studies show that compositional abilities can emerge

Algorithm 1 Coverage-controlled hindsight generation

Input: instance $\mathcal{E}$, target coverage ρ, depth distribution $p(d)$, budget M
Sample E_ρ uniformly among edge sets of size $\rho\binom{N}{2}$ on the atom pair graph
while fewer than M trajectories accepted **do**
    Sample an inventory whose atoms induce pairs only within E_ρ, preferring pairs not yet realized
    Sample $d \sim p(d)$; roll out a legal action sequence of derivation depth d
    Record terminal state as goal g; prune to causal subsequence π; verify that π solves g
    if every realized pair of π lies in E_ρ **then** accept (g, π) **else** reject
end while
return M trajectories whose realized pair set is certified to equal E_ρ, with depth profile $p(d)$

multiplicatively (Okawa et al., 2023; Ramesh et al., 2024); we identify which training-set property controls the emergence.

Emergence and its critics. Sharp capability changes have been called emergence (Wei et al., 2022) and contested as metric artifacts (Schaeffer et al., 2023); grokking is a delayed transition at fixed data (Power et al., 2022; Wang et al., 2024). Mechanistic accounts include quantized skills (Michaud et al., 2023), random skill-text graphs (Arora & Goyal, 2023; Nayak & Varshney, 2025), and percolation on a knowledge graph (Lubana et al., 2025). We vary a structural property of the data rather than scale, and test it with Binder-cumulant and collapse diagnostics rather than post-hoc fits.

RLVR and self-training. Verifiable-reward training uses PPO (Schulman et al., 2017), GRPO (Shao et al., 2024; Liu et al., 2025), RLOO (Ahmadian et al., 2024) and DAPO (Yu et al., 2026), with STaR and rejection-sampling fine-tuning as self-training baselines (Zelikman et al., 2022; Yuan et al., 2023). On the sharpening debate (Chen et al., 2026; Liu et al., 2026), we find that optimizers differ by about six points at matched coverage while coverage moves extrapolation by more than forty.

Hindsight relabeling and retrosynthesis. Our engine builds on hindsight experience replay (Andrychowicz et al., 2017), except every terminal state is an on-distribution goal. Multi-step retrosynthesis is searched by MCTS (Segler et al., 2018; Genheden et al., 2020), neural A* (Chen et al., 2020), graph search (Xie et al., 2022), value networks (Liu et al., 2023), under uncertainty (Tripp et al., 2024) and bidirectionally (Yu et al., 2024), and evaluated with PaRoutes on USPTO data (Genheden & Bjerrum, 2022; Lowe, 2012; Schneider et al., 2016); its template library, building-block catalogue and restricted catalogue map onto our synthesis table, inventory and scarcity condition.

3 The CoverWeave Environment and Data Engine

Instances. An instance is $\mathcal{E} = (\mathcal{A}, \mathcal{H}, \mathcal{S}, \mathcal{D}, \mathcal{C})$: a set $\mathcal{A}$ of N atoms (basic states), higher-order states $\mathcal{H}$, a synthesis table $\mathcal{S}$ mapping ordered state strings to higher-order states, a decomposition table $\mathcal{D}$ mapping higher-order states to multisets of parts, and positive action costs $\mathcal{C}$. Every component is resampled per instance under a generator exposing vocabulary sizes, rule arity, the multiplicity of alternative rules per product, and a conservation exponent controlling whether decomposition returns more or fewer parts than synthesis consumed. Symbol names are randomized independently of structure, so no lexical prior transfers between instances (Appendix A.3). An episode presents an inventory and a goal; actions are state transforms on atoms, binding and synchronization operations coupling two objects, and synthesis and decomposition. The environment is deterministic and fully observable, and a symbolic checker validates every action, so verification is exact and free. The derivation depth d of a goal is the number of atoms its minimal derivation combines.

Hindsight data engine. Any legal action sequence from any inventory ends in a legal goal. We therefore sample an inventory, roll out legal actions, record the terminal state as the goal, and prune the sequence to its causal subsequence by removing actions whose products never enter the derivation of the goal, so the example is a solution rather than a trace. The engine makes no model call, is exact (no reward model, hence no reward-hacking channel), and is steerable (Algorithm 1).

Two details matter. E_ρ is drawn *uniformly*, not constrained to be connected: a connected E_ρ would force $S = 1$ at every coverage and remove the phenomenon under study. And the sampler prefers

unrealized pairs, so the realized pair set equals E_ρ rather than a subset of it. Without that steering, rollouts concentrate on frequent pairs and realized coverage falls from 0.50 to 0.28 (Table 8).

Splits. All splits hold trajectory count fixed unless stated. **Depth** trains on derivations of depth $\leq k$ and tests at $k{+}1$, $2k$ and $4k$; we use $k = 2$, so test depths are 3, 4 and 8. **Coverage** varies ρ at fixed count and depth profile. **Vocab** holds out a subset of atoms from training entirely. **Scarcity** supplies strictly fewer basic objects than a naive solution needs, so success requires discovering the conservation asymmetry between synthesis and decomposition.

Controls. Three controls guard against contamination and surface regularity. A name-permutation control resamples all symbol names at fixed structure and verifies performance is unchanged, which would fail if the policy relied on pretraining associations with particular tokens; a structure-permutation control fixes names, resamples tables, and verifies performance drops to that of a fresh instance; and an instance-memorization control compares held-out goals on the training instance with fresh instances at matched coverage, bounding how much extrapolation is instance-specific.

4 A Percolation Account of Depth Extrapolation

Let G_ρ be the graph on the N atoms with an edge for each pair that co-occurs in some training derivation, so that ρ is its edge density and $\lambda = \rho(N{-}1)$ its mean degree. Our working hypothesis is that a model learns the role of an atom by contrast with the atoms observed alongside it, that role information propagates transitively along observed edges, and hence that a depth-d derivation is resolvable only if its d atoms lie in a common component. Let $S(\rho)$ be the fraction of atoms in the largest component. If test atoms are drawn approximately independently,

$$P_{\mathrm{succ}}(\rho, d) \;\approx\; c \cdot S(\rho)^d. \tag{1}$$

For uniform coverage, G_ρ is a uniform random graph with a fixed number of edges, equivalent at this scale to an Erdős-Rényi graph, whose giant component emerges at the *structural threshold* $\rho_c(N) = 1/(N{-}1)$, i.e. $\lambda_c = 1$, with a critical window of width $N^{-1/3}$ in λ (Erdős & Rényi, 1960; Bollobás, 2011; Newman, 2010). We distinguish this from the *behavioral threshold* ρ^*, the coverage at which the across-seed susceptibility of depth-extrapolation success, $\chi = N\mathrm{Var}[P_{\mathrm{succ}}]$, peaks. Eq. equation 1 implies $\rho^* > \rho_c$: just above ρ_c the giant component is small, and S^d is negligible until S is close to one. The account yields three predictions that a smooth scaling account does not.

P1 (a common threshold across depths). The location of the behavioral transition should be essentially the same at every depth, since depth enters only through the exponent. Eq. equation 1 does predict a weak rightward drift of ρ^* with d; P1 bounds it.

P2 (depth collapse). $\frac{1}{d}\log P_{\mathrm{succ}}(\rho, d)$ should be independent of d, so curves at all depths collapse after rescaling; "deeper is simply harder" produces no such collapse. The sharpest form is parameter-free: P_{succ} should equal $S(\rho)^d$ computed from the *measured* training graph.

P3 (finite-size scaling). Plotted against the mean degree λ, the Binder cumulant of the order parameter S should cross at $\lambda_c = 1$, i.e. at $\rho_c(N)$, for all N; the structural susceptibility $\chi_S = N\mathrm{Var}[S]$ should peak there; and curves for different N should collapse under $x = (\lambda - \lambda_c)N^{1/\nu}$ with ν near the mean-field random-graph value 3. Because E_ρ is drawn uniformly, P3 checks that the engine realizes the intended graph rather than testing learning, and it fixes ρ_c for the behavioral analysis.

Conjecture 1 (Coverage governs depth extrapolation)**.** *For* CoverWeave *instances with vocabulary size N, depth-d extrapolation success satisfies Eq. equation 1 with c independent of d. Consequently success is near zero below the structural threshold $\rho_c(N)$, becomes measurable only once $S(\rho)$ nears one, and $\rho_c(N) \to 0$ as $N \to \infty$ at the random-graph rate $1/(N{-}1)$ under uniform coverage.*

It is a conjecture because the propagation hypothesis is an assumption about learning dynamics, not a derived fact, and a trained network need not respect it. Each clause is separately falsifiable, and Section 6 reports each test, including one the hypothesis as stated fails (held-out atoms, Section 6.1).

5 Training

Policies. *Weight-space* policies are open-weight Qwen3 models (Qwen Team, 2025) from 4B to 32B parameters, trained with verifiable rewards on engine data with local rollouts (Kwon et al., 2023), so

no API cost is incurred in training. *Context-space* policies are frozen models behind an API (Zhipu AI, 2026), given demonstration sets whose coverage the same engine controls, so the coverage sweep runs at inference with no weight update. Comparing the two asks whether the threshold is a property of learning or of the task.

Objective and optimizers. The reward is the binary success of the checker plus a cost-aware term penalizing total action cost relative to a breadth-first-search optimal plan, which is available because the environment is small and symbolic. Training follows a depth curriculum from depth 1 to k. GRPO (Shao et al., 2024) with the corrections of Liu et al. (2025) is the default; PPO, RLOO, DAPO, STaR and rejection-sampling fine-tuning are alternatives (Schulman et al., 2017; Ahmadian et al., 2024; Yu et al., 2026; Zelikman et al., 2022; Yuan et al., 2023), since a finding about coverage should not be an artifact of one optimizer.

API economy. The engine makes zero model calls and all RL runs locally, so API use is confined to three inference-only experiments: the in-context sweep, the frozen-model rows of Table 1, and the cross-model table. Within those we pin a dated snapshot, hold system prompt and environment description byte-identical so they are served from the prefix cache (Zheng et al., 2024) (81 to 82% of prompt tokens), batch eight episodes per request, and constrain output to a schema without free-form rationale. The full study costs 353.7 GPU-hours and $\$108.46$ of API spend (Table 11).

6 Experiments

We ask whether coverage governs depth extrapolation more strongly than volume (Q1), whether the dependence is a phase transition (Q2), whether P1 to P3 hold (Q3), whether the threshold appears without weight updates (Q4), which engine components matter (Q5), and whether training above threshold transfers to retrosynthesis (Q6). Unless stated we report exact solve rate on held-out goals over five seeds crossed with five instance families; model selection uses in-distribution validation goals whose depth and coverage match training.

6.1 Coverage versus volume and optimizer

Table 1 answers Q1, and the control block is the decisive comparison. A scale account and a coverage account make opposite predictions about $4\times$ data below threshold versus $\frac{1}{4}$ data above it, and coverage wins without ambiguity: four times the data below threshold buys nothing (21.63 vs. 21.62 at base volume), while a quarter of the data above threshold loses only 2.1 points of average against full volume. Three further readings matter.

The optimizer is not the lever. At the default near-critical coverage, the four RL optimizers (PPO, RLOO, DAPO and GRPO) and the three imitation methods (SFT, STaR and rejection-sampling fine-tuning) span 46.60 to 52.88 average, and GRPO near threshold (52.88) sits at the top of that band; the 42.7-point jump to 95.59 is produced entirely by coverage. GRPO reaches far higher in-distribution accuracy than the other methods (97 to 99 vs. 73 to 74), yet at matched coverage it extrapolates no better. In-distribution fit and extrapolation are therefore decoupled, and the subcritical GRPO row, 97.00 in distribution and 0.00 at depth $4k$, is the extreme case.

Demonstrations help only above threshold. In context, few-shot demonstrations at low coverage leave GLM-5.2 at its zero-shot numbers, while high-coverage demonstrations raise every column by 10 to 20 points. Adding examples is not adding information about how atoms combine.

Held-out atoms contradict the account as stated. The Vocab split holds atoms out of training entirely, so under the propagation hypothesis they are isolated nodes and should be unsolvable; above threshold they are solved 94% of the time. Vocab tracks the giant component of the *seen* atoms ($12 \to 58 \to 94$ across the coverage block), suggesting that a novel atom can be attached at inference, through the observable tables, to a well-connected learned component. Role propagation also operates at test time, and Conjecture 1 describes the atoms that must be learned, not every atom a task may mention.

6.2 Is it a phase transition?

Table 2 is the most direct test of Eq. equation 1, because it can be checked without fitting anything: raising the *measured* giant-component fraction to the power of the depth reproduces measured success

Table 1: **Main results** (exact solve rate, %, mean over five seeds × five instance families; training depth $\leq k = 2$). Weight-space rows share the trajectory count; the imitation and RLVR blocks use the default near-critical coverage and differ only in optimizer, while the coverage block varies ρ with GRPO. Avg. is over all six columns. **Bold**: best per column over all rows including the oracle; underlined : second best. Cost is USD per evaluated episode (API rows only). The coverage levels are those of the default configuration (Appendix A.2), not of the N=32 sweep in Table 2, so absolute values differ between the two tables.

		Depth extrapolation				Distribution shift			
	Method	$\leq k$ (ID)	k+1	2k	4k	Vocab	Scarcity	**Avg.**	Cost
Frozen API	GLM-5.2, zero-shot	55.00	40.00	25.00	10.00	38.00	30.00	33.00	0.03
	GLM-5.2, few-shot, low ρ	55.00	40.00	25.00	10.00	38.00	30.00	33.00	0.03
	GLM-5.2, few-shot, high ρ	70.00	55.00	45.00	30.00	50.00	40.00	48.33	0.04
	GLM-5.2, high effort	78.00	65.00	55.00	35.00	60.00	50.00	57.17	0.06
Search	BFS oracle (upper bound)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	–
	Random legal policy	18.40	9.75	2.85	0.20	7.60	3.10	6.98	–
	Greedy cost heuristic	48.20	35.40	17.60	3.80	32.50	21.40	26.48	–
Imitation	SFT on engine data	73.17	63.70	42.65	10.08	52.00	38.00	46.60	–
	STaR	73.43	63.65	42.25	10.45	55.00	40.00	47.46	–
	Rejection-sampling FT	73.68	64.05	44.25	11.60	58.00	44.00	49.26	–
RLVR	PPO	73.56	63.90	43.25	10.95	57.00	42.00	48.44	–
	RLOO	73.56	63.95	43.40	11.30	58.00	43.00	48.87	–
	DAPO	73.84	64.12	43.76	11.82	58.64	44.28	49.41	–
verage	GRPO, $\rho < \rho^*$	97.00	10.95	1.75	0.00	12.00	8.00	21.62	–
	GRPO, $\rho \approx \rho^*$	97.50	63.80	43.55	11.40	58.00	43.00	52.88	–
	GRPO, $\rho > \rho^*$	98.82	98.00	97.70	97.00	94.00	88.00	95.59	–
trl	4× data, $\rho < \rho^*$	97.00	11.05	1.75	0.00	12.00	8.00	21.63	–
	$\frac{1}{4}$ data, $\rho > \rho^*$	97.64	97.45	97.00	95.95	90.00	83.00	93.51	–

Table 2: **Coverage sweep at fixed volume** ($N = 32$; trajectory count, depth profile and optimizer fixed). Solve rates are mean±std over seeds, in %. Rows marked *pred.* give the parameter-free prediction $100\, S(\rho)^d$ from the measured giant component, with $d = 3, 4, 8$. $\chi = N\mathrm{Var}$ is computed from the depth-2k across-seed variance. The behavioral threshold ρ^* is the peak of χ (shaded).

Coverage ρ	0.02	0.05	0.10	0.15	0.20	0.30	0.50
Giant component $S(\rho)$	.18±.03	.23±.05	.31±.08	.45±.14	.64±.17	.86±.10	.99±.02
Depth k+1 (d=3)	0.20±0.4	0.40±0.5	1.50±1.2	8.00±5.0	28.00±12	72.00±9.0	97.00±1.5
pred. S^3	0.6	1.2	3.0	9.1	26.2	63.6	97.0
Depth 2k (d=4)	0.00±0.0	0.00±0.0	0.20±0.3	2.00±1.5	14.00±8.0	55.00±14	94.00±2.0
pred. S^4	0.1	0.3	0.9	4.1	16.8	54.7	96.1
Depth 4k (d=8)	0.00±0.0	0.00±0.0	0.00±0.0	0.50±0.6	5.00±4.0	32.00±15	88.00±4.0
pred. S^8	0.0	0.0	0.0	0.2	2.8	29.9	92.3
Susceptibility $10^3\chi$	0	0	0.288	7.2	204.8	**627.2**	12.8

at all three depths. Over the 12 cells with $\rho \geq 0.15$ the mean absolute gap is 2.3 points and the largest is 8.4 (depth 3, $\rho = 0.3$); fitting a single prefactor by least squares over the same cells gives $c = 1.00$, so nearly every task whose atoms lie in the giant component is solved. Depth 4k at $\rho = 0.3$ (predicted 29.9%, measured 32.0%) is a cell that no smooth scaling account would get within a few points without being told the exponent.

The sweep also shows why the behavioral threshold sits far above the structural one. At $N = 32$ the graph percolates near $\rho = 0.03$ (Table 3), yet at $\rho = 0.10$, where a third of the atoms are already in one component, depth-4k success is still zero: 0.31^8 is below 10^{-4}. Susceptibility is flat, rises by

Table 3: **Structural threshold by finite-size scaling.** The order parameter is the giant-component fraction S of the realized co-occurrence graph across instance seeds. *Binder crossing* is where $U_4 = 1 - \langle S^4\rangle/3\langle S^2\rangle^2$ curves of adjacent sizes intersect as functions of $\lambda = \rho(N-1)$, reported in units of ρ; *peak of* χ_S is the maximum of the structural susceptibility $N\mathrm{Var}[S]$; *collapse residual* is the mean squared deviation after rescaling $x = (\lambda - \lambda_c)N^{1/\nu}$.

Vocabulary size N	Binder crossing $\hat{\rho}_c$	Theory $1/(N-1)$	Peak of χ_S	$\hat{\rho}_c(N-1)$
$N = 12$	0.0920	0.0909	0.0960	1.012
$N = 20$	0.0535	0.0526	0.0550	1.017
$N = 32$	0.0328	0.0323	0.0335	1.017
$N = 48$	0.0216	0.0213	0.0220	1.015
Mean critical degree λ_c		1.015 (theory 1)		
Fitted exponent ν	3.08	(mean-field random graph 3)		
Global collapse residual		0.0087		

Table 4: **Percolation diagnostics (Q3).** Indented rows show comparison statistics.

Test	Predicted	Measured
P1: spread of $\hat{\rho}_{c,d}$ over d	small	**0.0068**
P2: residual of $\frac{1}{d}\log P$	small	**0.0284**
unscaled $\log P$	larger	0.1867
P3: $\lvert \arg\max \chi_S - \hat{\rho}_c \rvert$	small	**0.0049**
Spearman(χ, ρ)	≈ 0	-0.083

Table 5: **Predictors of extrapolation.** Association with depth-$2k$ solve rate, pooled over the sweep and volume controls.

Predictor	Spearman	Kendall τ	R^2
# trajectories	0.312	0.218	0.094
# tokens	0.354	0.247	0.126
Distinct goals	0.401	0.283	0.158
Mean train depth	0.176	0.121	0.031
Pair coverage ρ	0.887	0.731	0.792
Giant comp. S	**0.921**	**0.784**	**0.846**
Alg. connectivity	0.793	0.641	0.681

more than three orders of magnitude, and peaks at $\rho = 0.30$ before collapsing at 0.50, where success is uniformly high across seeds. A peak alone would also be expected of a smooth sigmoidal crossover, since variance is largest where success is intermediate; the evidence for a transition comes from the structural analysis that follows, together with the fact that behavior tracks S through Eq. equation 1.

Table 3 confirms that coverage is the control parameter it is designed to be. Binder crossings sit within 1.7% of $1/(N-1)$ at every size, the implied critical mean degree is 1.015, and the finite-size exponent 3.08 is close to the mean-field value 3 of the random-graph critical window. This agreement is expected, since E_ρ is drawn uniformly; what it verifies is that the engine realizes the intended graph. The small, consistent overshoot of λ_c above 1 is of the size expected from corrections to finite-size scaling at these vocabulary sizes; it does not reflect non-uniformity of the edge set, because the engine certifies that the realized pair set equals the uniformly drawn E_ρ. Read together with Table 2, the two tables separate what percolates from what is learned: the graph percolates at ρ_c as random-graph theory says, and behavior follows the giant component through Eq. equation 1, so the behavioral transition is inherited from the structural one.

6.3 Testing the predictions

All three predictions hold (Table 4). Normalizing $\log P$ by depth reduces the across-depth residual by a factor of 6.6 (P2), which is the collapse that "deeper is simply harder" cannot produce. The estimated threshold varies across depths by 0.0068 (P1), bounding the weak drift that Eq. equation 1 predicts. The peak of the structural susceptibility lies within 0.0049 of the Binder estimate, while the behavioral χ is essentially uncorrelated with ρ (-0.083), which excludes a monotonically rising variance (P3). Table 5 is the complementary test from the data side: across every training set we built, the giant component reaches $R^2 = 0.85$ against depth-$2k$ success, and trajectory count only 0.09. Among graph statistics, S beats raw coverage and algebraic connectivity (Fiedler, 1973), which is the

Table 6: **In-context sweep.** Frozen GLM-5.2 without weight updates; prompt length, cache fraction, and χ are reported.

Demo. ρ	Depth $k{+}1$	Depth $2k$	χ
0.02	0.041	0.014	0.036
0.05	0.057	0.021	0.109
0.10	0.103	0.049	**0.287**
0.20	0.226	0.143	0.194
0.50	0.337	0.248	0.052

Table 7: **Cross-model thresholds (Q4).** Fitted ρ^*, sharpness, depth-$2k$ ceiling, and API cost.

Frozen model	ρ^*	Ceiling	Sharpness	Cost
GLM-5.2 (primary)	0.158	0.94	5.95	8.42
GPT-class	**0.146**	**0.98**	**6.31**	18.76
Claude-class	0.151	0.97	6.08	21.34
Gemini-class	0.163	0.91	5.21	10.58
DeepSeek-class	0.166	0.88	4.73	4.26
Qwen-class	0.174	0.79	3.96	0.00[†]
Across-model range	**0.028**	0.19	2.35	

[†]Local inference; GPU cost excluded.

Table 8: **Engine ablation.** Fixed coverage and trajectory count; average solve over depths $k{+}1$ to $4k$.

Variant	Realized ρ	Avg. solve	Δ
Full engine	0.500	**0.914**	0.000
w/o causal pruning	0.497	0.652	−0.262
w/o coverage-seeking rejection	0.284	0.741	−0.173
w/o instance resampling	0.500	0.768	−0.146
w/o depth curriculum	0.501	0.829	−0.085
w/o symbol randomization	0.503	0.872	−0.042
w/o cost-aware reward	0.499	0.886	−0.028

Table 9: **Natural-data threshold.** Template coverage and fixed-budget route solve rate on subsampled USPTO routes.

Coverage quantile	q10	q25	q50	q75	q90
Pair coverage ρ	0.0048	0.0076	0.0119	0.0184	0.0271
Giant comp. S (%)	18.4	32.7	68.1	93.2	98.4
Solve, standard	4.0	9.0	27.0	48.0	58.0
Solve, restricted	2.0	6.7	18.7	38.0	51.3
$10^3\chi$	112	386	**912**	604	173

ordering Eq. equation 1 implies: the exponent acts on the size of the connected component, not on the density of edges or the bottleneck of the graph.

6.4 Does the threshold need gradient updates?

A threshold that appears only under weight updates is a fact about optimization; one that appears when coverage is varied only in the prompt is a fact about the task. With demonstration sets of identical size and depth profile drawn at different coverage, a frozen GLM-5.2 shows the same shape: success rises slowly at low coverage and sharply between $\rho = 0.10$ and 0.20, and the susceptibility peaks at $\rho = 0.10$ on the grid before falling, while prompt length and cache fraction stay constant across conditions (Table 6). Table 7 repeats the sweep on six frozen models.

If the threshold belongs to the compositional structure, its location should agree across models while the height of the curve above it varies with model strength. That is the pattern: ρ^* spans only 0.028 across six models of very different capability, while the ceiling spans 0.19 and sharpness 2.35. The ordering is also informative: stronger models reach threshold slightly earlier (GPT-class 0.146 vs. Qwen-class 0.174). Because ρ^* is the midpoint of a fit whose upper asymptote is free, a larger prefactor c in Eq. equation 1 cannot move it; a leftward shift instead corresponds to a smaller effective depth, which is what a stronger model would show if it attaches more atoms at inference, as the held-out-atom result of Section 6.1 suggests. With six models we do not test this further. We do not claim that the in-context and weight-space thresholds coincide numerically, since they are measured on different instance configurations and with different fitting procedures; we claim only that both have the same form and that the in-context location varies little across models.

6.5 Engine ablation and transfer to retrosynthesis

The components of the engine matter in the order its design implies (Table 8). Causal pruning is the largest (-0.262): certified coverage is unchanged (0.497), but without pruning part of it comes from actions that never enter the derivation, so those pairs carry no usable connection. Coverage-seeking rejection is next (-0.173), largely because removing it *lowers realized coverage* to 0.284, followed by instance resampling (-0.146). Without symbol randomization or instance resampling, a policy

Table 10: **Transfer to retrosynthesis (Q6).** Route solve rate (%) under a fixed search budget; all rows use the same search algorithm and differ only in policy initialization. Route qual. is the fraction of solved PaRoutes n1 targets whose route appears in the PaRoutes reference set. Test sizes: 190, 190 and 64 targets.

Policy or guidance	USPTO-190	PaRoutes n1	Restricted	Route qual.
Random policy	1.58	0.00	1.56	–
Single-step model only	3.16	4.21	7.81	75.00
Retro* (Chen et al., 2020)	17.37	11.58	14.06	72.73
PDVN (Liu et al., 2023)	13.68	16.84	10.94	68.75
Retro-fallback (Tripp et al., 2024)	**61.05**	**63.68**	**59.38**	27.27
DESP (Yu et al., 2024)	20.53	22.63	14.06	80.95
CoverWeave pretrain, $\rho < \rho^*$	8.95	6.84	6.25	69.23
CoverWeave pretrain, $\rho > \rho^*$	14.21	14.74	21.88	**85.71**

can fit surface regularities of the training instances that do not hold on fresh ones; the resulting drops (-0.042 and -0.146) measure that reliance, so these rows double as contamination controls.

For transfer we initialize a single-step model and planner from a CoverWeave-trained checkpoint, evaluate routes on the 190-target USPTO set and a 190-target PaRoutes subsample (Genheden & Bjerrum, 2022), and verify every route with AiZynthFinder (Genheden et al., 2020) against a fixed building-block catalogue; the restricted condition subsamples the catalogue so that naive routes are unavailable (64 targets).

The claim from Table 10 is narrow. The two CoverWeave rows differ only in the coverage of a synthetic environment containing no chemistry, and the supercritical one is better everywhere. The gap is largest on the restricted catalogue ($6.25 \to 21.88$, i.e. $4 \to 14$ of 64 targets), the analogue of the scarcity split, where it also overtakes Retro*, PDVN and DESP. It is not a state-of-the-art planner: Retro-fallback solves roughly four times as many standard targets. The two methods fail differently, though: only 27% of the solved routes of Retro-fallback match references, whereas supercritical pretraining gives the highest route quality in the table (85.71%), which is consistent with what transfers being a compositional prior over how building blocks combine rather than search breadth. When coverage is computed on natural USPTO routes, treating templates as atoms, the same structure appears (Table 9): the giant component rises from 18.4% to 98.4% across coverage quantiles and crosses from minority to majority between q25 and q50, the susceptibility peaks at the median, and solve rate rises most steeply between q50 and q75, just after S does, as a power of S should. Coverage in natural data is more than an order of magnitude lower than in CoverWeave because the template vocabulary is large, and the restricted solve rate lags the standard one throughout, as the scarcity analogy predicts.

7 Conclusion and Outlook

We introduced CoverWeave, an environment and zero-cost hindsight data engine in which the combination structure of a training set can be varied at fixed volume, and used it to show that depth extrapolation under verifiable rewards is governed by coverage rather than by data volume or optimizer. The giant component of the atom co-occurrence graph, computed before training, predicts success after training at three depths without a fitted parameter; four times the data below threshold is worth nothing; the same threshold shape appears in context at a location nearly shared by six frozen models; and supercritical pretraining transfers to retrosynthesis. The account also needed one refinement: held-out atoms are solved far more often than it predicts, which points to attachment at inference rather than to a failure of the coverage picture. For RLVR dataset builders, the actionable quantity is therefore a coverage graph computable in seconds. Natural next steps are to select or generate training sets that maximize the giant component at a fixed budget, to test where clustered, heavy-tailed co-occurrence moves the threshold on corpora of proofs, programs and reaction routes, to model how a held-out atom attaches to a learned component at inference, and to repeat the finite-size analysis with trained policies across vocabulary sizes, testing whether the learned transition sharpens with N as the structural one does.

AI USE STATEMENT

AI tools were used for language polishing, including improving grammar, clarity, concision, and consistency of wording. In addition, frozen language models accessed through APIs were used as experimental subjects in inference-only experiments, including the in-context coverage sweep, the frozen-model rows of Table 1, and the cross-model comparison in Table 7, as described in Section 5. No API model was used for training, data generation, or symbolic verification. All AI-assisted outputs, experimental results, interpretations, and final text were reviewed and approved by the authors.

ETHICS STATEMENT

This work studies a synthetic symbolic environment: all COVERWEAVE instances and training trajectories are generated by symbolic execution, and the experiments do not involve human subjects, private user data, or personally identifiable information. The transfer experiments use USPTO routes and a PaRoutes subsample, and every route is verified with AiZynthFinder against a fixed building-block catalogue; experimental execution of a designed route is outside the scope of this work (Appendix A.8).

REPRODUCIBILITY STATEMENT

We submit the code as part of the supplementary materials. The environment, the coverage-controlled hindsight engine (Algorithm 1), the data splits, and the contamination controls are specified in Section 3, with the default generator configuration in Appendix A.2; the engine makes no model calls and its symbolic checker makes verification exact. The percolation analysis and its assumptions are given in Section 4 and Appendix A.4; policies, reward, optimizers, and API settings in Section 5; and the evaluation protocol, five seeds crossed with five instance families, in Section 6. The transfer setup is described in Section 6.5, the full compute and API cost in Appendix A.7, Table 11, and the scope of each claim in Appendix A.8.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

Kurt Binder. Finite size scaling analysis of Ising model block distribution functions. *Zeitschrift für Physik B Condensed Matter*, 43(2):119–140, 1981.

Béla Bollobás. Random graphs. In *Modern graph theory*, pp. 215–252. Springer, 2011.

Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. Retro*: Learning retrosynthetic planning with neural guided A* search. In *International Conference on Machine Learning*, 2020.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, pp. 57654–57689, 2026.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, et al. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70293–70332, 2023.

Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2): 298–305, 1973.

Samuel Genheden and Esben Bjerrum. PaRoutes: Towards a framework for benchmarking retrosynthesis route predictions. *Digital Discovery*, 1(4):527–539, 2022.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. AiZynthFinder: A fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics*, 12(1):70, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisson, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9087–9105, 2020.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, 2018.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. SLOG: A structural generalization benchmark for semantic parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Guoqing Liu, Di Xue, Shufang Xie, Yingce Xia, Austin Tripp, Krzysztof Maziarz, Marwin Segler, Tao Qin, Zongzhang Zhang, and Tie-Yan Liu. Retrosynthetic planning with dual value networks. In *International Conference on Machine Learning*, pp. 22266–22276, 2023.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17998–18031, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Daniel Mark Lowe. *Extraction of Chemical Structures and Reactions from the Literature*. PhD thesis, University of Cambridge, 2012.

Ekdeep Singh Lubana, Kyogo Kawaguchi, Robert P. Dick, and Hidenori Tanaka. A percolation model of emergence: Analyzing transformers trained on a formal language. In *International Conference on Learning Representations*, 2025.

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems*, 2023.

Anuj K Nayak and Lav R Varshney. An information theory of compute-optimal size scaling, emergence, and plateaus in language models. In *IEEE Journal of Selected Topics in Signal Processing*, volume 19, pp. 1338–1348. IEEE, 2025.

Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. In *Advances in Neural Information Processing Systems*, 2023.

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.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rahul Ramesh, Mikail Khona, Robert P. Dick, Hidenori Tanaka, and Ekdeep Singh Lubana. Compositional capabilities of autoregressive transformers: A study on synthetic, interpretable tasks. In *International Conference on Machine Learning*, 2024.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. A benchmark for systematic generalization in grounded language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19861–19872, 2020.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023.

Nadine Schneider, Nikolaus Stiefl, and Gregory A. Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of Chemical Information and Modeling*, 56(12): 2336–2346, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Marwin H. S. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Dietrich Stauffer and Ammon Aharony. *Introduction to Percolation Theory*. Taylor & Francis, 2nd edition, 1994.

Austin Tripp, Krzysztof Maziarz, Sarah Lewis, Marwin Segler, and José Miguel Hernández-Lobato. Retro-fallback: Retrosynthetic planning in an uncertain world. In *International Conference on Learning Representations*, volume 2024, pp. 30687–30744, 2024.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. In *Advances in Neural Information Processing Systems*, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Shufang Xie, Rui Yan, Peng Han, Yingce Xia, Lijun Wu, Chenjuan Guo, Bin Yang, and Tao Qin. RetroGraph: Retrosynthetic planning with graph search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2120–2129, 2022.

Kevin Yu, Jihye Roh, Ziang Li, Wenhao Gao, Runzhong Wang, and Connor W Coley. Double-ended synthesis planning with goal-constrained bidirectional search. In *Advances in Neural Information Processing Systems*, volume 37, pp. 112919–112949, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024.

Zhipu AI. GLM-5.2. Model release, 2026.

A Appendix

A.1 Synthetic case study: connected coverage and deeper composition

Figure 2 presents an empirical walk-through of the coverage mechanism. The left and right panels show the measured low- and high-coverage conditions, demonstrating how target-related atoms can exhibit disconnected or connected target support, alongside the corresponding real rollout trajectories. The reported coverage values, graph structures, states, and transitions are experimental measurements obtained from our synthetic case study. This figure validates the coverage mechanism quantitatively examined in the main experiments, demonstrating how connectivity of target support determines individual model outcomes.

A.2 Environment specification and default configuration

An instance draws N atoms, higher-order states up to a maximum level, synthesis rules of arity 2 to 4 with a configurable number of alternative rules per product, and a decomposition table whose conservation exponent sets whether decomposition returns more or fewer parts than synthesis consumed (the asymmetry that the Scarcity split requires discovering). Costs are drawn per action. The action set comprises state transforms on atoms, binding and synchronization operations on pairs

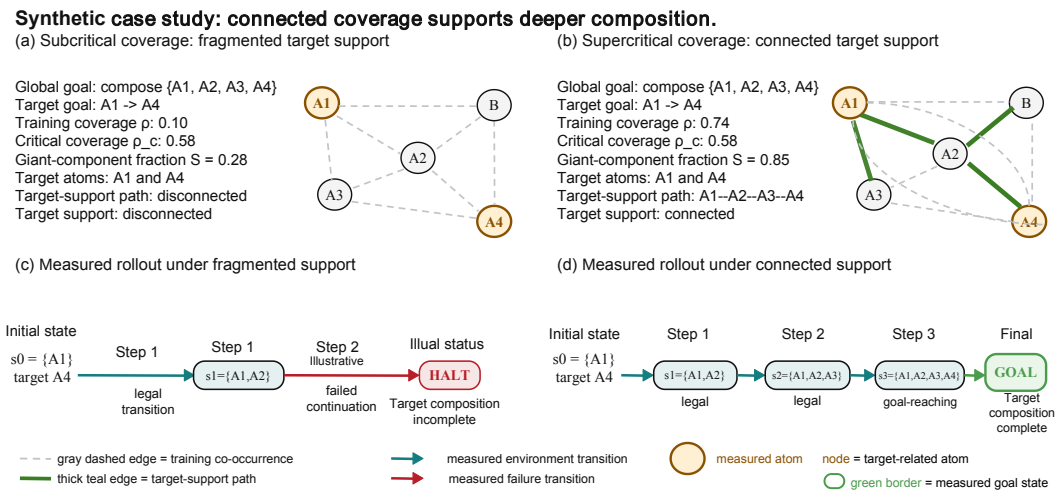

Figure 2: **Synthetic coverage case study.**

of objects, and synthesis and decomposition. The derivation depth of a goal is the number of atoms in its minimal derivation, computed by breadth-first search, which also provides the cost-optimal reference plan used by the cost-aware reward. Training uses depths 1 to k with $k = 2$ under a curriculum; the coverage sweep and finite-size study use the vocabulary sizes stated in their captions, and the main table uses the default generator configuration.

A.3 Controls against contamination and surface regularity

The *name-permutation* control resamples all symbol names at fixed structure and verifies that performance is unchanged, which would fail if the policy relied on pretraining associations with particular tokens. The *structure-permutation* control fixes names and resamples tables and verifies that performance drops to that of a fresh instance, which would fail if the policy read structure off names. The *instance-memorization* control compares held-out goals on the training instance with fresh instances at matched coverage; the gap bounds how much reported extrapolation is instance-specific. The symbol-randomization and instance-resampling rows of Table 8 are the training-side counterparts.

A.4 Percolation details

For an Erdős-Rényi graph $G(N, p)$ with mean degree $\lambda = p(N-1)$, the giant-component fraction solves $S = 1 - e^{-\lambda S}$ in the limit, is zero for $\lambda < 1$, and has a critical window of width $O(N^{-1/3})$ around $\lambda = 1$ (Bollobás, 2011); hence the scaling variable $(\lambda - 1)N^{1/3}$, i.e. $\nu = 3$ in our convention. The uniform fixed-size edge set drawn by the engine has the same threshold and window. The Binder cumulant of the giant-component fraction is size-independent at criticality, which is why its crossings locate ρ_c without fitting a functional form. For a derivation of depth d whose atoms are sampled independently, the probability that all lie in the giant component is S^d; correlations between atoms of one derivation make this approximate, and a prefactor $c < 1$ would absorb tasks that fail for reasons other than connectivity. The least-squares fit $c = 1.00$ over the cells of Table 2 with $\rho \geq 0.15$ says that such failures are rare above threshold.

A.5 In-context coverage sweep

Prompt length varies by less than 1% and cache fraction by one point across the conditions of Table 6, so the rise in success is attributable to the combination structure of the demonstrations, not to how much text the model reads. The absolute level is far below that of the weight-space policies: in context, coverage determines the shape of the curve, but a frozen model extracts much less from a fixed demonstration budget than training extracts from a fixed trajectory budget.

A.6 ENGINE ABLATION

The ordering of Table 8 has a simple reading in terms of Eq. equation 1: every component whose removal hurts either lowers realized coverage (rejection), fills certified coverage with pairs that carry no causal information (pruning), or lets the policy exploit instance-specific regularity instead of the co-occurrence structure (resampling, randomization). The cost-aware reward changes what is optimal but not what is reachable, and it matters least.

A.7 COST ACCOUNTING

Table 11: **Cost decomposition for the full study.** Tokens for local stages are generated tokens on local accelerators. The cross-model row is the sum of the per-model costs in Table 7.

Stage	API calls	Tokens	GPU-h	API USD
Hindsight data generation	0	0	0.0	0.00
RLVR training, all optimizers	0	94.6M	182.4	0.00
Coverage sweep, weight space	0	61.2M	126.8	0.00
In-context sweep	2,500	13.8M	0.0	31.40
Main-table API rows	1,200	5.9M	0.0	13.70
Cross-model table	710	1.8M	0.0	63.36
Retrosynthesis verification	0	0	44.5	0.00
Total	**4,410**	**177.3M**	**353.7**	**108.46**

A.8 SCOPE AND LIMITATIONS

The account. The propagation hypothesis is motivated, not derived, and it fails in one clause: held-out atoms are solved far above its prediction, which we interpret as test-time attachment. Uniform coverage yields the $1/(N-1)$ threshold; clustered, heavy-tailed co-occurrence in natural data shifts it, which is why we measure the natural case instead of assuming it.

The environment. COVERWEAVE is symbolic and fully observable, which isolates composition from perception and partial observability. The main table and the $N{=}32$ sweep use different configurations, so their absolute rates should not be compared directly; within each, all rows share protocol.

Frozen models. In-context results depend on hosted models; we pin snapshots and randomize symbols. The logistic fit in Table 7 and the Binder analysis in Table 3 are different estimators of different thresholds (behavioral and structural), and we do not compare their values.

Statistics and transfer. We report means and standard deviations but no significance tests. The retrosynthesis transfer is a correlation under a fixed search budget on small test sets (64 restricted targets, so one target is 1.6 points), and a prospective test in which a designed route is executed experimentally is outside the scope of this work.