Diversity Drives, Coverage Follows: Decomposing What "More Data" Buys in Compositional Generalization

Ed Henry

2026-05-24

Abstract

What property of a training set causally drives out-of-distribution (OOD) generalization in compositional semantic parsing? On COGS with T5-small (60M parameters), training diversity (the count of unique training examples under random selection) is the dominant causal driver, exposure (forward passes per unique example) modulates floor-escape probability at intermediate diversity, and pair coverage has no detectable independent effect within the tested band: an initial +26.24pp easy-split coverage effect decomposes additively into +30.20pp from a diversity confound, −4.33pp from a typicality penalty introduced by greedy coverage-maximizing selection, and +0.37pp from coverage as an independent variable. Extending the diversity-versus-exposure decomposition to a 12M-parameter transformer trained from scratch on four further benchmarks (SCAN, ReCOGS, gSCAN-compositional, CFQ-mcd1), the qualitative finding "diversity drives, exposure refines" replicates on three of four datasets; CFQ-mcd1 is a flat-floor boundary case at this configuration, with no detectable dose-response across a 10× scaling in N_{unique} (bootstrap slope CI $[-2 \times 10^{-5}, -1 \times 10^{-5}]$). We treat this as conditional on both model capacity and on CFQ's non-canonical surface measure of compositional structure, not as a pure capacity diagnosis. Cross-dataset *magnitude* comparisons turn out to be fragile. The per-dataset measures used in compositional-generalization research (`template_id`, `extract_pairs`, `compute_depth`) measure operationally different constructs across datasets, and naive cross-dataset effect-size comparisons inherit the choice of bucketization, the choice of effect-size statistic, and ~69% of their spread from baseline arithmetic. We propose a four-step audit framework (per-adapter measure audit; unified surface signature; bucket-boundary robustness sweep; baseline-arithmetic null calibration), apply it to our own results, and find that the qualitative "diversity drives generalization in learnable regimes" finding survives every robustness check, but quantitative cross-dataset learnability rankings should not be inferred at the five-seed-per-cell level common in the field. The audit framework is the methodological contribution; the COGS decomposition and the cross-dataset qualitative classification are the empirical contributions.

1 Introduction

Compositional generalization, the ability to combine familiar parts in novel arrangements, remains a persistent failure mode of neural sequence models (Lake and Baroni 2018; Kim and Linzen 2020; Lake et al. 2017; Fodor and Pylyshyn 1988). Benchmarks including SCAN (Lake and Baroni 2018), COGS (Kim and Linzen 2020), CFQ (Keysers et al. 2020), gSCAN (Ruis et al. 2020), and ReCOGS (Wu, Manning, and Potts 2023) document this failure across surface forms, semantic structures, and grounded settings. Practitioners building training sets for these tasks face a tactical question: at a fixed compute budget, what should they do? Should they engineer training sets for high *pair coverage* (the fraction of attested compositional pairs in test seen at training)? For high *diversity* (the count of unique training examples)? For high *exposure* (forward passes per unique example)? Each prescription implies a different intervention.

This work answers three questions, each at a different scope.

Q1 (within-dataset, causal). On a single benchmark with a single model, which of these three properties is the causal driver of OOD performance? On COGS with T5-small (60M parameters), we show that diversity is the causal driver, exposure modulates floor-escape probability at intermediate diversity, and pair coverage has no detectable independent effect. An initial 2×2 factorial appears to show a 26.24pp easy-split coverage advantage, exactly the result the pair-coverage hypothesis predicts; a six-experiment confound-elimination arc decomposes that apparent effect additively into +30.20pp from a diversity confound, −4.33pp from a typicality penalty introduced by greedy coverage-maximizing selection, and +0.37pp from coverage as an independent variable. Equivalence testing at high N supports a coverage null within ±3 percentage points.

Q2 (cross-dataset, qualitative). Does the qualitative finding "diversity drives, exposure refines" replicate beyond COGS? We extend the diversity-versus-exposure factorial to a 12M-parameter transformer on SCAN length-split, ReCOGS gen, gSCAN-compositional adverb_1, and CFQ-mcd1. Three of four datasets show the same qualitative pattern, with diversity dominating exposure as a predictor of OOD performance. The fourth, CFQ-mcd1, shows a *flat-floor* signature at the 12M-parameter configuration: OOD test accuracy is approximately 10.6% with standard deviation under 0.07 percentage points, across a 10× scaling in N_{unique} (bootstrap slope CI $[-2\times10^{-5}, -1\times10^{-5}]$). The diversity dose is delivered to the model in unified-signature terms ($96 \rightarrow 151$ signatures across the sweep), but at this model scale and under CFQ's adapter-specific surface-pair measure no measurable conversion into generalization is observed. We deliberately do not read this as a pure capacity diagnosis; it is jointly conditional on model size and on the non-canonical surface measure that CFQ's adapter currently uses for compositional structure.

Q3 (cross-dataset, quantitative). Can we read learnability rankings off cross-dataset effect-size comparisons? We argue no, and the argument is the methodological contribution of the paper. The per-dataset measures used in compositional-generalization research (`template_id`, `extract_pairs`, `compute_depth`) measure operationally different constructs across datasets: bindings in COGS/ReCOGS, parse-tree structure in SCAN, Cartesian products in gSCAN, surface co-occurrence in CFQ. A naive comparison of per-dataset Cohen's d inherits these heterogeneous constructs; a comparison of bucketed "closed headroom" metrics inherits the bucketization choice; the specific rank ordering of three positive-slope datasets is preserved in only 34% of 44 alternative bucketization schemes. About 69% of the observed cross-dataset spread in our preferred metric is attributable to baseline arithmetic rather than signal. At the five-seed-per-cell level common in the field, no pair of datasets in our sweep is pairwise distinguishable by permutation test.

We turn these observations into a four-step audit framework: a per-adapter measure audit, a unified surface signature that re-bucketizes each dataset into a common 4-dimensional grid, a bucket-boundary robustness sweep, and a baseline-arithmetic null calibration. Applied to our own results, the framework cleanly separates the robust claims (diversity drives, exposure refines, CFQ is a flat-floor boundary case) from the fragile ones (specific magnitude rankings, specific slope values), and prescribes which kinds of cross-dataset claims should be made with what level of confidence.

1.1 Contributions

1. **A within-dataset causal decomposition of pair coverage.** On COGS with T5-small, we causally decompose an apparent +26.24pp pair coverage effect into a +30.20pp diversity confound, a −4.33pp typicality penalty, and a +0.37pp residual that is statistically equivalent to zero under TOST at ±3pp. Pair coverage is established as a *byproduct* of training diversity, not an independent mechanism.

2. **A cross-dataset external validity test of the qualitative diversity-versus-exposure dichotomy.** On a 12M-parameter transformer, three of four further benchmarks replicate the qualitative pattern (diversity-dominant, exposure as a refinement), and the fourth (CFQ-mcd1) is identified as a flat-floor boundary case (zero observed slope across a 10× diversity scaling, robust across bucketization choices) jointly conditional on model capacity and on CFQ's non-canonical surface-pair measure.

3. **A construct-validity gap in cross-dataset compositional generalization research.** The per-dataset measures of "diversity" used in the literature are operationally heterogeneous; naive cross-dataset effect-size comparisons are fragile to the choice of bucketization, the choice of effect-size statistic, and the number of seeds per cell.

4. **A four-step audit framework.** Per-adapter measure audit; unified surface signature; bucket-boundary robustness sweep; baseline-arithmetic null calibration. Applied to our own results, the framework recovers a clean separation between robust qualitative findings (which transfer cross-dataset) and fragile quantitative rankings (which do not).

Figure 1 summarizes the within-dataset COGS decomposition, the foundation for the broader cross-dataset analysis.

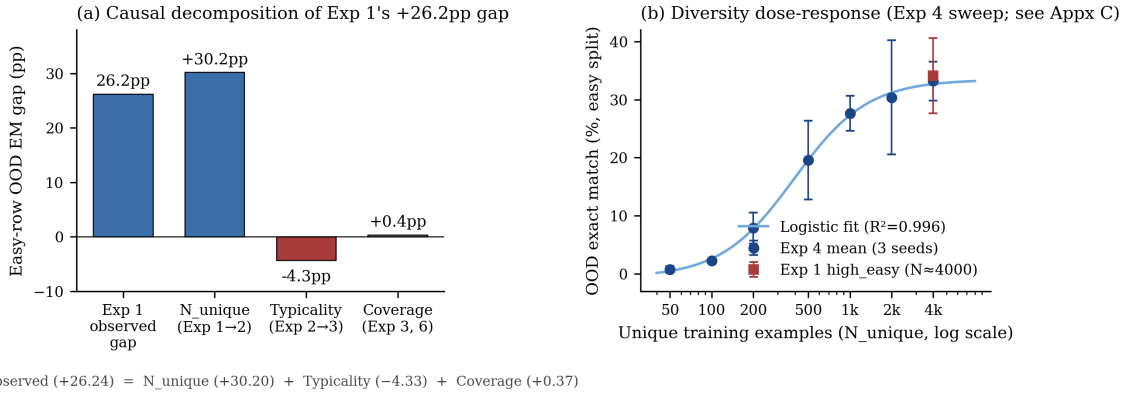

Figure 1: Within-dataset decomposition of an apparent pair-coverage effect on COGS with T5-small. (a) The initial 2×2 factorial's +26.24pp easy-split coverage advantage decomposes additively, once diversity and structural typicality are controlled, into +30.20pp from a diversity confound, −4.33pp from a typicality penalty introduced by greedy coverage-maximizing selection, and +0.37pp from coverage as an independent variable. (b) Diversity dose-response from the Experiment 4 pre-bugfix N_{unique} sweep (see Appendix C); post-bugfix anchors validate the high-N points within 1pp at $N \in \{1000, 4000\}$. The initial factorial's high-coverage easy cell is overlaid as a red square. A logistic fit on $\log_{10} N_{\text{unique}}$ explains 99.6% of the variance.

The remainder of the paper is structured as follows. Section Section 2 situates the work in compositional-generalization benchmarks, cross-benchmark concurrence, LLM construct validity, and data-diversity measurement. Section Section 3 defines the task, model, and notation. Section Section 4 presents the within-dataset COGS decomposition (Q1). Section Section 5 presents the cross-dataset qualitative replication (Q2). Section Section 6 presents the construct-validity gap and the four-step audit framework (Q3). Section Section 7 discusses implications for dataset construction and SFT data curation. Section Section 8 collects the limitations. Section Section 9 concludes.

2 Related work

Compositional generalization benchmarks. SCAN (Lake and Baroni 2018) showed near-zero generalization on novel length splits despite high i.i.d. accuracy. COGS (Kim and Linzen 2020) extends the failure mode to semantic parsing with richer compositional structure. CFQ (Keysers et al. 2020) introduces a principled compound divergence measure that supports graded OOD evaluation, and gSCAN (Ruis et al. 2020) situates compositional generalization in a grounded instruction- following setting. ReCOGS (Wu, Manning, and Potts 2023) re-annotates COGS with a cleaner logical-form grammar that removes incidental memorization artifacts. These benchmarks have documented the failure mode in depth. They have not directly addressed the data-side question of what training properties enable generalization without specialized architectural bias (Hupkes et al. 2020).

Cross-benchmark concurrence and validity. Sun, Williams, and Hupkes (2023) examine the extent to which different compositional-generalization benchmarks measure the same underlying capability and find limited concurrence. Their result motivates caution about cross-dataset effect-size comparisons. Our construct-validity audit (Section Section 6) is an operational mechanism for that caution: it gives a procedure for distinguishing robust qualitative cross-dataset claims from fragile quantitative ones.

LLM construct validity. A growing line of work flags construct-validity problems in LLM evaluation (Biderman et al. 2025). Our concern is narrower: not whether benchmarks measure the *generalization capacity* construct, but whether the per-dataset *diversity* measures used as predictors in our own (and others') causal analyses measure a common construct. The audit framework we propose is dataset-construction-focused rather than capability-focused.

Training-data composition and data diversity. Training-data composition is known to matter substantially for downstream generalization in pretraining and instruction tuning (Longpre et al. 2023; UBC NLP Group 2025). Recent measurement-focused work proposes data-diversity metrics for instruction tuning (Nguyen and Ploeger 2025; Fudan NLP Lab 2025). Our contribution to this literature is a *unified surface signature* (Section Section 6.2) that, in the limited domain of compositional generalization benchmarks, gives a cross-dataset comparable diversity measure with quantified robustness properties. The unified signature is a practical mechanism that recovers the qualitative finding of these prior works on our four-benchmark slice while making explicit the fragility of the quantitative rankings.

SFT data curation. A practitioner-side literature on supervised fine-tuning has produced both "less is more" results (Zhou et al. 2023; Chen et al. 2023) and quality-focused diversity metrics (Liu et al. 2024). The empirical fact our work contributes is the diversity-versus-exposure decomposition: at a fixed model scale, "more data" buys generalization through the distinct-example-count axis rather than the repeat-pass axis. Whether this decomposition explains the apparent disagreements between SFT-curation papers is a separate empirical question that requires the same diversity and exposure measurements be taken in those settings; we offer the decomposition as one lens (Section Section 7), not as a resolution of the SFT debate.

Confound isolation in ML experiments. Our within-dataset methodology is related to arguments for more rigorous causal inference in ML experiments (D'Amour et al. 2020; Lipton and Steinhardt 2019). The specific challenge we address in Section Section 4 is that two experimental factors (coverage and unique count) are exactly collinear under naturalistic data collection; the cross-dataset challenge in Section Section 6 is the higher-order one of comparing causal effect sizes across studies that measure operationally different constructs.

3 Setup and notation

3.1 Task family

The five benchmarks studied here all share a sequence-to-sequence form: a natural-language or symbolic input is mapped to a structured output, and an OOD test set is constructed to expose compositional failures. We summarize each briefly.

- **COGS** (Kim and Linzen 2020) maps English sentences to typed logical forms. Example: *"A cat saw the boy"* → `cat(x_1) ; *boy(x_4) ; see.agent(x_2, x_1) AND see.theme(x_2, x_4)`. The structured generalization split tests novel compositional depth, novel predicate-argument structures, and familiar atoms in novel combinations.
- **ReCOGS** (Wu, Manning, and Potts 2023) re-annotates COGS with a cleaner LF grammar that removes incidental memorization artifacts. The evaluation split is denoted *gen*.
- **SCAN** (Lake and Baroni 2018) maps English-like commands to action sequences. The length-split tests OOD generalization to longer action sequences than seen at training.
- **gSCAN** (Ruis et al. 2020) situates SCAN in a grounded grid world with adverbs. We use the *adverb_1* compositional split.
- **CFQ** (Keysers et al. 2020) maps natural-language questions to SPARQL queries over a knowledge base. We use the *mcd1* split (one of three maximum-compound-divergence splits).

All OOD evaluation is on the appropriate generalization split. Primary metrics are *exact match* (EM, every output token correct) and *target accuracy* (TA, the fraction of canonical output content tokens correct, more granular than EM and used where EM saturates at or near zero).

3.2 Models

We use two model configurations. Section Section 4 uses T5-small (Raffel et al. 2020) (60M parameters) on COGS, fine-tuned from a pretrained T5 checkpoint with teacher-forced cross-entropy and AdamW. Section Section 5 uses a 12M-parameter encoder-decoder transformer (architecture matched across benchmarks) on SCAN, ReCOGS, gSCAN-comp, and CFQ, trained from scratch (no pretrained initialization is loaded) with the same teacher-forced cross-entropy and AdamW setup. Seeds are drawn from $\{42, 123, 456, 789, 1024\}$ unless otherwise noted; section Section 4 uses 3-seed or 5-seed cells depending on experiment, and section Section 5 uses 5 seeds per cell.

3.3 Notation and definitions

Let $T = (e_1, \ldots, e_n)$ be the (possibly repeating) training manifest and $U(T)$ its underlying set of distinct examples.

Training diversity:

$$N_{\text{unique}} \equiv |U(T)|.$$

Budget and exposure. With $B = |T|$ the total number of seen examples (counting repetition), *exposure* is

$$E \equiv B/N_{\text{unique}},$$

reported in forward passes per unique example (fp/unique). N_{unique} and E are independently controllable: holding the manifest fixed and varying the number of optimizer updates changes E without changing N_{unique}, and vice versa.

Compositional pair. A pair (a, b) is a co-occurrence of two primitives within a single training example at a specified analysis layer. The atomic primitives differ by dataset (Section Section 6 makes this explicit) but the formal definition is shared.

Pair coverage ratio. For a training manifest T and a test distribution D,

$$C(T) \equiv \frac{\left|\{(a,b) : (a,b) \in e \in U(T)\}\right|}{\left|\{(a,b) : (a,b) \in e \in D\}\right|}.$$

$C(T)$ is stored as analysis-only metadata; the model never observes it. By construction $C(T) \in [0,1]$, and $C(T)$ increases monotonically with N_{unique} under random sampling.

Structural typicality. The typicality of T is the closeness of its distribution over example difficulty to the pool distribution. We operationalize typicality by two summary statistics: mean compositional depth $\bar{d}(T)$ and mean target-sequence length $\bar{L}(T)$. T is *typical at tolerance* $(\delta_d, \delta_L) = (0.05, 5 \text{ chars})$ when both statistics lie within those bounds of the pool medians.

Dose-response. The functional relationship between a graded input (here N_{unique}) and an outcome (here OOD EM or TA) with covariates held constant.

TOST equivalence test. The two one-sided tests procedure (Lakens 2017) establishes statistical equivalence to zero within a prespecified margin $\pm\Delta$; we use $\Delta = 3$pp where explicitly noted.

Closed headroom. For two manifests at low and high diversity producing OOD test accuracies TA_{low} and TA_{high}, the *closed headroom* is

$$\text{HC} \equiv \frac{\text{TA}_{\text{high}} - \text{TA}_{\text{low}}}{1 - \text{TA}_{\text{low}}}.$$

HC normalizes the absolute gain by the available room for improvement, removing first-order dependence on the noise floor that distorts cross-dataset Cohen's d comparisons (Cohen 1988) (Section Section 6.4).

4 Causal decomposition on COGS (Q1)

This section establishes the within-dataset causal decomposition that forms the foundation for the broader cross-dataset analysis. The chronology of six experiments is the natural way to present the evidence: each successive design controls a variable the previous result had failed to isolate.

4.1 E1: the apparent coverage effect

A 2×2 factorial of Coverage (high / low) × Difficulty (easy / hard) at fixed budget 4000 examples. High-coverage cells use greedy selection over 4000 unique examples; low-coverage cells use 200 unique examples repeated 20×. Three seeds per cell.

On the easy split, high-coverage outperforms low-coverage by +26.2pp (34.1% vs. 7.9%; partial $\eta^2 = 0.978$, $F(1,8) = 347$, $p < 10^{-4}$). The result is consistent in direction and magnitude across all three seeds. By the registered criterion, the coverage hypothesis is confirmed. Pre-execution review of the design identified two confounds: high cells have 20× more unique examples than low cells, and greedy coverage maximization preferentially picks structurally atypical examples.

4.2 E2: equalizing N_{unique} reverses the effect

The same 2×2 factorial, with $N_{\text{unique}} = 200$ for all cells. High-coverage cells use greedy-selected 200 examples (97.2% coverage at easy); low-coverage cells use random-selected 200 examples (44% coverage at easy). All cells use budget 4000, repetition rate 20×. Three seeds per cell.

The high cells collapse from 34.1% to 3.9% on easy; the low cells replicate seed-for-seed (7.9% in both Experiment 1 and Experiment 2). The coverage contrast is now −4.0pp, significant

in the opposite direction ($F(1,8) = 27.57$, $p = 0.0008$, partial $\eta^2 = 0.775$). At fixed N_{unique}, greedy coverage maximization hurts: it spends the repetition budget consolidating structurally unusual examples. The seed-level exact replication of the low cells confirms no other pipeline variable changed between Experiment 1 and Experiment 2.

4.3 E3: typicality-controlled coverage at $N = 200$

All cells use random selection, removing the typicality penalty from greedy. Coverage is varied by pre-screening 1000 random seeds and retaining top-5% and bottom-5% by pair coverage, *conditional on* typicality (mean depth within ± 0.05, mean target length within ± 5 chars of the pool median). The procedure achieves a 14.9pp natural coverage gap (53.7% vs. 38.8%) while holding mean depth within 0.01 units.

The coverage contrast at easy is +0.37pp ($F(1,8) = 0.001$, $p = 0.97$, partial $\eta^2 = 2 \times 10^{-4}$). The coverage null is declared at $N_{\text{unique}} = 200$. Combining Experiments 1, 2, and 3 yields the additive decomposition of Table 1.

Table 1: Additive decomposition of Experiment 1's apparent easy-split coverage effect; Experiment 6 independently bounds the coverage contrast at high N_{unique} (the $N \in \{1000, 2000\}$ contrasts are both within 0.4 percentage points in absolute value and pass TOST at a 3 percentage point margin), and the $N = 200$ Experiment 3 contrast of +0.37pp is separately ANOVA-null at $p = 0.97$

Component	Source	Easy contribution
N_{unique} confound (200 → 4000)	E1 → E2 swing	+30.20pp
Typicality (greedy → random at $N = 200$)	E2 → E3 restoration	−4.33pp
Coverage (typicality controlled)	E3 direct	+0.37pp
Sum		+26.24**pp E1 observed**

4.4 E4: the diversity dose-response

A 7-level N_{unique} sweep ($\{50, 100, 200, 500, 1000, 2000, 4000\}$) with budget fixed at 4000. Easy EM increases monotonically from 0.71% (N=50) to 33.18% (N=4000), Spearman $\rho = 1.0$. A logistic in $\log_{10} N_{\text{unique}}$ fits with $R^2 = 0.996$, RMSE 0.90pp, ceiling 34.6%, inflection near $N \approx 400$. The $N = 4000$ result is statistically indistinguishable from Experiment 1's high-coverage easy cell (−0.92pp, 95% CI $[-4.27, +2.43]$). Post-bugfix anchor cells at $N \in \{1000, 4000\}$ reproduce the high-N Experiment 1 advantage under random selection (within 1pp of the corresponding pre-bugfix sweep points); the dose-response shape from the pre-bugfix sweep carries through but should be read with the Appendix C caveat, since the original Experiment 4 manifests were generated with the greedy strategy due to a pipeline default and the $N = 200$ point shifts by −2.44pp on the post-bugfix re-run.

4.5 E5: the diversity × exposure factorial

A fully crossed 6×4 on easy ($N_{\text{unique}} \in \{100, 200, 500, 1000, 2000, 4000\} \times E \in \{100, 400, 1600, 6400\}$) plus a 4×3 hard companion grid. Manifest size equals N_{unique} (no in-manifest repetition); max-updates is varied independently to set E. 5 seeds per cell: 120 easy runs plus 60 hard, 180 total.

ANOVA gives N_{unique} partial $\eta^2 = 0.882$, exposure $\eta^2 = 0.196$, interaction $\eta^2 = 0.263$. N_{unique} is the dominant axis: no level of exposure rescues $N \leq 200$, and at $N = 4000$ a 64× ex-

posure increase shifts mean EM by less than 1pp. Figure 2 shows the mean-EM grid and the matched escape-probability matrix from a Gaussian-mixture audit. At the transition zone ($N \in [500, 1000]$, $E \leq 400$) several cells exhibit within-cell standard deviation above 5pp; a 2-component GMM with BIC model comparison gives positive ΔBIC (with $\Delta\text{BIC} > 6$ treated as strong evidence), classifying them as mixtures of a floor mode and an escaped mode. A logistic for $P(\text{escape})$ on $\log N$ and $\log E$ gives coefficients $+3.73$ and $+1.06$; the 50% escape boundary moves from $N \approx 885$ at $E = 100$ to $N \approx 270$ at $E = 6400$. Escaped seeds converge *later* than stuck seeds (mean step 3000 vs. 1625 at $N = 1000$, $E = 400$), not faster: exposure raises the probability that a given initialization finds a productive trajectory, not the capacity it can reach.

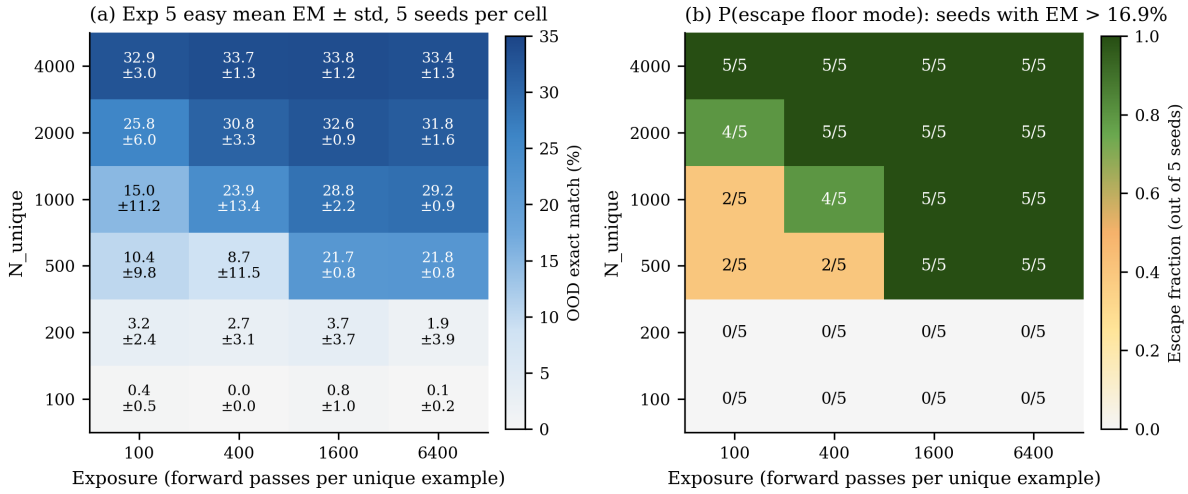

Figure 2: Experiment 5 diversity × exposure factorial (easy split). (a) Mean EM with within-cell standard deviation, 5 seeds per cell; rows are N_{unique}, columns are exposure. (b) Escape probability: fraction of seeds exceeding the empirically estimated escape threshold (16.9% EM).

4.6 E6: closing the arc, coverage at high N

A 2×4 factorial: $N_{\text{unique}} \in \{1000, 2000\}$ × coverage quantile $\in \{q_1, q_{25}, q_{75}, q_{99}\}$. For each cell, 10,000 random seeds are pre-screened; the 5 seeds whose coverage is closest to the target quantile *and* which satisfy the typicality constraints are retained. Coverage manipulation: 11.7pp gap at $N = 1000$, 10.3pp at $N = 2000$.

Primary contrasts: q_{99} vs. q_1 is -0.39pp at $N = 1000$ ($t(8) = -0.61$, $p = 0.56$) and -0.07pp at $N = 2000$ ($t(6) = -0.07$, $p = 0.94$). TOST equivalence to zero at ± 3pp is supported at both levels; 90% CIs are $[-1.58, +0.80]$pp and $[-2.01, +1.87]$pp. The two-way ANOVA gives coverage-quantile $F(3, 32) = 0.14$, $p = 0.93$, partial $\eta^2 = 0.013$; the N main effect $F(1, 32) = 24.23$, $p < 10^{-3}$, partial $\eta^2 = 0.431$. Diversity matters; coverage, controlled for typicality, does not.

4.7 Findings (Q1)

Figure 3 shows the causal structure consistent with these findings. Diversity is a direct cause of OOD generalization on COGS at this model scale. Coverage is a correlate of diversity under natural sampling but does not act on OOD performance independently within the tested band. Typicality is a downstream property of the selection strategy and becomes a confound on coverage contrasts when the strategy is varied at fixed N_{unique}.

Thesis (Q1). *Training diversity, operationalized as N_{unique} under random selection with exposure above the empirical floor ($E \geq 1600$ for $N_{unique} \in [500, 1000]$, $E \geq 400$ at $N_{unique} = 2000$,*

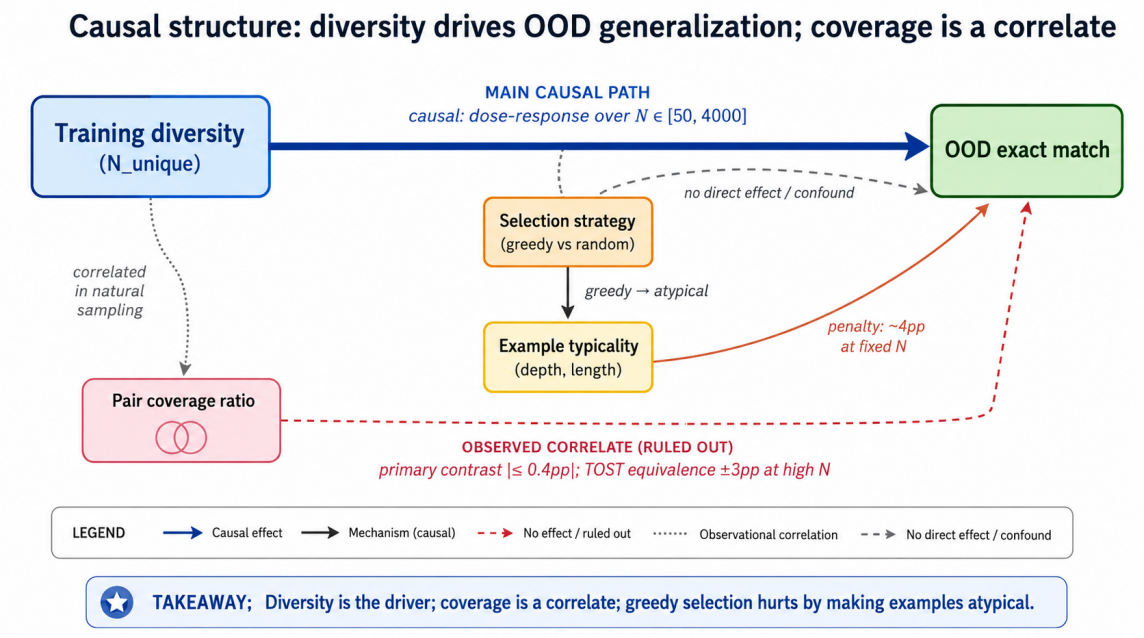

Figure 3: Causal structure consistent with Experiments 1 through 6 on COGS. Solid blue: the causal effect we measure (diversity → OOD). Dashed red: candidate pathways for which we find no detectable effect, or that act as confounds when present. Dotted grey: observational correlation between diversity and coverage under natural sampling.

$E \geq 100$ at $N_{unique} = 4000$), is the primary causal driver of OOD compositional generalization on COGS with T5-small. Pair coverage ratio, as a standalone variable, has no detectable independent causal effect within the band $[38.8\%, 88\%]$ across $N_{unique} \in \{200, 1000, 2000\}$.

5 Cross-dataset external validity (Q2)

This section extends the diversity-versus-exposure decomposition to a 12M-parameter transformer on four further compositional benchmarks. The goal is qualitative: does "diversity drives, exposure refines" replicate? Quantitative cross-dataset effect-size comparisons are deferred to Section Section 6, where we show they are fragile.

5.1 Setup

For each of SCAN length-split, ReCOGS gen, gSCAN-compositional adverb_1, and CFQ-mcd1, we run a diversity × exposure factorial at the 12M-parameter scale, 5 seeds per cell. SCAN, ReCOGS, and gSCAN-comp use a 3×3 design with $N_{\text{unique}} \in \{500, 1000, 2000\}$ and $E \in \{400, 1600, 6400\}$, giving 9 cells × 5 seeds = 45 runs each. CFQ-mcd1 is run as a scaling sweep: a 4-cell single-row design with $N_{\text{unique}} \in \{2000, 5000, 10000, 20000\}$ at $E = 400$, giving 4 × 5 = 20 runs. The CFQ design is restricted because the EM metric saturates at zero at the 12M scale (see Section Section 5.5).

ReCOGS has 43 of 45 runs completed at write time (two seeds of the N=2000, E=6400 cell in flight); the other datasets are complete.

5.2 SCAN length-split

EM ranges from 6.6% to 10.2% across the 9 easy-cell mean values. The diversity Cohen's d between the N=500 and N=2000 endpoint cells is large and positive across all exposure levels:

$d = 3.95$ at $E = 400$, $d = 2.84$ at $E = 1600$, $d = 4.08$ at $E = 6400$. The exposure Cohen's d between the E=400 and E=6400 endpoint cells is small to moderate: maximum $d = 0.98$ at $N = 500$, vanishing at $N \geq 1000$. The qualitative finding "diversity dominant, exposure as a refinement" holds; we register this as Pattern P (full qualitative replication of the diversity-versus-exposure finding) for SCAN.

A construct-validity caveat that will matter in Section Section 6: SCAN's `compute_depth` is parse-tree depth, not the variable count that the same name measures in COGS. The Cohen's d values above are not directly comparable to COGS or ReCOGS values without adjustment.

5.3 ReCOGS gen

EM is at the metric floor (0% across all 43 completed runs) because the 12M model never produces a fully-correct logical form at this scale on the ReCOGS gen split. TA, the more granular metric, is non-trivial: 51.81% to 52.12% across the OOD partitions at the N=2000, E=400 cell, with far_OOD iid (indicating genuine generalization rather than overfitting).

Diversity Cohen's d on TA is enormous: $d = 20.67$ at $E = 400$, $d = 24.80$ at $E = 1600$, $d = 16.46$ at $E = 6400$ between the N=500 and N=2000 endpoint cells. Exposure Cohen's d on TA is essentially zero and slightly negative: -0.71 to $+0.18$ across N levels. Pattern P.

The very large diversity d values reflect the small within-cell standard deviation of TA at this scale, not a substantively larger underlying effect than SCAN's. This non-portability of Cohen's d across datasets is exactly the construct-validity issue Section Section 6 addresses.

5.4 gSCAN-compositional adverb_1

EM is again low; TA is the primary metric. Endpoint cell TA values range from 17–20% at N=500 to 22–30% at N=2000, with a non-monotonic dip at $N = 1000$ (TA 5–15%, standard deviation up to 0.14). The endpoint cells also carry substantial seed variance (standard deviation up to 0.105 at N=500 and 0.074 at N=2000 in the canonical report). The dip at $N = 1000$ is the most acute case: five seeds is insufficient to average over the bimodal distribution there, and a single bimodal seed dominates the cell mean. Endpoint means remain positive across exposures, and the N=500 $\rightarrow$ N=2000 comparison is positive at every exposure level, but the magnitudes carry seed-variance caveats.

Diversity Cohen's d on TA between endpoint cells: $d = 0.55$ at $E = 400$, $d = 1.53$ at $E = 1600$, $d = 0.62$ at $E = 6400$. Moderate but positive. Per-adapter `template_id` saturates at 32 total templates across all N values, which would naively suggest no diversity dose is being delivered; in fact the unified surface signature (Section Section 6.2) registers a +4.2-signature increase from $N = 500$ to $N = 2000$, confirming that the diversity dose IS being administered but the per-adapter measure is too coarse to detect it. Pattern P with the caveat that the high within-cell variance at $N = 1000$ would benefit from additional seeds before the magnitude claim is treated as definitive.

5.5 CFQ-mcd1 scaling sweep

EM is zero across all four cells (0.0000 ± 0.0000); the 12M model emits boilerplate SPARQL prefixes without producing complete correct queries. TA across the four cells:

Table 2: CFQ-mcd1 scaling sweep, 5 seeds per cell. TA values are essentially identical across a 10× scaling in N_{unique}

N_{unique}	TA
2000	0.1066 ± 0.0007
5000	0.1060 ± 0.0003
10000	0.1052 ± 0.0002
20000	0.1057 ± 0.0002

The bootstrap slope CI for TA against $\log N_{\text{unique}}$ is $[-2 \times 10^{-5}, -1 \times 10^{-5}]$, width 10^{-5}, numerically negligible at the scale of TA (the CI does not include zero, but its absolute magnitude is four orders of magnitude below the positive-slope datasets' CIs). The unified-signature dose administered across this sweep is +55 signatures (96 → 151, the largest unified-signature dose of any dataset in the sweep). The dose is delivered; no metric movement results.

We register this as Pattern P-negative for CFQ-mcd1 at the 12M-parameter configuration. The flat line is 100% stable across all bucket-boundary schemes we test in Section Section 6.3; the *qualitative* finding "no dose-response detected" is robust to the bucketization choices that sweep examines. The interpretation is jointly conditional on (i) the 12M-parameter model size and (ii) CFQ's non-canonical surface-pair measure: a larger model might recover a slope without changing the data, and a canonical rule-trace measure of CFQ atoms and pairs might recover a slope without changing the model. We do not treat this result as a pure capacity diagnosis. What travels for practitioners is the empirical signature: zero observed slope across a wide diversity scaling under the current model and measure.

5.6 Cross-dataset summary

Table 3: Cross-dataset summary of Q2 pattern outcomes. The diversity-versus-exposure qualitative pattern replicates on the three learnable further datasets (SCAN, ReCOGS, gSCAN-comp); CFQ-mcd1 at the 12M-parameter configuration is a flat-floor boundary case, jointly conditional on model capacity and on CFQ's non-canonical surface-pair measure

Dataset	Design	Pattern outcome (Q2)	Primary metric	Diversity d at endpoint cells	Exposure d at endpoint cells
COGS (Section Section 4)	6×4 + 2×4	P (diversity dominant)	EM	$\eta^2 =$ 0.882 in ANOVA	$\eta^2 =$ 0.196
SCAN length-split	3×3	P (diversity dominant)	EM	2.84 to 4.08	0.14 to 0.98
ReCOGS gen	3×3	P (diversity dominant on TA)	TA (EM floor)	16.46 to 24.80	−0.71 to +0.18
gSCAN-comp adverb_1	3×3	P (diversity dominant, $N = 1000$ dip)	TA (EM low)	0.55 to 1.53	−0.47 to +0.46
CFQ-mcd1	4-cell sweep	P-negative (flat floor at 12M, this measure)	TA (EM zero)	bootstrap slope CI $[-2\times10^{-5}, -1\times10^{-5}]$	not varied

Thesis (Q2). *The qualitative diversity-versus-exposure decomposition replicates on three of*

four further compositional benchmarks at the 12M-parameter from-scratch configuration; CFQ-mcd1 shows a flat-floor signature with zero detectable dose-response across a 10× scaling in N_{unique}. The flat-floor classification is bucket-robust, but the underlying mechanism is jointly conditional on model capacity and on CFQ's non-canonical surface-pair measure.

Figure 4 collects the per-dataset dose-response curves on a common axis. The axis is the count of *unique surface signatures* present in the training manifest. A surface signature is a bucketized 4-tuple over the atom count, pair count, compositional depth, and output length of each training example, intentionally abstracting away each dataset's idiosyncratic per-adapter constructs; the construction and bucket boundaries are defined in Section Section 6.2.

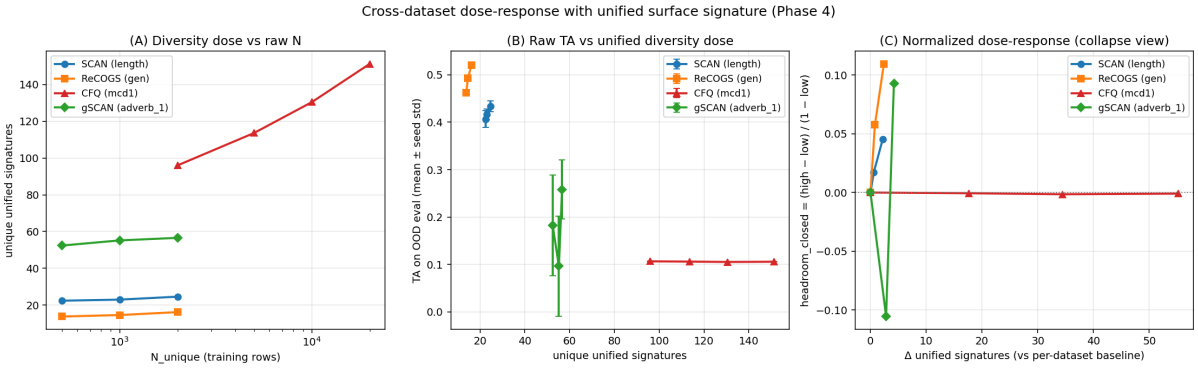

Figure 4: Per-dataset OOD dose-response on the unified-signature axis (count of unique surface signatures in the training manifest; defined in Section Section 6.2). SCAN, ReCOGS, and gSCAN-comp show positive slopes; CFQ-mcd1 is flat at TA 0.106 across a 10× scaling. The vertical axis is target accuracy (TA), the cross-dataset comparable granular metric. The specific magnitudes of the three positive slopes are not directly comparable across datasets; that fragility is the subject of Section Section 6. The qualitative classification (positive slope vs. flat floor) is robust to all bucketization choices we test.

6 Construct-validity audit (Q3)

The cross-dataset results of Section Section 5 admit a natural follow-up: can we read learnability rankings off these effect sizes? Is SCAN harder than ReCOGS? Is gSCAN-comp harder than SCAN? At what slope does each dataset gain TA from each additional unit of unified-signature dose?

This section makes the case that, at the five-seed-per-cell level common in the compositional-generalization literature, such quantitative cross-dataset claims are not defensible. The argument is structured as the four-step audit framework we propose for any such cross-dataset analysis: (i) per-adapter measure audit; (ii) unified surface signature; (iii) bucket-boundary robustness sweep; (iv) baseline-arithmetic null calibration.

6.1 Step 1: per-adapter measure audit

Each benchmark in our sweep ships with a per-adapter `extract_atoms`, `extract_pairs`, and `compute_depth` function that the analysis pipeline calls without modification. Inspecting these functions reveals that they measure operationally different constructs.

- **COGS / ReCOGS.** `extract_pairs` returns LF-variable co-occurrence pairs: (cat, see.agent), (boy, see.theme), etc. The construct is *bindings*: which entities participate in which thematic role of which event.

- **SCAN.** `extract_pairs` returns token-pair compositions from the input command and action sequence: (walk, LOOK), (twice, WALK WALK). The construct is *parse-tree composition / structural*.
- **gSCAN.** `extract_pairs` returns Cartesian-product co-occurrences over the coordinate × adjective × verb space: ((2, 3), red square), (push cautiously, walk verb). The construct is *grounded Cartesian product*.
- **CFQ.** `extract_pairs` returns SPARQL-token co-occurrences: (?x, a.actor), (ns:music, ns:track). The construct is *surface co-occurrence* of SPARQL tokens, not the semantic binding equivalent.

A 2-row table of "unique pair count" across datasets at fixed N_{unique} silently merges these four constructs. The construct names happen to be the same; the operational definitions are not.

The most acute symptom is the *cardinality spread* per adapter. At $N_{\text{unique}} = 2000$, the number of unique per-adapter-`template_id` values ranges from 32 (gSCAN, saturated) to 1,927 (CFQ pre-fix), a 60× spread. Most of this spread reflects the differing granularity of the per-adapter `template_id` rather than a real difference in compositional richness. Per-adapter diversity measures as comparators across these benchmarks are *not on the same scale*.

6.2 Step 2: the unified surface signature

To put the benchmarks on a common scale, we define a *unified surface signature*: a four-dimensional bucketized descriptor of each training example,

$$\text{sig}(e) \;=\; (b_a(\#\text{atoms}(e)),\; b_p(\#\text{pairs}(e)),\; b_d(\text{depth}(e)),\; b_L(\text{len}(e))),$$

where b_a, b_p, b_d, b_L are fixed bucket assignments on counts of atoms, counts of pairs, compositional depth, and output sequence length. With six atom-count bins, six pair-count bins, five depth bins, and six output-length bins, the unified signature has 1080 possible values. We compute the *number of unique signatures* in each training manifest as the unified diversity measure, intentionally abstracting away from each dataset's idiosyncratic per-adapter construct.

Observed unique signature counts in the cells at $N_{\text{unique}} \in \{500, 2000\}$:

Table 4: Unified-signature cardinality at the endpoint diversity levels of each Tier 2a sweep. CFQ-mcd1's endpoint range is $N \in \{2000, 20000\}$; the others' is $N \in \{500, 2000\}$

Dataset	$N_{\text{unique}} = 500$	$N_{\text{unique}} = 2000$	Δ unique signatures
SCAN	22	25	+2
ReCOGS	14	16	+2
gSCAN-comp	52	57	+4 (mid-cell)
CFQ-mcd1	96 (N=2000)	151 (N=20000)	+55

The unified signature reduces the per-adapter cardinality spread from roughly 40× to roughly 6× (as measured by the ratio of high-N unique-signature counts across datasets). The growth-rate spread across datasets tightens from per-adapter ratios of 1.0× to 5.3× into unified ratios of 1.10× to 1.57×. The unified signature does not *eliminate* construct heterogeneity; it bounds it and makes it quantifiable.

6.3 Step 3: bucket-boundary robustness sweep

The unified signature depends on the choice of bucket boundaries. Different reasonable choices (equal-frequency quantiles, linear spacing, logarithmic spacing, information-theoretic bin selection, random Latin-hypercube assignments) produce different unique-signature counts. We test 44 alternative bucketization schemes (4 quantile + 3 linear + 3 logarithmic + 4 information-theoretic + 30 random Latin-hypercube schemes) and report what fraction of schemes preserve the qualitative findings of Section Section 5.

Robust across all 44 schemes (100%):

- CFQ-mcd1 has zero slope (the flat-floor classification).
- The three learnable datasets (SCAN, ReCOGS, gSCAN-comp) all have positive slope.
- The partition "learnable vs flat-floor" is preserved.

Fragile across schemes:

- The full magnitude rank ordering ReCOGS > SCAN gSCAN > CFQ is preserved in only **34.1%** of schemes.
- The specific claim "ReCOGS > SCAN" holds in 36.4% of schemes.
- The specific claim "ReCOGS > gSCAN" holds in 50.0% of schemes.

Figure 5 shows the per-dataset slope distribution across all 44 schemes. ReCOGS sits at the 90th percentile of its own distribution when the hand-chosen scheme is used; SCAN sits at the 25th percentile. A reader looking at the hand-chosen-scheme slope of ReCOGS = 4.56% per unified-signature and SCAN = 2.07% per unified- signature would conclude ReCOGS has more than 2× the diversity- to-TA slope of SCAN; the same reader looking at the median scheme in each distribution would conclude the slopes are within a factor of 1.3 of each other.

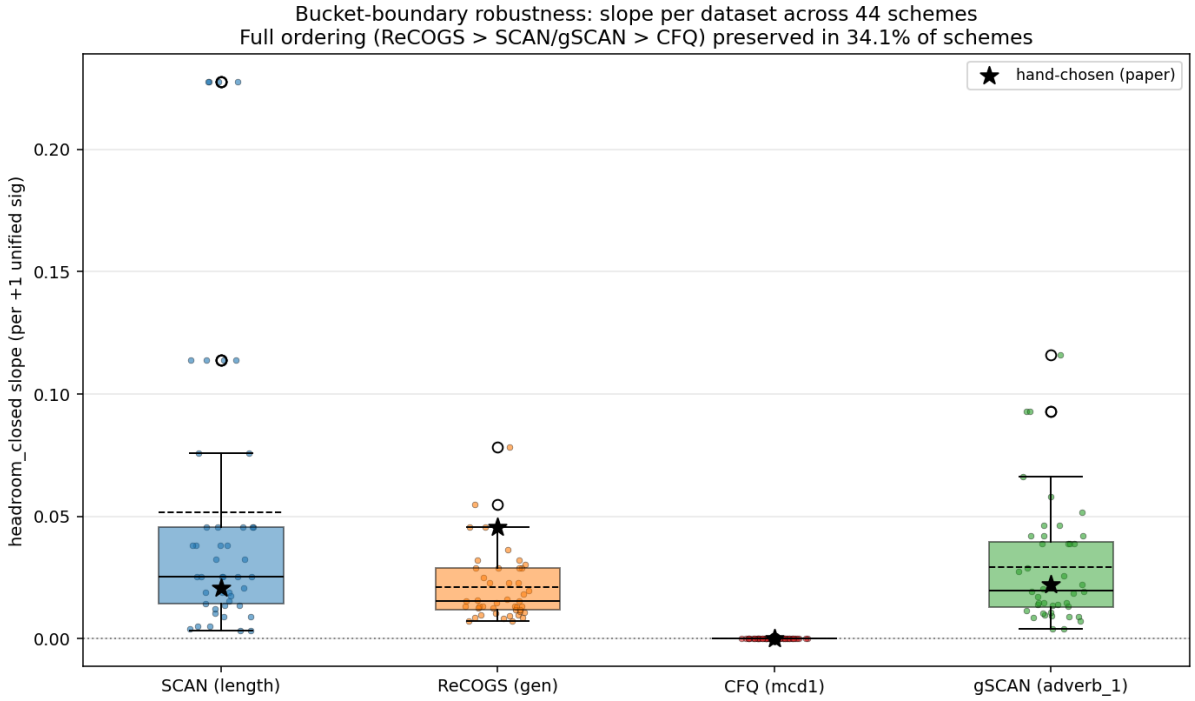

Figure 5: Per-dataset slope distribution across 44 alternative bucketization schemes. CFQ-mcd1 is tightly concentrated at zero across all schemes (100% flat-floor classification at this model and measure). SCAN, ReCOGS, and gSCAN-comp all have positive median slopes, but the per-scheme variability across the three learnable datasets is large enough that magnitude rank orderings are preserved in only 34% of schemes.

The bucket-robustness test is cheap (one shell-level loop over the 44 schemes, no model retraining) and high-yield: it transforms a brittle point estimate into a distribution over modeling choices and flags which conclusions are scheme-dependent.

6.4 Step 4: baseline-arithmetic null calibration

The closed-headroom metric HC defined in Section Section 3.3 is sensitive to the baseline TA. For a fixed absolute gain ΔTA, a *higher* baseline TA produces a larger HC, because the denominator $1 - \text{TA}_{\text{low}}$ is smaller. Two datasets with identical absolute gains but different baselines will therefore land at different HC values, and the cross-dataset spread that results is denominator arithmetic rather than differential signal.

To quantify how much of the observed cross-dataset spread in HC is baseline arithmetic, we run a *synthetic-uniform ΔTA null calibration*: we hold ΔTA fixed at a constant value (we use $+0.03$, a typical mid-range generalization gain) across all datasets, and recompute HC using each dataset's actual baseline TA. The spread of the synthetic HC values across datasets is then attributable entirely to baseline arithmetic.

Observed cross-dataset spread in HC: 2.41× ratio between maximum and minimum. Synthetic-uniform null spread: 1.66× ratio. The synthetic spread accounts for roughly $1.66/2.41 \approx 0.69$ of the observed spread, i.e., about 69% of the cross-dataset HC spread we observe is baseline arithmetic and only about 31% is differential signal.

This does not invalidate the closed-headroom metric, which is the correct first-order normalization for cross-dataset effect-size comparison. It does establish that cross-dataset HC-ranking claims at the magnitude level should be made with explicit reference to the null spread: a 2× ranking difference in HC is only about 1.2× the synthetic-null spread for our four datasets (2.0 / 1.66), comparable in magnitude to what identical absolute gains would produce, and so should not on its own be interpreted as a substantive ranking.

6.5 Step 4b: bootstrap CIs and pairwise permutation tests

The Phase 5c follow-up complements the bucket-robustness sweep (Step 3) and the null calibration (Step 4) with classical statistical infrastructure. Bootstrap slope CIs (B = 10000) at our chosen bucketization:

Table 5: Bootstrap CIs on per-dataset slopes at the hand-chosen bucketization. CFQ-mcd1's CI is essentially zero; the other three CIs are positive and bounded away from zero, but their relative magnitudes carry the Step-3 robustness caveat

Dataset	Slope (HC per unified-signature)	Bootstrap 95% CI
SCAN	$+0.021$	$[+0.013, +0.030]$
ReCOGS	$+0.046$	$[+0.038, +0.056]$
gSCAN-comp	$+0.022$	$[+0.006, +0.038]$
CFQ-mcd1	-0.000015	$[-2 \times 10^{-5}, -1 \times 10^{-5}]$

Pairwise permutation tests at $\alpha = 0.05$ shuffle the per-seed slope labels between two datasets and ask how often the observed difference is exceeded under the null:

Table 6: Pairwise permutation tests for slope-difference significance at 5 seeds per cell. Closest pair to significance is ReCOGS vs CFQ-mcd1 at $p = 0.104$; no pair clears $\alpha = 0.05$

Pair	Permutation p	Distinguishable at $\alpha = 0.05$?
SCAN vs ReCOGS	0.389	no
SCAN vs gSCAN-comp	0.983	no
ReCOGS vs gSCAN-comp	0.515	no
SCAN vs CFQ-mcd1	0.228	no
ReCOGS vs CFQ-mcd1	0.104	no (closest to significance)
gSCAN-comp vs CFQ-mcd1	0.143	no

At five seeds per cell, no pair of datasets in our sweep is pairwise distinguishable by permutation test, despite three of the four bootstrap CIs being disjoint at the visual level. The pooled null distribution under permutation is wider than the individual bootstrap CIs, because the permutation test correctly accounts for the between-dataset variance under exchanged labels.

Adding seeds is the obvious power lever. Whether 10 or 20 seeds per cell would clear $\alpha = 0.05$ on the closest pairwise contrasts is an empirical question we have not run, and the gain depends substantially on whether the additional seeds tighten the per-cell variance or simply extend the existing distribution. The five-seed convention in the field is, by construction, underpowered for the cross-dataset *magnitude* claims that the literature attempts. We do not assert any specific p-value at 10 seeds; we do assert that 5 seeds is below the threshold at which the magnitude claims could be defensible.

6.6 What survives, and what doesn't

Applying the four steps to our own data sorts the claims into two lists.

Robust qualitative claims (transfer cross-dataset):

- Diversity drives, exposure refines, on every learnable benchmark in the sweep.
- CFQ-mcd1 at the 12M-parameter from-scratch configuration shows a flat floor under the current surface-pair measure; no dose-response is detectable across 10× scaling.
- The partition "learnable vs flat-floor" is 100% bucket-robust.

Fragile quantitative claims (do not transfer at 5 seeds per cell):

- Specific magnitude rankings among the three learnable datasets (only 34% bucket-robust).
- Specific slope values (e.g., "ReCOGS at 4.56% per signature") as primary findings (point estimates from one bucketization).
- Direct Cohen's d comparisons across datasets (different metric regimes, different noise floors).
- Cross-dataset "diversity slope" hierarchies inferred from 5-seed effect-size point estimates (no pair pairwise distinguishable by permutation test).

Thesis (Q3). *Cross-dataset quantitative learnability rankings should not be inferred from compositional-benchmark experiments at the 5-seed-per-cell level common in the field. The qualitative diversity-drives-generalization-in-learnable-regimes finding survives every robustness check; the magnitude-ranking claims do not. The four-step audit framework (per-adapter measure audit; unified surface signature; bucket-boundary robustness sweep; baseline-arithmetic null calibration) is the operational mechanism for separating these.*

7 Implications for dataset construction

7.1 SFT data curation: one lens on a known disagreement

A recurring debate in supervised fine-tuning data curation is the tension between "less is more" results (Zhou et al. 2023; Chen et al. 2023), which find that small, quality-curated datasets match or beat larger generic datasets, and "more is more" results, which find continued gains from scale. We do not claim our compositional-benchmark results directly explain that debate, which spans different model scales, different task families, and different evaluation regimes. But the diversity-versus-exposure separation we observe in our regime is one lens that could be applied. If a small curated dataset is more *diverse* (broader coverage of a unified-signature-like space) than a larger dataset is *unique* (high repetition rate per example), the small-curated configuration delivers more diversity dose; if both are high-diversity, the comparison should mostly track other axes. Empirically testing this prediction against the LIMA / AlpaGasus / DEITA literature requires the same diversity and exposure measurements be taken in those papers, which they are typically not.

The diagnostic practitioners can use is the diversity dose: count unique examples (or, more carefully, unique unified signatures) per fixed training budget. The diversity axis is what drives generalization in the learnable regime; the exposure axis is a refinement that matters at intermediate diversity and disappears at saturation.

7.2 The flat-floor signature

The CFQ-mcd1 result demonstrates a *flat-floor signature*: zero dose-response across a wide diversity scaling, robust across the 44 bucketizations we test. The bucket-robustness establishes the *signature itself* (no slope detectable under any reasonable bucketization at this model and measure); it does not by itself establish that capacity is the binding constraint, because the flat floor is also consistent with the surface-pair measure missing the structure that the diversity dose actually administers. A practitioner observing this pattern on a new benchmark should run two follow-ups before concluding capacity is binding: increase model scale, and substitute a canonical (rule-trace or representation-derived) measure for the surface adapter. If the flat floor survives both, capacity is the most parsimonious explanation. If either lifts the slope, the original flat-floor signature was joint between capacity and measure.

7.3 Audit as standard practice for cross-dataset claims

The four-step audit framework of Section Section 6 has a low cost per study: the per-adapter measure audit is a code review; the unified surface signature is a one-time bucketization plus a recomputation of per-cell statistics; the bucket-robustness sweep is a single shell loop over alternative bucket boundaries; the null calibration is a few lines of arithmetic comparing synthetic to observed cross-dataset spread. Applied to our own data, the four steps produced the explicit two-list separation in Section Section 6. We expect the framework to be cheap enough that cross-dataset magnitude claims in subsequent work can be made with explicit reference to the bucket-robustness percentage and the null-calibration share, rather than to a single hand-chosen bucketization with implicit confidence.

7.4 Within-dataset bounds for COGS

For practitioners building COGS training sets at the T5-small scale, our Section Section 4 results yield concrete bounds. For target OOD EM 30%, Experiment 6 $N_{\text{unique}} = 1000$ cells span 27.9%–29.6% at $C \geq 70\%$ (below the 30% target), while $N_{\text{unique}} = 2000$ cells clear it; interpolation suggests $N_{\text{unique}} \approx 1100$–$1200$ is sufficient but not directly observed. For 32%, $N_{\text{unique}} = 2000$

at any $C \geq 77\%$ is directly attested. The observed ceiling is 34% at $N_{\text{unique}} = 4000$. The practical exposure floor is $E \geq 1600$ at $N_{\text{unique}} \in [500, 1000]$; at $N_{\text{unique}} = 2000$, $E \geq 400$ is sufficient; at $N_{\text{unique}} = 4000$, $E \geq 100$ suffices.

The priority order for a fixed COGS dataset budget at T5-small is: (1) raise N_{unique}; (2) ensure $E \geq 1600$ at intermediate N; (3) do not invest in coverage optimization above $C \approx 65\%$, where the observed marginal gain is < 0.5pp per 10pp of additional coverage above $C \approx 70\%$. Extended bounds tables are in Section B.

8 Limitations

Single architecture per scale. Section Section 4 uses T5-small (60M parameters) only. Section Section 5 uses a 12M-parameter transformer only. Whether the diversity-dose-response sigmoidal shape, the exposure escape-probability structure, the coverage null, and the CFQ-mcd1 flat-floor signature all hold at larger scales is open. A larger model might shift the saturation point left (less data buys the same OOD performance) or right (more capacity demands more data); both directions are plausible. The CFQ-mcd1 flat-floor signature is explicitly scale-conditional: we expect it to lift at sufficient model scale, but the present result does not by itself locate the cause in capacity rather than measure (see also the next limitation).

Coverage band not fully tested. The coverage null on COGS is established for $C \in [38.8\%, 88\%]$, but the band below 69.7% is observed only at $N_{\text{unique}} = 200$. At high N the coverage null is not directly tested below 69.7%; below 69.7% at high N, and especially below 65%, a threshold effect cannot be ruled out.

Five seeds per cell is below threshold for cross-dataset magnitude claims. Section Section 6 makes this point quantitatively: no pair of datasets in our sweep is pairwise distinguishable by permutation test at five seeds. The qualitative classification (learnable vs flat-floor) is robust; the magnitude rankings are not.

The unified signature is a surface measure. It uses atom counts, pair counts, depth, and output length, all defined at the surface level of each benchmark's input/output strings. For benchmarks with a well-defined symbolic representation (e.g., CFQ with its rule trace), a canonical-atoms-and-compounds measure would be more authoritative. Phase 3c of the work (canonical CFQ atoms and compounds via rule trace) is deferred. A canonical measure could strengthen, weaken, or revise the CFQ-mcd1 flat-floor finding; the present result is explicitly conditional on the surface adapter.

The CFQ flat floor is conditional on capacity and measure. Our sweep tops out at $N_{\text{unique}} = 20\,000$ on CFQ-mcd1, well within the training pool size; a $10\times$ extension to $N_{\text{unique}} = 200\,000$ would rule out a delayed dose-response under the same model and measure. Independently, a canonical rule-trace measure of CFQ atoms and pairs could lift the slope without changing the data. Our claim is the flat-floor signature itself, not a mechanism: at the 12M-parameter from-scratch configuration and under the current surface-pair adapter, no dose-response is detectable. The next-step disambiguation is two follow-ups: scale-up and canonical measure.

ReCOGS has two seeds in flight. As of write time, the $N_{\text{unique}} = 2000$, $E = 6400$ cell has three of five seeds complete (43 of 45 total runs). The qualitative finding is robust to the missing seeds; the specific Cohen's d at the $E = 6400$ column will be revised on a final write.

Engineered tails / targeted coverage. Our coverage null targets the *global* pair coverage ratio. A more restricted hypothesis remains testable: targeted inclusion of examples covering low-frequency compositional pairs may improve generalization beyond what N_{unique} alone pre-

dicts. We treat this as an open question; a follow-up experiment (the engineered-tails follow-up) tests it directly.

Hard-split findings are weaker. The COGS hard split shows the same qualitative N_{unique} dose-response from a higher threshold toward a lower ceiling (24% at $N = 4000$, $E = 6400$), with higher within-cell variance. We report easy-split findings with higher confidence than hard-split findings.

Scope of "coverage" tested. This work falsifies a *standalone* coverage main effect at COGS / T5-small. The original "Coverage Beats Scale" framing was a *coverage-by-scale interaction* hypothesis, which is logically distinct and remains testable at larger scales (Milestone B Tier 3 covers this).

9 Conclusion

The pair-coverage hypothesis for compositional generalization is a byproduct of training diversity, not an independent mechanism. On COGS with T5-small, an apparent +26.24pp easy-split coverage effect decomposes additively into a +30.20pp diversity confound, a −4.33pp typicality penalty, and a +0.37pp residual that is statistically equivalent to zero under TOST at ±3pp. The qualitative "diversity drives, exposure refines" finding replicates on three of four further compositional benchmarks at the 12M-parameter scale, and CFQ-mcd1 emerges as a flat-floor boundary case at the 12M-parameter from-scratch configuration, with a robust zero-slope signature across a 10× diversity scaling that is jointly conditional on model capacity and on CFQ's non-canonical surface-pair measure.

The harder question is what to make of cross-dataset *magnitudes*. We argue: not much, at the five-seed-per-cell level common in the field. The per-dataset measures of "diversity" are operationally heterogeneous; the unified signature reduces their cardinality spread from 40× to 6× but does not eliminate it; specific magnitude rankings among the learnable datasets are preserved in only 34% of 44 alternative bucketization schemes; about 69% of the observed cross-dataset spread in our closed-headroom metric is baseline arithmetic; and no pair of datasets in our sweep is pairwise distinguishable by permutation test. The four-step audit framework we propose (per-adapter measure audit; unified surface signature; bucket-boundary robustness sweep; baseline-arithmetic null calibration) is the operational mechanism for separating the robust qualitative claims that transfer cross-dataset from the fragile quantitative ones that do not.

For practitioners building training sets, the take-away is concrete: diversity is the axis to invest in; exposure is a refinement at intermediate diversity; global pair-coverage optimization above the tested COGS band did not produce gains once diversity and typicality were controlled. We leave open whether targeted inclusion of examples covering low-frequency compositional pairs (engineered tails) has a non-zero effect at fixed N_{unique}; the COGS null targets the global ratio, not the rare-pair tail. For researchers making cross-dataset claims about compositional generalization, the take-away is methodological: the audit framework gives a cheap way to make magnitude claims with explicit robustness percentages, and the five-seed-per-cell convention is underpowered for the magnitude claims that appear frequently in the literature.

References

Biderman, Stella et al. 2025. "Lessons from the Trenches on Reproducible Evaluation of Language Models." In *Proceedings of NeurIPS Datasets and Benchmarks Track*.

Chen, Lichang, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, et al. 2023. "AlpaGasus: Training a Better Alpaca with Fewer Data." https://arxiv.org/

abs/2307.08701.
Cohen, Jacob. 1988. "Statistical Power Analysis for the Behavioral Sciences." Routledge.
D'Amour, Alexander, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, et al. 2020. "Underspecification Presents Challenges for Credibility in Modern Machine Learning." *arXiv Preprint arXiv:2011.03395*.
Fodor, Jerry A., and Zenon W. Pylyshyn. 1988. "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition* 28 (1–2): 3–71.
Fudan NLP Lab. 2025. "NovelSum: A Data Diversity Measure."
Hupkes, Dieuwke, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. "Compositionality Decomposed: How Do Neural Networks Generalise?" In *Journal of Artificial Intelligence Research*, 67:757–95.
Keysers, Daniel, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, et al. 2020. "Measuring Compositional Generalization: A Comprehensive Method on Realistic Data." In *International Conference on Learning Representations (ICLR)*.
Kim, Najoung, and Tal Linzen. 2020. "COGS: A Compositional Generalization Challenge Based on Semantic Interpretation." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9087–9105.
Lake, Brenden M., and Marco Baroni. 2018. "Generalization Without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks." In *International Conference on Machine Learning (ICML)*, 2879–88.
Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. "Building Machines That Learn and Think Like People." *Behavioral and Brain Sciences* 40: e253.
Lakens, Daniël. 2017. "Equivalence Tests: A Practical Primer for t-Tests, Correlations, and Meta-Analyses." *Social Psychological and Personality Science* 8 (4): 355–62.
Lipton, Zachary C., and Jacob Steinhardt. 2019. "Troubling Trends in Machine Learning Scholarship: Some ML Papers Suffer from Flaws That Could Mislead the Public and Stymie Future Research." In *Queue 17(1)*, 45–77.
Liu, Wei, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. "What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning." In *International Conference on Learning Representations (ICLR)*.
Longpre, Shayne, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, et al. 2023. "The Flan Collection: Designing Data and Methods for Effective Instruction Tuning." In *International Conference on Machine Learning (ICML)*.
Nguyen, and Ploeger. 2025. "Measuring Data Diversity for Instruction Tuning." In *Proceedings of EMNLP*.
Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21 (140): 1–67.
Ruis, Laura, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. 2020. "A Benchmark for Systematic Generalization in Grounded Language Understanding." In *Advances in Neural Information Processing Systems (NeurIPS)*.
Sun, Kaiser, Adina Williams, and Dieuwke Hupkes. 2023. "Validity-Preserving Delexicalisation Methods for Compositional Generalisation Benchmarks." In *Proceedings of EMNLP*.
UBC NLP Group. 2025. "Facts in Stats: Pretraining Data Diversity and Downstream Generalization."
Wu, Zhengxuan, Christopher D. Manning, and Christopher Potts. 2023. "ReCOGS: How Incidental Details of a Logical Form Overshadow an Evaluation of Semantic Interpretation." In *Transactions of the Association for Computational Linguistics (TACL)*.

Zhou, Chunting, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, et al. 2023. "LIMA: Less Is More for Alignment." In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Appendix A: COGS cell-level results

Table 7: Quick-reference of COGS cells used in Section Section 4

Experiment cell	N_{unique}	Coverage	Mean EM (easy)	Coverage effect
E1 high_easy	4000	high, greedy	34.1%	—
E1 low_easy	200	44%, random	7.9%	+26.2pp (confounded)
E2 high_easy (greedy-200)	200	97.2%	3.9%	−4.0pp (reversed)
E2 low_easy (random-200)	200	44%	7.9%	—
E3 high_easy (random, high cov)	200	53.7%	6.31%	+0.37pp (null)
E3 low_easy (random, low cov)	200	38.8%	5.94%	—
E4 $N = 200$	200	natural	7.87%	—
E4 $N = 1000$	1000	natural	27.63%	—
E4 $N = 4000$	4000	natural	33.18%	—
E5 $N = 4000$, $E = 400$	4000	natural	33.68%	—
E5 $N = 1000$, $E = 1600$	1000	natural	28.84%	—
E5 $N = 200$, $E = 6400$	200	natural	1.91%	—
E5 hard $N = 4000$, $E = 6400$	4000	natural	24.26%	—
E6 $N = 1000$, q_1	1000	69.7%	29.61%	—
E6 $N = 1000$, q_{99}	1000	81.5%	29.22%	−0.39pp (null)
E6 $N = 2000$, q_1	2000	77.0%	31.91%	—
E6 $N = 2000$, q_{99}	2000	87.3%	31.83%	−0.07pp (null)

B Appendix B: empirical bounds for COGS dataset construction

Table 8: Diversity targets per OOD EM goal on COGS at T5-small

OOD EM target	Required N_{unique}	Coverage required	Basis
20%	400 to 500	any 65%	E5 N=500, E=1600: 21.67%
25%	600 to 700 (interp.)	any 65%	Interpolated E4/E5
30%	**1100 to 1200 (interp.)**	**any 70%**	E6 N=1000 spans 27.9%–29.6%; N=2000 clears 30%
32%	**2000**	**any 77%**	E6 N=2000, q_1: 31.91% at 77.0% (directly attested)
34% (observed ceiling)	4000	natural (96%)	E5 N=4000, E=400: 33.68%

Table 9: Exposure floors that produce reliable convergence at each COGS diversity level

N_{unique}	E for near-universal escape	Notes
200	not achievable	floor mode; no exposure rescues
500	1600	2/5 escape at $E = 400$, 5/5 at 1600
1000	1600	4/5 at $E = 400$, 5/5 at 1600
2000	400	full escape at $E = 400$
4000	100	near-universal regardless

C Appendix C: pipeline-bugfix audit for COGS Experiment 4

Two pipeline bugs were identified by pre-execution code review between Experiments 4 and 5, and corrected before any Experiment 5 run.

Epoch shuffle bug. The training loop used `itertools.cycle(loader)`, which memoizes the first epoch's shuffle order and replays it identically on every subsequent epoch. Every pre-bugfix run saw the same batch sequence from epoch 2 onward. The fix replaces the call with a generator `while True: yield from loader` that calls `iter(loader)` once per pass, letting the DataLoader's `RandomSampler` draw a fresh permutation per epoch. A regression test (`tests/test_training/test_loop.py::test_epoch_cycle_reshuffles_per_epoch`) serves as a permanent guard.

Selection-strategy default. Experiment 4 manifests had been generated with the greedy strategy (`fixed_coverage_scaled_repetition`) rather than `random_baseline`. This was corrected before any Experiment 5 run, so Experiments 5 and 6 use the registered `random_baseline` selection throughout.

Effect on E4 anchors. Re-validation of Experiment 4 anchor points with both fixes applied showed negligible differences at $N \geq 1000$ (less than 1pp) but a larger shift at $N = 200$ (-2.44pp, with increased within-cell variance from genuine stochastic resampling). Conclusions of Experiments 1 through 4 at $N \geq 500$ are unaffected; $N = 200$ results before the fix should be read with this caveat. Experiments 5 and 6 use the corrected pipeline. The bugfix-validated Experiment 4 anchors are 28.56% at N=1000 and 33.84% at N=4000, consistent with Experiment 5's matched cells (28.84% and 33.68% respectively).

D Appendix D: COGS Experiment 5 bimodal cell audit

The high within-cell variance at some transition-zone cells of Experiment 5 is not measurement noise. A 2-component GMM with BIC model comparison identifies several cells as mixtures of a floor mode and an escaped mode.

Table 10: Experiment 5 easy-split cells with $\Delta\text{BIC} > 0$ favouring a 2-component Gaussian mixture; one hard-split cell (N=4000, E=6400) also has $\Delta\text{BIC} > 0$ but is not analyzed here

Cell	Mean EM	Std	Lower mode	Upper mode	Escape rate
N=1000, E=400	23.95%	13.36%	0.06%	29.93%	4/5
N=500, E=400	8.69%	11.54%	0.3%	21.3%	2/5
N=500, E=100	10.37%	9.78%	2%	21%	2/5
N=2000, E=100	25.77%	6.05%	16.64%	28.06%	4/5

E Appendix E: COGS experiment storyboard

Figure 6 collects per-experiment views of the COGS arc covered in Section Section 4. Each panel re-renders the headline cell means or distribution for one of the six experiments, in the chronological order of the arc.

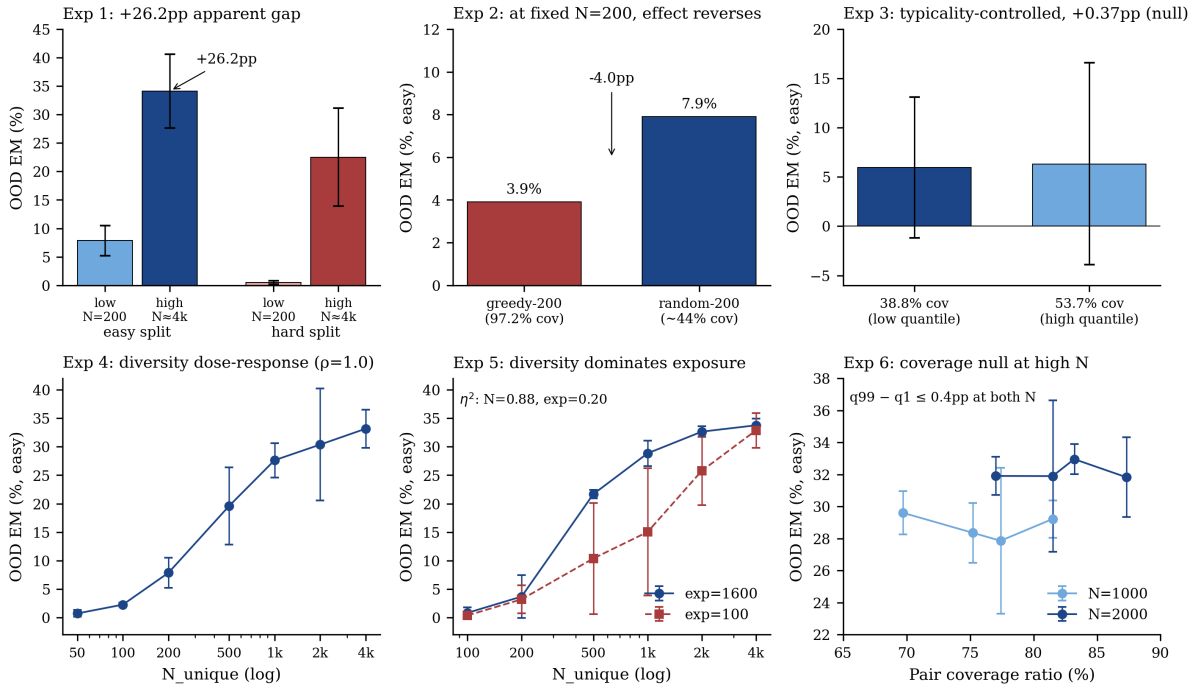

Figure 6: Experiments 1 through 6 storyboard for the COGS section. Top row: Experiment 1 starting point (left), Experiment 2 reversal at fixed N (centre), Experiment 3 typicality-controlled null at $N = 200$ (right). Bottom row: Experiment 4 diversity dose-response (left), Experiment 5 diversity-versus-exposure separation showing diversity dominance (centre), Experiment 6 coverage-at-high-N null (right).

F Appendix F: cross-dataset Tier 2a per-dataset details

This appendix records per-dataset cell means used in Section Section 5. Headline numbers are reproduced in the body; this appendix gives the full grids.

F.1 SCAN length-split (3×3, 5 seeds per cell, 45 total runs)

EM on length_test ranges 6.6%–10.2% across the 9 cells. Diversity Cohen's d at endpoints ($N = 500 \to 2000$): 3.95 at $E = 400$, 2.84 at $E = 1600$, 4.08 at $E = 6400$. Exposure Cohen's d at endpoints ($E = 400 \to 6400$): up to 0.98 at $N = 500$; vanishes at $N \geq 1000$. Canonical source: `experiments/milestone_b/tier2a_scan/SUMMARY.md`.

F.2 ReCOGS gen (3×3, 5 seeds per cell, 43 of 45 complete)

EM at 0% across all completed runs (metric floor at this scale); primary metric is TA. TA at N=2000, E=400: far_OOD 0.5212, near_OOD 0.5214, iid 0.5181, depth_OOD 0.5153 (far_OOD iid indicates genuine generalization). Diversity Cohen's d at endpoints on TA: 20.67 at $E = 400$, 24.80 at $E = 1600$, 16.46 at $E = 6400$. Exposure Cohen's d on TA: -0.71 to $+0.18$. Canonical source: `experiments/milestone_b/tier2a_recogs/SUMMARY.md`.

F.3 gSCAN-comp adverb_1 (3×3, 5 seeds per cell, 45 total runs)

TA endpoint cells: $N = 500$ gives 17–20% TA; $N = 2000$ gives 22–30% TA. $N = 1000$ middle cell shows a pathological dip (TA 5–15%, standard deviation up to 0.14) attributable to optimization variance at this benchmark. Diversity Cohen's d at endpoints on TA: 0.55 at $E = 400$, 1.53 at $E = 1600$, 0.62 at $E = 6400$. Per-adapter `template_id` saturates at 32 across all N levels (a per-adapter false-negative); unified-signature dose across endpoints is $\Delta = +4$ signatures ($52 \rightarrow 57$), confirming a real diversity dose is being administered. Canonical source: `experiments/milestone_b/tier2a_gscan_comp/SUMMARY.md`.

F.4 CFQ-mcd1 scaling sweep (4 cells, 5 seeds per cell, 20 total runs)

EM at 0.0000 ± 0.0000 across all four cells (the 12M model emits boilerplate SPARQL without producing complete correct queries). TA per cell: 0.1066 ± 0.0007 (N=2000), 0.1060 ± 0.0003 (N=5000), 0.1052 ± 0.0002 (N=10000), 0.1057 ± 0.0002 (N=20000). Unified-signature dose administered: $\Delta = +55$ signatures across the sweep. Bootstrap slope CI on TA against $\log N$: $[-2 \times 10^{-5}, -1 \times 10^{-5}]$, width 10^{-5}. Surface template-per-N ratio improved from 0.84–0.96 (pre-Phase-3a fix) to 0.35–0.66 (post-fix) at $N = 20000$, but remains ~2.3× higher than SCAN/ReCOGS, reflecting CFQ's surface co-occurrence construct (rather than binding). Unique pairs plateau at 95–168 across the 10× N scaling, suggesting the compositional pair space does not scale with N at this benchmark under the current adapter. Canonical source: `experiments/milestone_b/tier2a_cfq_scaling/SUMMARY.md`.

G Appendix G: unified surface signature and harmonized metrics

The unified surface signature is defined as a 4-tuple $\text{sig}(e) = (b_a, b_p, b_d, b_L)$ over the bucketized counts of atoms, pairs, compositional depth, and output sequence length of each training example. The bucket boundaries used in the body of the paper:

- Atom count: {[0], [1–2], [3–4], [5–7], [8–11], [12+]}
- Pair count: {[0], [1–2], [3–4], [5–7], [8–11], [12+]}
- Depth: {[1], [2], [3], [4], [5+]}
- Output length: {[≤ 5], [6–10], [11–25], [26–50], [51–100], [101+]}

The space has $6 \times 6 \times 5 \times 6 = 1080$ possible signatures. Per-dataset coverage at the endpoint cells:

Table 11: Unified-signature cardinality at the endpoint cells of each Tier 2a sweep

Dataset	low-N unique sigs	high-N unique sigs
SCAN (N=500−2000)	22	25
ReCOGS (N=500−2000)	14	16
gSCAN-comp (N=500−2000)	52	57
CFQ-mcd1 (N=2000−20000)	96	151

The closed-headroom metric is $\text{HC} = (\text{TA}_{\text{high}} - \text{TA}_{\text{low}})/(1 - \text{TA}_{\text{low}})$. The harmonized cross-dataset HC values used in the body:

Table 12: Harmonized cross-dataset closed-headroom metric, using the Phase 5b post-null-calibration cell averages (`SUMMARY_phase5b_headroom_null.md`). The earlier Phase 3d table used a different gSCAN baseline (TA 0.1996 instead of 0.1825) and is superseded here. The specific HC values carry the bucket-robustness caveat of Section Section 6.3 and the baseline-arithmetic caveat of Section Section 6.4

Dataset	low-N → high-N	Δ sigs	TA range	Δ TA	HC
SCAN	500 → 2000	+2.2	0.4064 → 0.4334	+0.0270	+4.55%
ReCOGS	500 → 2000	+2.4	0.4620 → 0.5210	+0.0590	+10.97%
CFQ-mcd1	2000 → 20000	+55.2	0.1066 → 0.1057	−0.0008	−0.09%
gSCAN-comp	500 → 2000	+4.2	0.1825 → 0.2583	+0.0758	+9.27%