RECALL, REASON, REGRESS: HOW REASONING DATA MIXTURES SHAPE GENERALIZATION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

How does increasing the share of reasoning supervision change an LLM's ability to use what it knows? We investigate this question in a controlled relational reasoning framework that separates fact recall, within-pattern generalization, and cross-pattern transfer. Across four model sizes (4B–32B) and 15 nonzero data mixtures, we vary the number of reasoning demonstrations under a fixed training-row budget while holding the demonstrated composition patterns constant. We identify three empirical regimes: Recall, Reason, and Regress. Knowledge-only training supports perfect fact and rule recall but fails to produce correct composite answers. Sparse reasoning demonstrations bridge this gap, enabling generalization to new instances and undemonstrated compositions, with within-pattern errors decreasing approximately as a power law. Yet further increasing supervision can reverse cross-pattern gains while within-pattern accuracy remains near saturation: doubling the reasoning-data share from 6.4% to 12.8% reduces cross-pattern accuracy by 8.82–20.31 percentage points across all four models, with within-pattern changes of at most 0.16 points. This decline persists when independently probed prerequisites remain accessible; pattern-specific shifts in generated solutions are consistent with specialization to demonstrated patterns. At a common training endpoint, model scale raises observed peak transfer from 55.94% at 4B to 96.74% at 32B, but the grid-maximizing mixture does not vary monotonically with model size. These findings make cross-pattern transfer an essential objective for selecting reasoning mixtures: continued gains on demonstrated patterns can conceal a narrowing of compositional generalization.

Keywords: Large Language Models, Reasoning Supervision, Compositional Generalization, Data Mixtures, Knowledge Utilization

1 INTRODUCTION

The promise of reasoning supervision is to make learned knowledge usable beyond familiar examples. A large language model (LLM) may recall every fact needed for a question yet fail to combine those facts into an answer, exposing a gap between knowledge retrieval and manipulation (Allen-Zhu & Li, 2025b). Worked examples seek to bridge this gap by demonstrating how to reason (Wei et al., 2022; Zelikman et al., 2022; Hsieh et al., 2023), and the success of large reasoning datasets makes scaling such supervision appealing (Toshniwal et al., 2025). However, increasing the number of demonstrations need not expand the composition patterns they cover. More instances of familiar patterns may improve proficiency without broadening the use of knowledge.

Understanding this distinction requires separating *fact recall*, *within-pattern generalization*, and *cross-pattern transfer*: retrieving learned atomic facts, solving new instances of demonstrated patterns, and solving compositions whose rules are known but whose solutions have never been demonstrated. Success on familiar compositions does not guarantee systematic generalization to novel combinations (Lake & Baroni, 2018; Keysers et al., 2020); even reversing a learned relation can remain difficult (Berglund et al., 2024). Measuring these capabilities separately is therefore essential: an aggregate improvement can conceal different responses to the same supervision.

Scaling research characterizes how performance changes with model size and data availability (Kaplan et al., 2020; Muennighoff et al., 2023), while data-mixing research studies how to allocate

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.

training resources across distributions (Xie et al., 2023; Ye et al., 2025a). Closest to our setting, Gu et al. (2025) show that data mixing can induce phase transitions in knowledge acquisition as mixture proportion and model size vary. We ask a complementary question: *once prerequisite knowledge is accessible, how does increasing the share of reasoning supervision change its compositional use?* This shifts the focus from acquiring knowledge to generalizing its use beyond demonstrated patterns.

We investigate this question in a controlled relational reasoning framework. Models first learn atomic facts and composition rules, then receive varying numbers of reasoning demonstrations covering a fixed subset of patterns. Evaluation separates new instances of demonstrated patterns from undemonstrated compositions, with all test entities disjoint from those in reasoning demonstrations. Both groups draw on facts and rules included in knowledge training. Independent probes assess each query's required facts and rule. We study four model sizes (4B, 8B, 14B, and 32B), 15 nonzero mixtures, and a knowledge-only continuation. The training-row budget is fixed: increasing reasoning supervision reduces fact replay while preserving factual coverage.

We organize the results into three empirical regimes of knowledge utilization: **Recall, Reason, and Regress**. These describe a separation among capabilities across mixtures, not shared transition thresholds or a smooth progression at every adjacent mixture. Three findings connect this progression to supervision quantity and model scale.

First, sparse reasoning supervision unlocks compositional use. The knowledge-only control achieves perfect fact and rule recall but zero composite accuracy under our evaluation. Reasoning demonstrations enable both within-pattern generalization and cross-pattern transfer. Within-pattern errors decrease approximately as a power law, quantifying the benefit of additional examples of the demonstrated compositions.

Second, further supervision can reverse transfer gains. Doubling the reasoning-data share from 6.4% to 12.8% reduces cross-pattern accuracy by 8.82–20.31 percentage points across all four models, while within-pattern accuracy changes by at most 0.16 points. The decline persists after conditioning on successful prerequisite probes, whose coverage exceeds 99.4%. Pattern-specific changes in generated solutions are consistent with specialization to demonstrated patterns, although these diagnostics do not establish its causal role.

Third, model scale raises peak transfer without monotonically shifting the best mixture. At the common training endpoint, observed peak cross-pattern accuracy rises from 55.94% at 4B to 96.74% at 32B. The corresponding grid-maximizing shares are 6.4%, 6.4%, 1.6%, and 6.4%. Larger capacity therefore improves attainable transfer without providing a simple rule for choosing the reasoning mixture.

Together, these findings make cross-pattern transfer an explicit objective for reasoning-mixture selection. Near-saturated performance on demonstrated patterns can coexist with declining transfer, even when prerequisite knowledge remains accessible. Evaluating which compositions a model can solve is therefore central to determining whether additional reasoning supervision broadens its use of knowledge.

2 Experimental Design

Our design separates *knowing the ingredients of a solution* from *combining them*, then varies reasoning supervision without expanding the demonstrated pattern set.

Explicit facts and composition rules. Following controlled studies of synthetic knowledge (Gu et al., 2025), we construct 32 maps, each an 8×8 grid with distinct resident names and four neighbor relations: north, south, east, and west. Map-specific relations are assigned by construction, reducing dependence on pretrained factual knowledge. The grid provides a unique answer and enumerable supporting facts for every valid path; boundary-crossing paths are excluded. Atomic facts are (resident, direction, neighbor) triples, while eight labels, A–H, encode ordered direction sequences.

For example, Type-A means north then east. If Alice Hall's north neighbor is Peter Martinez and Peter Martinez's east neighbor is Laura Martin, applying Type-A to Alice Hall yields Laura Martin. Neither an individual fact nor the rule gives this composite answer. A reasoning demonstration combines them by stating the rule, intermediate residents and directions, and final answer. A com-

Table 1: Composition patterns and evaluation groups. All facts and rules are taught; worked demonstrations cover A–D only on demonstration maps. All tests use target maps. N/S/E/W denote north/south/east/west.

Evaluation group	Types and directions	Demos	Items
Within-pattern	A: N→E; B: S→E; C: S→W; D: N→W	Yes	3,136
Cross-pattern, two-hop	E: N→N; F: E→E	No	1,536
Cross-pattern, three-hop	G: N→E→E; H: S→S→W	No	1,344

posite query supplies only the map, starting resident, and Type label. Thus, facts and rules must be recovered from learned knowledge; they are not provided as context.

Separating instance novelty from pattern novelty. We partition the maps into 16 demonstration maps and 16 target maps with disjoint residents. Knowledge training includes facts from *both* sets and *all eight* rules, but reasoning demonstrations cover only A–D on demonstration maps. Every composite test uses a target map. Consequently, A–D tests require applying demonstrated patterns to new entities, while E–H tests require compositions whose rules are taught but whose solutions are never demonstrated (Table 1). This split prevents direct reuse of demonstration answers while keeping test prerequisites within the training knowledge. E/F test unseen two-hop compositions; G/H additionally require a third hop. Because demonstrated A–D patterns contain turns, coverage and path structure are not fully orthogonal; we retain group- and type-specific results rather than attributing every difference to depth.

Establishing knowledge before varying supervision. We use dense Qwen3 models at 4B, 8B, 14B, and 32B (Qwen Team, 2025). Stage 1 teaches 7,168 distinct facts through five question–answer variants each, plus eight rules repeated 40 times: 35,840 fact rows and 320 rule rows. The variants change wording while keeping factual content fixed. Training duration is selected using fact and rule probes, excluding composite accuracy, so knowledge acquisition is not selected for favorable compositional performance. Within each model and seed configuration, all mixture branches start from the same Stage-1 parent. This shared starting point makes the subsequent comparison concern supervision applied to a common knowledge-trained model.

Varying reasoning share while preserving knowledge coverage. Stage 2 trains each branch for three epochs on 20,000 question–answer rows per epoch:

$$N_{\text{demo}} = R, \qquad N_{\text{rule}} = 176, \qquad N_{\text{fact}} = 19{,}824 - R, \qquad r = \frac{R}{20{,}000}. \tag{1}$$

We evaluate 15 nonzero values of R from 8 to 2,560 (0.04%–12.8%), plus a knowledge-only continuation ($R = 0$) to test whether further knowledge training alone enables composition. Demonstration sets are nested, contain distinct A–D instances, and repeat across epochs; all 16 demonstration maps are covered at $R \geq 16$. Thus, larger R adds instances without introducing new demonstrated patterns. Each mixture retains every fact two or three times and 22 independent rows per rule per epoch, preventing increased reasoning share from removing prerequisite facts. Higher R nevertheless reduces fact replay, and demonstrations add further A–D rule statements. The comparison therefore fixes rows and knowledge coverage, not tokens, compute, or total rule exposure.

Measuring access separately from compositional success. Fact recall uses 3,584 questions covering all target-map neighbor relations; rule recall uses eight Type questions. Composite evaluation enumerates all valid starting residents (Table 1). We report final-answer accuracy on A–D (A_S) and E–H (A_U), weighting questions equally and counting failed requests as incorrect. For each composite query, the knowledge gate $\mathcal{G}$ passes only if independent probes recover *every* required fact and the corresponding rule. A three-hop query, for example, requires three correct fact answers and one correct rule answer. Gate coverage $\Pr(\mathcal{G})$ and conditional accuracy $\Pr(C \mid \mathcal{G})$, where C denotes a correct composite answer, distinguish prerequisite accessibility from successful composition. Probe answers are never supplied to the composite query. Main results use epoch-3 checkpoints, with the best-performing mixtures identified by cross-pattern accuracy on the evaluated grid. Complete mixture grids and additional probe and scoring details appear in Appendices A and B.

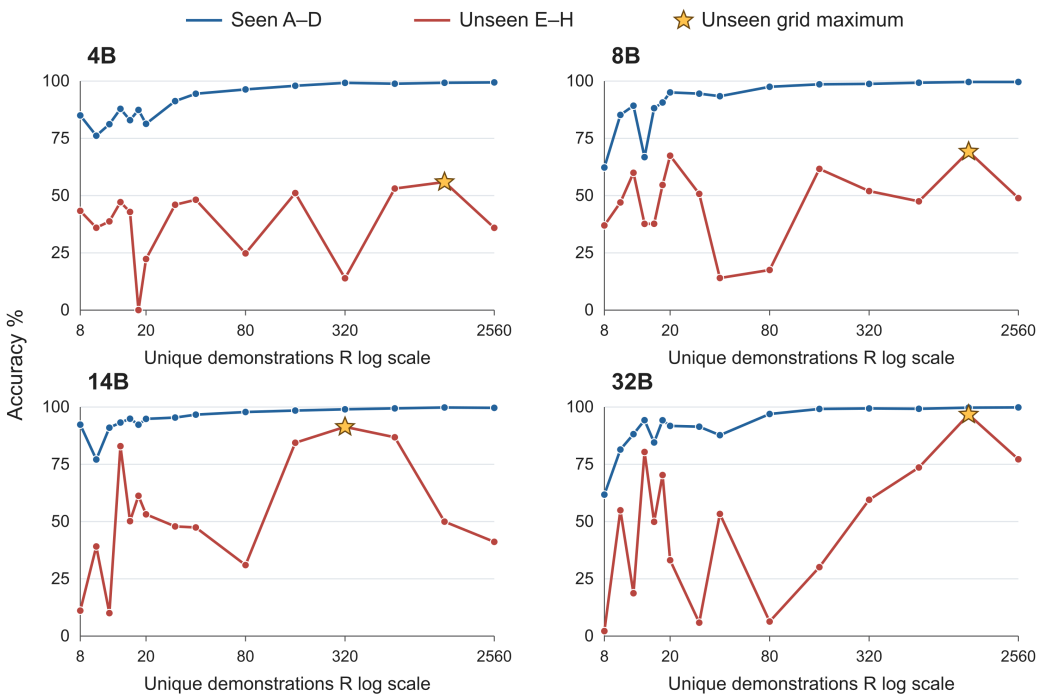


Figure 1: Within-pattern gains can coexist with declining cross-pattern transfer. Blue: new instances of demonstrated patterns A–D; red: undemonstrated patterns E–H, whose rules are taught. Stars mark observed cross-pattern maxima at epoch 3. The logarithmic horizontal axis counts distinct demonstrations R in a 20,000-row mixture; lines connect observations.

3 Recall, Reason, Regress

Increasing reasoning supervision produces different responses in within-pattern generalization and cross-pattern transfer (Figure 1). We organize these responses as *Recall, Reason, and Regress*, using Stage-2 epoch-3 checkpoints unless stated otherwise. Since both evaluation groups use entities absent from the demonstrations, their separation concerns the reach of generalization beyond demonstrated composition patterns.

Recall: accessible knowledge does not ensure composition. In the knowledge-only continuation ($R = 0$), all four models achieve 100% fact and rule accuracy but 0% composite accuracy across the three evaluated epochs. Thus, continued exposure to the required facts and rules does not close the recall–composition gap under our evaluation. During mixed training, factual recall remains high but is not uniformly perfect: it ranges from 97.80% to 100% at the recorded $R = 8$–40 checkpoints. Prerequisite access must therefore be measured at each checkpoint; Section 4 tests whether the transfer decline persists when the joint fact-and-rule gate passes.

Reason: sparse demonstrations enable generalization. At $R = 80$, worked examples constitute only 0.4% of training rows, yet within-pattern accuracy reaches 96.36%–97.83%. This is generalization to new entities, although transfer beyond the patterns remains uneven: at this mixture, 32B scores 96.97% on A–D but only 6.35% on E–H. With a suitable mixture, transfer can nevertheless be strong: 32B reaches 96.74% on E–H at $R = 1{,}280$, despite never receiving E–H solution demonstrations. Such recombination of known components beyond demonstrated instances and patterns is central to compositional-generalization evaluation (Hupkes et al., 2020).

Figure 2 quantifies the within-pattern benefit. Pooling four sizes and 15 nonzero mixtures gives

$$\widetilde{E}_S \approx 0.1751(R/8)^{-0.741}, \tag{2}$$

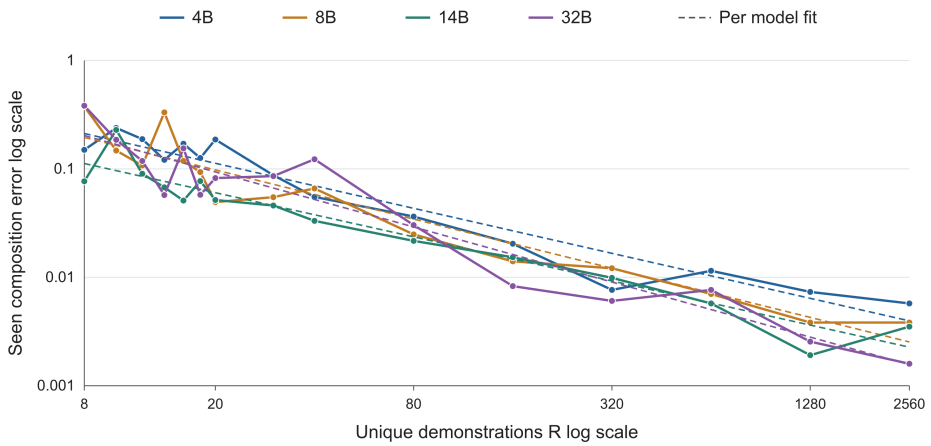


Figure 2: Approximate within-pattern error scaling at epoch 3. Solid curves show raw A–D errors; dashed curves show per-model power laws fitted to continuity-corrected errors. Both axes are logarithmic. This regularity concerns new instances of demonstrated patterns and coexists with the cross-pattern reversals in Figure 1.

Table 2: Cross-pattern regression under the same mixture increase. Changes compare $R = 2{,}560$ with $R = 1{,}280$ at epoch 3, in percentage points. Intervals use 10,000 paired map resamples of fixed predictions.

Model	ΔA_S	ΔA_U	95% map interval	Maps declining
4B	0.16	-20.00	$[-22.67, -17.60]$	16/16
8B	0.00	-20.31	$[-23.30, -17.57]$	16/16
14B	-0.16	-8.82	$[-12.12, -4.69]$	14/16
32B	0.10	-19.55	$[-20.90, -18.26]$	16/16

where $\widetilde{E}_S = (n_{\text{wrong}} + 0.5)/(3{,}136 + 1)$ is continuity-corrected A–D error; reported accuracies are uncorrected. The log-error fit has $R^2 = 0.898$ and predicts approximately 40% less remaining error per doubling of R. Holding out an entire R level across sizes gives an accuracy RMSE of 5.86 percentage points, versus 9.39 for a constant predictor. Per-model exponents range from 0.676 to 0.843, with log-error R^2 of 0.920–0.941, showing similar approximate decay within each size.

This trend allows small reversals: across 24 doubling comparisons from $R = 40$ to 2,560, accuracy increases 20 times, ties once, and falls three times, with a largest loss of 0.38 points. At $R = 2{,}560$, A–D accuracy reaches 99.43%–99.84%. The exponent depends on training progress and fitting choices; restricting epoch-3 data to $R \geq 80$ gives approximately 0.600. The result is an approximate within-pattern scaling relation over the evaluated range, not a universal exponent or a scaling law for cross-pattern transfer.

Regress: further supervision can reverse transfer gains. The same final doubling, $R = 1{,}280 \rightarrow 2{,}560$ (6.4% to 12.8%), lowers cross-pattern accuracy by 8.82–20.31 percentage points across all four models, while within-pattern accuracy changes by at most 0.16 points (Table 2). This comparison uses a common starting mixture rather than a separately selected peak for each model. The decline occurs on all 16 target maps for 4B, 8B, and 32B, and 14 of 16 for 14B; all map-resampling intervals lie below zero. Thus, the transfer loss extends across evaluation maps while remaining largely invisible in demonstrated-pattern accuracy.

All four observed transfer peaks occur below the largest mixture. For 14B, increasing R from 320 to 2,560 lowers cross-pattern accuracy from 91.35% to 41.15%, a 50.21-point loss, while within-pattern accuracy rises from 99.01% to 99.65%. The other models lose 19.55–20.31 points from their peaks. The common final doubling establishes regression under a matched mixture increase;

Table 3: Cross-pattern regression persists after conditioning on prerequisite access. All entries concern E–H at epoch 3 and are percentages. The gate requires every supporting fact probe and the corresponding rule probe to pass.

	Gate coverage		Conditional accuracy	
Model	$R1280$	$R2560$	$R1280$	$R2560$
4B	99.90	99.93	55.96	35.89
8B	99.86	99.86	69.26	48.99
14B	99.90	99.48	49.95	41.19
32B	99.97	100.00	96.73	77.19

peak-to-endpoint losses measure its cumulative extent. Much of 14B's loss occurs before the final doubling, explaining its smaller high-end contrast despite greater overall regression.

Three regimes, distinct generalization outcomes. The regimes describe a separation between recall, compositional success, and declining transfer, with substantial local fluctuations. Training progress also matters: at fixed $R = 20$, 8B's cross-pattern accuracy falls from 76.35% in epoch 2 to 67.43% in epoch 3. Accordingly, these observations do not imply universal thresholds or irreversible decline; Section 5 examines mixture and checkpoint selection. Appendix A records complete grids and supporting diagnostics. Because fact replay, rule exposure, and supervised tokens also vary with R, the scan establishes the behavioral separation without isolating its cause. We next examine prerequisite access and generated solutions to localize the failures.

4 Anatomy of Compositional Regression

Why does cross-pattern transfer deteriorate while demonstrated-pattern performance remains near saturation? We examine three observable components of a solution: access to prerequisite knowledge, declaration of the required rule, and composition of the corresponding relations. This progression localizes the behavioral failure and tests whether its structure resembles the patterns reinforced by supervision.

Regression persists when prerequisite probes pass. At $R = 1{,}280$ and 2,560, the E–H fact-and-rule gate covers 99.48%–100% of questions, yet conditional accuracy declines in every model (Table 3). For 32B at $R = 2{,}560$, all 2,880 unseen questions pass the gate, but only 77.19% receive correct composite answers. The high-end decline therefore cannot be explained by failures on the independently tested prerequisites alone. This distinction concerns measured accessibility: it does not imply that retrieval or knowledge representations within composite prompts remain unchanged.

Correct rule declarations can accompany incorrect answers. Independent probes leave open whether the required rule is available inside a composite response. We therefore inspect the direction sequence declared in generated solutions. At 32B/$R = 30$, fact recall, rule recall, and gate coverage are all 100%, but E–H accuracy is only 5.87% (169/2,880). Only 56 of all 6,016 A–H responses fail the declared-path metric. Even assigning every declaration failure to E–H leaves at least 2,824 unseen responses (98.06%) with correct declarations, and at least 2,655 with correct declarations but incorrect final answers. These conservative bounds establish that stating the right sequence is insufficient for applying it successfully; their derivation and a second case appear in Appendix B. Because this separation already occurs at low dose, it identifies a compositional failure mode rather than a cause specific to high-dose regression.

Pattern identity matters beyond required depth. Figure 3 separates unseen two-hop E/F from unseen three-hop G/H, revealing distinct responses hidden by pooled accuracy. Depth alone is insufficient to explain these responses. For 14B, increasing R from 320 to 1,280 lowers G accuracy from 93.30% to 1.19%, while H improves from 89.14% to 93.60%, although both require three steps (Table 4). For 32B, the final doubling lowers G from 94.05% to 32.14%, while F increases slightly from 99.74% to 99.87%. The collapse of one unseen pattern can therefore coexist with essentially perfect performance on another. Regression therefore selectively disrupts particular compositions, rather than uniformly removing the ability to solve a given depth.

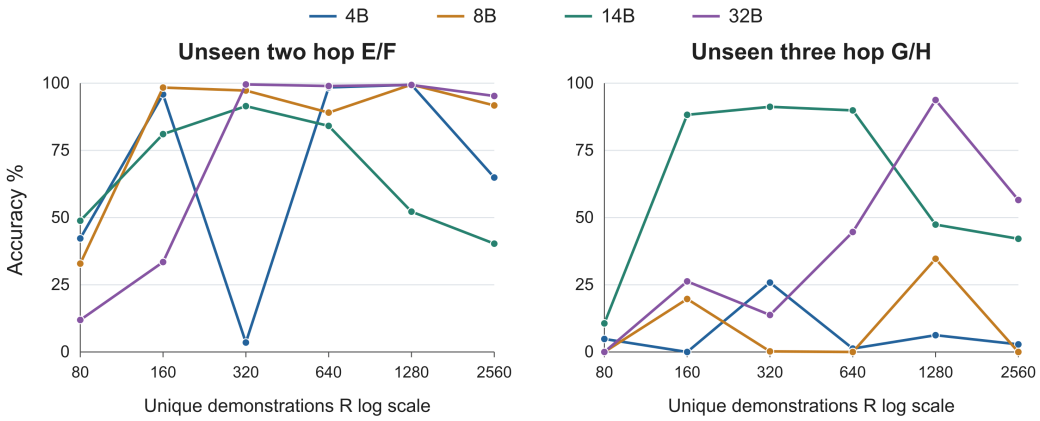


Figure 3: Cross-pattern accuracy by required composition length at epoch 3, for $R = 80$–2,560. E/F are unseen two-hop patterns; G/H are unseen three-hop patterns. All rules are taught, but none of these patterns receives solution demonstrations. E/F excludes demonstrated two-hop patterns A–D.

Table 4: Per-type cross-pattern accuracy (%) at epoch 3. G and H require the same depth but respond differently to increased supervision; unseen two-hop F can also regress.

Model	R	E	F	G	H
14B	320	97.40	85.55	93.30	89.14
14B	1280	99.87	4.56	1.19	93.60
14B	2560	69.66	10.94	3.87	80.36
32B	1280	98.96	99.74	94.05	93.45
32B	2560	90.62	99.87	32.14	80.95

Generated solutions partly shift toward the demonstrated form. All demonstrations use two steps. At 14B/$R = 1{,}280$, 662/672 G responses (98.51%) are parsed as two steps despite the three-step rule. At 32B/$R = 1{,}280$, 671/672 responses in each of G and H are parsed as three steps; at $R = 2{,}560$, 376/672 G responses are instead parsed as two. The accompanying accuracy losses are consistent with an increased preference for the demonstrated solution form.

However, response length does not explain the full failure pattern. At 14B, unseen two-hop F also drops from 85.55% to 4.56% between $R = 320$ and 1,280. At 14B/$R = 2{,}560$, G outputs instead include 367 with no parsed steps, 282 with more than three, and only 23 with exactly three. These are recognized output forms, not measurements of internal reasoning depth: unsupported phrasing and repetition can affect counts, and generation truncation has not been excluded. The counts characterize changes in generated behavior, with parsing details documented in Appendix B.

The supervision imbalance suggests pattern specialization. Demonstrations reinforce both a solution form and its rule definition. At the balanced high-dose points, each A–D Type receives $22 + R/4$ rule statements per epoch, whereas each E–H Type receives 22: 342 versus 22 at $R = 1{,}280$, and 662 versus 22 at $R = 2{,}560$. Thus, holding independent rule rows fixed does not equalize total rule exposure. Higher mixtures jointly reinforce demonstrated rules and worked solutions while reducing standalone fact replay.

Preserved prerequisite access, correct declarations with incorrect answers, and selective output changes together support specialization to the demonstrated pattern distribution as an explanation. They do not isolate its causal contribution: rule exposure, instance diversity, fact replay, and supervised tokens vary together. Exchanging demonstration coverage under matched budgets would test whether the failure follows which patterns are demonstrated. The evidence here establishes the behavioral separation; we next examine how model scale changes attainable transfer and the mixtures that support it.

Table 5: Model scale and cross-pattern transfer at epoch 3. R_U^* and r_U^* denote the best observed mixture among the same 15 nonzero R values evaluated for each model. Accuracies and reasoning-row proportions are percentages.

Model	Peak A_U	R_U^*	r_U^* (%)	A_U at 12.8%
4B	55.94	1280	6.40	35.94
8B	69.24	1280	6.40	48.92
14B	91.35	320	1.60	41.15
32B	96.74	1280	6.40	77.19

5 Strategies for Reasoning Mixture Selection

The preceding results distinguish prerequisite availability from successful composition. We now ask how model scale changes attainable transfer and the mixture needed to achieve it.

Scale raises peak transfer without eliminating regression. At epoch 3, the highest cross-pattern accuracy on the evaluated grid rises from 55.94% at 4B to 96.74% at 32B (Table 5). All four sizes nevertheless regress at the largest reasoning share. Even 32B loses 19.55 percentage points when that share increases from 6.4% to 12.8%. Larger models thus attain stronger transfer under suitable mixtures, while remaining sensitive to supervision composition.

Mixture choice can reverse model rankings. The best observed reasoning shares are 6.4%, 6.4%, 1.6%, and 6.4% across the four sizes, with no monotonic trend. At $R = 320$, 14B outperforms 32B (91.35% versus 59.51%); at $R = 1{,}280$, 32B outperforms 14B (96.74% versus 49.97%) (Figure 1). A comparison at one fixed mixture can therefore give a different ranking from a comparison of each model's best observed performance. Model size alone does not identify a suitable reasoning share.

Selection depends on both checkpoint and transfer objective. Epoch-3 comparisons hold the training endpoint fixed, but earlier checkpoints can transfer better: 8B reaches 76.35% at $R = 20$ in epoch 2, exceeding its epoch-3 grid maximum of 69.24% (Table A4). The evaluation objective also changes the preferred mixture. For 14B at epoch 3, unseen two-hop accuracy peaks at 94.27% with $R = 14$, whereas unseen three-hop accuracy peaks at 91.22% with $R = 320$. Mixture selection consequently depends on the intended weighting of composition types as well as training progress.

These observations motivate jointly selecting the mixture and checkpoint on cross-pattern validation accuracy, subject to predefined requirements for prerequisite coverage and within-pattern accuracy. Validation should reflect the intended composition types, with final evaluation on maps excluded from selection. This proposed protocol requires separate validation; the reported maxima summarize the evaluated grid. The results support coordinating model scale, mixture, and training duration, but do not establish an adaptive mixing schedule. Improving within-pattern accuracy alone does not justify increasing the reasoning share.

6 Related Work

Knowledge access and compositional transfer. Controlled studies distinguish factual storage and retrieval (Allen-Zhu & Li, 2024; 2025a) from flexible knowledge manipulation (Allen-Zhu & Li, 2025b; Berglund et al., 2024). The compositionality gap shows that answering constituent questions correctly does not guarantee answering their composition (Press et al., 2023). Studies of latent reasoning and grokking further highlight dependence on learned solution structure and shortcuts (Yang et al., 2024; 2025; Wang et al., 2024; Lin et al., 2025). Compositional benchmarks formalize generalization to novel combinations of familiar components (Hupkes et al., 2020; Lake & Baroni, 2018; Keysers et al., 2020; Kim & Linzen, 2020), while counterfactual and mathematical evaluations expose brittle transfer (Dziri et al., 2023; Wu et al., 2024; Li et al., 2024; Mirzadeh et al., 2025). We separate new instances of demonstrated patterns from undemonstrated compositions, independently probing the facts and rules required by each query.

Reasoning supervision and transfer. Intermediate solutions support prompting, bootstrapping, and distillation (Wei et al., 2022; Zelikman et al., 2022; Hsieh et al., 2023). Reasoning datasets emphasize scale and diversity (Yue et al., 2024a;b; Toshniwal et al., 2025), difficulty-aware alloca-

tion (Tong et al., 2024), or compact curation (Ye et al., 2025b; Muennighoff et al., 2025); test-time compute provides a distinct scaling axis (Snell et al., 2025). Post-training studies examine differences between supervised fine-tuning and reinforcement learning (Chu et al., 2025; Guo et al., 2025), objectives addressing overfitting and stability (Wu et al., 2025; Zhu et al., 2026), and transfer that varies with training duration, data, and base-model capability (Ren et al., 2026). Our study isolates an important distinction within data scaling: additional demonstrations expand instance diversity while pattern coverage remains fixed. Under this constraint, within-pattern improvement can coexist with cross-pattern regression, without implying that supervised training inherently impairs generalization.

Scaling laws and mixture selection. Scaling studies relate predictive loss and downstream performance to model size, compute, data quantity, and repeated exposure (Kaplan et al., 2020; Hoffmann et al., 2022; Muennighoff et al., 2023; Gadre et al., 2025). Mixture optimization selects domain weights through robust objectives, generalization estimates, or predictive models (Xie et al., 2023; Fan et al., 2024; Ye et al., 2025a; Liu et al., 2025; Tan et al., 2026). Closest to our setting, Gu et al. (2025) identify mixture- and size-dependent transitions in knowledge acquisition. We examine the complementary question of how supervision shapes compositional use when prerequisites remain accessible. Our results distinguish mixtures that improve within-pattern accuracy from those that maximize observed cross-pattern transfer; the Recall–Reason–Regress regimes do not themselves establish critical transitions.

Interpreting reasoning traces. Process supervision demonstrates the value of evaluating intermediate steps (Lightman et al., 2024), but generated explanations need not faithfully represent internal computation (Turpin et al., 2023; Parcalabescu & Frank, 2024). We therefore combine prerequisite probes with rule declarations, intermediate relations, and parsed step counts to characterize observable failures. These measurements constrain explanations of regression while leaving specialization as a causal hypothesis for controlled intervention.

7 Implications and Limitations

Our results distinguish possessing knowledge from generalizing its use. With disjoint demonstration and evaluation entities, within-pattern improvements reflect genuine generalization, yet do not ensure broader compositional transfer. The coexistence of accessible prerequisites, near-saturated within-pattern accuracy, and declining cross-pattern performance shows why neither recall nor familiar-pattern proficiency alone adequately measures reasoning progress. This distinction motivates evaluating reasoning supervision by both proficiency and compositional breadth, and selecting mixtures and checkpoints against the compositions models are expected to solve. It also motivates testing whether broader pattern coverage can sustain transfer as supervision increases. More generally, reasoning data should be characterized by which compositions they teach, alongside how many examples they provide.

The evidence is limited to one model family, synthetic tasks, and a finite training horizon. A fixed row budget couples demonstration quantity with fact replay, rule exposure, and token count; pattern coverage also partly overlaps with path structure and possible shortcuts. Independent probes and generated traces constrain explanations without establishing an internal mechanism. Mixture selection requires validation across training seeds and held-out maps, and longer training may alter the observed regression (Ren et al., 2026).

8 Conclusion

We identify an empirical Recall–Reason–Regress progression: reasoning demonstrations enable generalization, but increasing supervision over fixed patterns can erode cross-pattern transfer despite accessible knowledge and strong within-pattern performance. Larger models achieve higher transfer peaks while remaining sensitive to the mixture. These findings provide an empirical foundation for studying how LLMs learn to reason with acquired knowledge, highlighting supervision composition as a consequential factor in whether that knowledge transfers to unfamiliar reasoning patterns. They motivate a concrete principle for data design and evaluation: progress should be judged by the range of compositions a model can solve, alongside its proficiency on demonstrated patterns.

AI USE STATEMENT

Generative AI tools were used only for literature exploration, reference checking, and language polishing of selected sentences in the final version of the manuscript. The research ideas, overall logic, methodological design and description, and experimental analysis were developed independently by the authors. All technical content, citations, equations, implementations, experimental results, and scientific claims were manually reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 2 describes the synthetic task construction, knowledge and demonstration datasets, two-stage training procedure, reasoning-mixture grid, and evaluation protocols. Sections 4 and 5 specify the prerequisite-conditioned analyses and mixture comparisons, with detailed numerical results and supplementary analyses provided in the appendices. The controlled setting makes the underlying facts, composition rules, and demonstration coverage explicit, supporting independent reconstruction of the experimental tasks. Upon acceptance, we will release the data-generation code, training configurations, and evaluation scripts, including those for prerequisite probing, output-trace analysis, and map-level resampling, to facilitate independent reproduction and further investigation.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *International Conference on Machine Learning*, pp. 1067–1077, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations*, 2025a.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. In *International Conference on Learning Representations*, 2025b.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations*, 2024.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *International Conference on Machine Learning*, pp. 10818–10838, 2025.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. DoGE: Domain reweighting with generalization estimation. In *International Conference on Machine Learning*, pp. 12895–12915, 2024.

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Luca Soldaini, Jenia Jitsev, Alex Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, Thomas Kollar, Yair Carmon, Achal Dave, Reinhard Heckel, Niklas Muennighoff, and Ludwig Schmidt. Language models scale reliably with over-training and on downstream tasks. In *International Conference on Learning Representations*, 2025.

Xinran Gu, Kaifeng Lyu, Jiazheng Li, and Jingzhao Zhang. Data mixing can induce phase transitions in knowledge acquisition. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67, 2020. doi: 10.1613/jair.1.11674.

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. URL `https://arxiv.org/abs/2001.08361`.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, 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*, 2020. doi: 10.18653/v1/2020.emnlp-main.731.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 2018.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-Plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 2961–2984, 2024. doi: 10.18653/v1/2024.acl-long.163.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Tianhe Lin, Jian Xie, Siyu Yuan, and Deqing Yang. Implicit reasoning in transformers is reasoning through shortcuts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9470–9487, 2025. doi: 10.18653/v1/2025.findings-acl.493.

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.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, 2025.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025. doi: 10.18653/v1/2025.emnlp-main.1025.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.acl-long.329.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023. doi: 10.18653/v1/2023.findings-emnlp.378.

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

Qihan Ren, Peng Wang, Ruikun Cai, Shuai Shao, Dadi Guo, Yuejin Xie, Yafu Li, Quanshi Zhang, Xia Hu, Jing Shao, and Dongrui Liu. Rethinking generalization in reasoning SFT: A conditional analysis on optimization, data, and model capability. In *Conference on Language Modeling*, 2026. Accepted paper; arXiv:2604.06628.

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.

Haoru Tan, Sitong Wu, Yanfeng Chen, Jun Xia, Ruobing Xie, Bin Xia, Samm Sun, and Xiaojuan Qi. Fast data mixture optimization via gradient descent. In *International Conference on Learning Representations*, 2026.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. DART-Math: Difficulty-aware rejection tuning for mathematical problem-solving. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data. In *International Conference on Learning Representations*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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*, volume 37, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Sitong Wu, Haoru Tan, Jingyao Li, Shaofeng Zhang, Xiaojuan Qi, Bei Yu, and Jiaya Jia. DLoFT: Gradient-decoupled fine-tuning for generalizable long chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1819–1862, 2024. doi: 10.18653/v1/2024.naacl-long.102.

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*, volume 36, 2023.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 10210–10229, 2024. doi: 10.18653/v1/2024.acl-long.550.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3971–3992, 2025. doi: 10.18653/v1/2025.findings-acl.205.

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*, 2025a.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is more for reasoning. In *Conference on Language Modeling*, 2025b.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MAmmoTH: Building math generalist models through hybrid instruction tuning. In *International Conference on Learning Representations*, 2024a.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. MAmmoTH2: Scaling instructions from the web. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. Anchored supervised fine-tuning. In *International Conference on Learning Representations*, 2026.

A Supporting Results for Reasoning-Supervision Dynamics

A.1 Complete accuracy grids and control accounting

Tables A1 and A2 provide the complete epoch-3 values plotted in Figure 1. All 15 nonzero mixtures are retained without smoothing or removal of local reversals. Mixture proportions refer to training rows, $r = R/20{,}000$. Final-answer accuracy counts failed requests as incorrect. The within-pattern group contains 3,136 A–D questions; the cross-pattern group contains 2,880 E–H questions. Each question has equal weight, so

$$A_U = \frac{1536A_{EF} + 1344A_{GH}}{2880}. \tag{3}$$

The unseen two-hop group E/F contains 1,536 items and is distinct from the all-two-hop group A–F, which contains 4,672 items.

The $R = 0$ condition is a separate Stage-2 knowledge-only continuation, not a substitute Stage-1 checkpoint. Across four sizes and three evaluated epochs, it yields 100% fact and rule accuracy and 0% composite accuracy. It is excluded from the 180 nonzero model–mixture–epoch observations and from the logarithmic fits.

The main-text doubling summary uses the six transitions $40 \to 80$, $80 \to 160$, $160 \to 320$, $320 \to 640$, $640 \to 1280$, and $1280 \to 2560$ for each of four models. Among these 24 comparisons, within-pattern accuracy increases 20 times, ties once, and decreases three times; the largest decrease is 0.38 percentage points. At $R = 2560$, the A–D error counts are 18, 12, 11, and 5 for 4B, 8B, 14B, and 32B, respectively. These subsets contain no marked request failures.

Table A1: Complete within-pattern accuracy grid at epoch 3. Accuracies and mixture proportions are in percent. Values correspond to the A–D curves in Figure 1.

R	Mix (%)	4B	8B	14B	32B
8	0.04	85.01	62.28	92.32	61.77
10	0.05	76.15	85.24	77.10	81.44
12	0.06	81.19	89.25	90.98	88.17
14	0.07	87.88	66.80	93.24	94.26
16	0.08	82.94	88.17	94.90	84.53
18	0.09	87.40	90.66	92.28	94.23
20	0.10	81.35	95.06	94.83	91.77
30	0.15	91.26	94.52	95.41	91.42
40	0.20	94.48	93.40	96.68	87.72
80	0.40	96.36	97.51	97.83	96.97
160	0.80	97.96	98.60	98.47	99.17
320	1.60	99.23	98.79	99.01	99.39
640	3.20	98.85	99.30	99.43	99.23
1280	6.40	99.27	99.62	99.81	99.74
2560	12.80	99.43	99.62	99.65	99.84

Table A2: Complete cross-pattern accuracy grid at epoch 3. Accuracies and mixture proportions are in percent. Bold entries mark each model's observed maximum. Values correspond to the E–H curves in Figure 1.

R	Mix (%)	4B	8B	14B	32B
8	0.04	43.33	36.98	11.11	2.19
10	0.05	35.97	47.01	39.13	54.93
12	0.06	38.75	59.97	10.03	18.72
14	0.07	47.15	37.67	82.92	80.38
16	0.08	42.88	37.71	50.17	49.90
18	0.09	0.03	54.65	61.25	70.31
20	0.10	22.33	67.43	53.16	33.12
30	0.15	46.01	50.76	47.88	5.87
40	0.20	48.19	14.03	47.43	53.33
80	0.40	24.79	17.53	31.01	6.35
160	0.80	51.11	61.67	84.41	30.10
320	1.60	13.92	51.98	**91.35**	59.51
640	3.20	53.09	47.50	86.81	73.61
1280	6.40	**55.94**	**69.24**	49.97	**96.74**
2560	12.80	35.94	48.92	41.15	77.19

A.2 Map resampling and peak-to-endpoint records

The main-text comparison in Table 2 computes $A(R = 2560) - A(R = 1280)$ for each model. For its reported intervals, 10,000 paired map resamples are used. Each resample recomputes pooled E–H accuracy at both mixtures using the same sampled maps and multiplicities, then takes their difference. The comparison therefore preserves map pairing and question weights. Intervals describe map-level variation conditional on fixed model predictions, not training-seed variation, and do not adjust for selection of grid maxima.

Table A3 records the peak-to-endpoint comparisons summarized in the main text. The peak is the highest observed cross-pattern accuracy on the epoch-3 grid. For 4B, 8B, and 32B, its location is $R = 1280$, so this comparison coincides with the fixed high-end contrast. For 14B, the peak occurs at $R = 320$; its 50.21-point peak-to-endpoint loss and 8.82-point final-doubling loss therefore use different starting mixtures. All differences are computed before rounding displayed accuracies. The maxima are grid observations rather than independently validated recipe optima.

Table A3: Observed cross-pattern maxima and accuracy at the largest mixture, at epoch 3. Mix and accuracies are percentages; loss is the peak minus the $R = 2{,}560$ accuracy, in percentage points.

Model	Peak R	Mix (%)	Peak A_U	$R2560$ A_U	Loss (pp)
4B	1280	6.40	55.94	35.94	20.00
8B	1280	6.40	69.24	48.92	20.31
14B	320	1.60	91.35	41.15	50.21
32B	1280	6.40	96.74	77.19	19.55

A.3 EPOCHWISE MAXIMA

Table A4 records the highest cross-pattern accuracy and corresponding mixture separately for each epoch. At fixed $R = 20$, 8B's accuracy changes from 76.35% in epoch 2 to 67.43% in epoch 3, as reported in the main text. The epoch-3 maximum for this model is instead 69.24% at $R = 1280$. These values distinguish the fixed-mixture comparison from the comparison of maxima across the full grid. Columns are successive checkpoints, not independent training runs.

Table A4: Observed cross-pattern grid maxima by epoch. Each entry is R / accuracy (%). Columns describe successive checkpoints, not independent training runs.

Model	Epoch 1	Epoch 2	Epoch 3
4B	160 / 44.31	1280 / 53.12	1280 / 55.94
8B	20 / 69.72	20 / 76.35	1280 / 69.24
14B	160 / 86.91	320 / 92.60	320 / 91.35
32B	1280 / 83.33	1280 / 95.35	1280 / 96.74

A.4 Scaling-fit specifications and numerical records

Response and pooled fit. For $n = 3{,}136$ A–D questions, let n_{wrong} denote the incorrect final-answer count. The log-error fit is

$$\log \widetilde{E}_S = \beta_0 - \alpha_R \log(R/8), \qquad \widetilde{E}_S = \frac{n_{\text{wrong}} + 0.5}{n + 1}. \tag{4}$$

Continuity correction applies only to fitting; the reported accuracy grids are uncorrected. The pooled epoch-3 fit uses 60 cells (four sizes and 15 nonzero mixtures), yielding Equation 2 with log-error $R^2 = 0.898$. The logarithmic predictor excludes $R = 0$. Doubling R multiplies fitted error by approximately $2^{-0.741}$, corresponding to the roughly 40% relative reduction stated in the main text.

Per-model parameters and fitting-range checks. Table A5 reports the parameters underlying the main-text scaling plot (Figure 2). The pooled data exponents at epochs 1, 2, and 3 are 0.387, 0.503, and 0.741. Restricting the epoch-3 fit to $R \geq 80$ gives an exponent of approximately 0.600. These are the training- and range-dependent fits supporting the main-text description of approximate scaling.

Held-out prediction. Holding out an R level removes that entire level across all four sizes; holding out a size removes all 15 mixtures for that size. Each fit uses the remaining cells, and prediction error is evaluated as accuracy RMSE in percentage points. Table A6 lists all reported functional forms. The R-only fit uses only demonstration count; the remaining nonconstant comparisons use both model size and demonstration count.

Adding nominal model size P, measured in billions of parameters, to the log-error fit gives

$$\widetilde{E}_S \approx 0.2144(P/4)^{-0.201}(R/8)^{-0.741}. \tag{5}$$

Its log-error R^2 is 0.909; held-out-size and held-out-R RMSEs are 6.73 and 5.90 points, respectively. The identical data exponent in Equations 2 and 5 follows from orthogonality of $\log P$ and $\log R$ in the balanced grid, rather than constituting an independent replication.

Objective and error-floor checks. Table A7 compares R-only fits using log-error regression and nonlinear least squares on raw error. The zero-floor raw-error fit gives an exponent of approximately 0.989 and held-out-R RMSE of 5.437 points. A free error floor gives RMSE of 5.418 points, with a fitted floor of approximately 0.651% on the full grid and 0.349% on $R \geq 80$. These numerical checks document dependence on the fitting objective and range; they are not estimates of a universal exponent or an established irreducible error floor.

Range-extrapolation check. Fitting the log-error power law on $R \leq 40$ and predicting $R = 80$–2,560 gives an accuracy RMSE of 0.383 percentage points. A fixed 100%-accuracy predictor gives 1.443 points on the same held-out range. This is a retrospective within-experiment check, not a prospectively registered forecast or an evaluation on new tasks.

Table A5: Per-model within-pattern scaling fits at epoch 3. The exponent α_R and goodness of fit are estimated in log-error space.

Model	Data exponent α_R	Log-error R^2
4B	0.690	0.941
8B	0.753	0.920
14B	0.676	0.933
32B	0.843	0.920

Table A6: Held-out prediction of within-pattern accuracy. RMSE is in percentage points. The R-only power law uses R alone; the other nonconstant comparisons use both P and R.

Functional form	Hold out size	Hold out R
R-only power law	6.13	5.86
Joint P/R power law	6.73	5.90
Accuracy linear in log sizes	7.27	7.03
Constant predictor	9.09	9.39
Exponential error decay	9.38	8.86
Accuracy linear in raw sizes	9.99	9.87

Table A7: Comparison of R-only power-law fits at epoch 3. Held-out-R RMSE is in accuracy percentage points. The log-error result is shown to three decimal places here and rounded to 5.86 in the main text.

Power-law fit	Hold out R RMSE
Log-error fit	5.862
Raw-error fit, zero floor	5.437
Raw-error fit, free floor	5.418

A.5 PREREQUISITE-RECALL RECORDS AND STATISTICAL UNITS

Table A8 lists epoch-3 factual recall for the nine low-dose recipes with available checkpoint summaries. Each cell evaluates 3,584 target-map facts and matches the corresponding model–mixture cell in the complete accuracy grids. Independent rule-probe accuracy is 100% throughout this table.

Across 36 checkpoints per epoch in this range, factual recall spans 96.5960%–98.6328% at epoch 1, 99.2188%–99.8884% at epoch 2, and 97.7958%–100% at epoch 3. Exactly 10 of these 108 checkpoints attain perfect factual recall, all at epoch 3. Independent rule accuracy is 100% at 106/108 checkpoints; the exceptions are 4B/$R = 16$/epoch 1 and 8B/$R = 20$/epoch 1, both at 87.5%.

Marginal fact and rule accuracy differs from joint gate coverage: a composite query passes the gate only when every required fact and its rule pass. High-dose E–H gate measurements appear in Table 3 and are analyzed in Section 4.

The tabulated values are individual-checkpoint measurements, not cross-seed averages. The 180 nonzero model–mixture–epoch cells include successive checkpoints from shared training trajectories and do not represent 180 independent training replications. Likewise, paired map intervals condition on fixed predictions and do not estimate between-training-run uncertainty.

Table A8: Factual recall (%) at epoch 3 for $R = 8$–40. Each entry evaluates 3,584 target-map atomic facts. Independent rule-probe accuracy is 100% for every cell in this table. High-dose joint-gate measurements are reported separately in Table 3.

R	4B	8B	14B	32B
8	100.0000	99.9721	100.0000	99.9163
10	99.9163	99.9721	100.0000	99.9721
12	99.9721	99.9721	100.0000	99.9721
14	99.9442	99.8884	100.0000	99.9442
16	99.9721	100.0000	100.0000	99.9721
18	99.9721	99.9442	100.0000	99.9721
20	99.9163	99.8884	99.8326	99.9442
30	100.0000	99.8605	99.9721	100.0000
40	99.9721	99.9163	99.9442	97.7958

B Rule Declarations and Generated Solution Structure

What the process metrics measure. The evaluator distinguishes the direction sequence stated in a Type definition from the directions and entities in subsequent solution steps. The declared-path metric checks whether the extracted direction sequence equals the gold Type path. Step-count exactness checks the number of recognized steps; it does not establish that their directions or entities are correct. A fully faithful output additionally requires the declared path, ordered source–direction–target triples, and final answer all to match the gold solution. Direction and intermediate-entity accuracies use the number of gold hops as their denominator; predicted-edge validity instead uses the number of parsed predicted steps. These metrics assess observable output consistency, not the model's hidden computation.

Bounds on correct rule declarations with incorrect answers. Let $N_U = 2{,}880$ be the number of E–H questions, M_D the number of declared-path failures among all $6{,}016$ composite responses, and C_U the number of correct E–H final answers. Write D for a response whose declared path passes the metric and C for a correct final answer. Since at most M_D declaration failures can belong to E–H,

$$n_U(D) \geq N_U - M_D, \qquad n_U(D \cap \neg C) \geq \max\{0, N_U - M_D - C_U\}. \tag{6}$$

The bounds require no assumption that declaration failures are evenly distributed across types.

For 32B/$R = 30$, the declared-path score is $5960/6016 = 99.0691\%$, while E–H final accuracy is $169/2880 = 5.8681\%$. Equation 6 gives an unseen declared-path pass rate of at least $2824/2880 = 98.0556\%$. Fact and rule probes, and hence the joint gate, are all perfect at this checkpoint. For 4B/$R = 18$, the analogous declared-path lower bound is $2813/2880 = 97.6736\%$, although only one unseen answer is correct. This latter checkpoint has one failed fact probe out of 3,584; it is not a perfect-recall example. Both cases concern low-dose recipes and therefore demonstrate a recall–composition separation without establishing its cause or its specificity to high-dose training.

Reported step-count observations. Table B2 preserves the individual observations discussed in Section 4. Every row refers to 672 target-map questions of one Type at epoch 3. A dash denotes a count not separately tabulated in the available summary; it is not a zero. For 14B/$R = 2560$/G, the reported zero-, three-, and more-than-three-step categories exhaust all 672 outputs.

Parsing and interpretation limits. The scorer recognizes a finite set of step forms, including the training template *"from [source], move [direction] to [target]"*, lookup statements, and specified arrow forms. A semantically complete relation expressed as *"[source] goes north to [target]"* may not be recognized. Comma-joined steps can also be merged by the parser. Consequently, a zero count means that no supported step was extracted, and a two-step count alone does not establish that all third-step content is absent. The reported counts have not been converted into manually adjudicated semantic step counts. They also do not independently distinguish an early stop in generation from a learned preference for shorter solutions. Final-answer scoring and trace matching should therefore be interpreted as separate measurements.

The final-answer scorer examines the last explicit answer claim, or a final-clause fallback, using normalized entity matching. This reduces credit for entities mentioned only in intermediate steps,

Table B1: Conservative E–H process bounds derived from aggregate counts at epoch 3. M_D counts declaration failures across A–H; C_U counts correct answers on E–H. The last two columns are lower bounds, not directly tabulated joint frequencies.

Model / R	M_D	C_U	Fact recall (%)	$n_U(D)$	$n_U(D \cap \neg C)$
4B / 18	67	1	99.9721	2813	2812
32B / 30	56	169	100.0000	2824	2655

Table B2: Recognized solution-step counts for selected unseen three-hop conditions. Counts refer to parser outputs, not internal reasoning depth.

Model	R	Type	Zero	Two	Three	More than three
14B	1280	G	—	662	—	—
14B	2560	G	367	0	23	282
32B	1280	G	—	—	671	—
32B	1280	H	—	—	671	—
32B	2560	G	—	376	—	—

but does not guarantee correct treatment of every negated or multi-candidate answer. The present process observations should not be read as a completed audit of all such scoring ambiguities.

Rule-definition exposure within traces. Each independent rule row contains one Type definition, and every trace adds another definition for its demonstrated Type. At the balanced high-dose points, each A–D Type therefore receives $22 + R/4$ definition statements per epoch, while each E–H Type receives 22. This yields 342 versus 22 at $R = 1280$ and 662 versus 22 at $R = 2560$, ratios of approximately 15.55 and 30.09. These count statements rather than model tokens. They characterize a supervision imbalance, not an independently identified causal variable. The traces also supply atomic relations on demonstration maps, whereas the standalone fact component covers both demonstration and target maps.

Evaluation groups and prospective selection. In this task, the all-two-hop group contains A–F (4,672 items), so it must not be relabeled as unseen two-hop E/F (1,536 items). Unseen-group aggregation uses $A_U = (1536A_{EF} + 1344A_{GH})/2880$. The v5 generator writes the same prerequisite and composite items to its validation and test directories; Stage-1 selection is restricted to knowledge/gate measurements and excludes composite accuracy. These directory names therefore do not provide independent composition-validation and final-test sets. The mixture/checkpoint selection procedure proposed in Section 5 would require separately held-out maps and was not used to obtain the reported grid maxima.