BEYOND FACT FREQUENCY: CONTROLLED EVIDENCE ON EXPRESSION ALLOCATION AND FACTUAL CALLABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Fact frequency does not specify how a fixed fact's exposure is distributed across linguistic expressions. We isolate this *within-fact expression-allocation* estimand by matching fact identity, exposure count, answer occurrence, tokenizer and training budgets, and corpus-level template marginals. In controlled synthetic settings, distributing fixed exposure across faithful expressions instead of repeating one substantially improves held-out Forward callability: +28.75 percentage points across ten Qwen3-0.6B training seeds, with replication across independently generated worlds from the same programmed generator, larger Qwen and Llama models, natural held-out wording, and a same-family Qwen3 persistence test through the largest tested 8B checkpoint. We then apply the same intervention to a five-domain natural-source panel at two Qwen sizes, where generated behavior and teacher-forced target-answer likelihood are heterogeneous across domains and model sizes, with strict exact match remaining at floor. Dedicated stress tests do not support a universal relation-preservation mechanism. Thus, in our controlled construction, equal fact frequency need not imply behaviorally equivalent training exposure, while natural-data, bidirectional-transfer, scaling-law, and internal-mechanism claims remain bounded.

1 INTRODUCTION

Language-model training data are commonly summarized by size, quality, or how often a fact, entity, or concept appears. Frequency is important, but incomplete: it does not say whether repeated exposure uses one expression or several. From a general learning perspective, count and allocation are distinct coordinates of the training signal. Two corpora can contain the same facts and answers, the same number of rows, and the same training and token budget, yet attach each fact's fixed exposure to one expression in one corpus and several expressions in the other. We therefore ask: *holding factual content and exposure fixed, does within-fact expression allocation change whether a fact becomes callable?*

Knowledge augmentation, synthetic continued pretraining, contextual diversity, multi-format supervision, and broad restatements can improve factual extraction or robustness (Allen-Zhu & Li, 2023; Yang et al., 2024; Chen et al., 2024; Liu et al., 2026; O'Neill, 2026). A generic "diversity helps" claim is therefore neither new nor sufficiently precise. Most prior natural-language and data-level comparisons do not isolate our estimand because expression breadth changes together with per-item exposure, answer realization, token budget, format allocation, or associative support. Our contribution is not the discovery that varied contexts can improve accessibility; it is the matched identification of a within-fact expression-allocation effect under fixed exposure and held-out per-fact Forward evaluation.

We estimate this quantity with paired training conditions. In SINGLE-REPEAT, each fact is presented four times through one expression; in DIVERSE, the same exposure is allocated across four faithful expressions. Fact identities, answer occurrences, optimization and tokenizer-measured budgets, and corpus-level template marginals are matched. Held-out Forward interfaces ask for the trained object through an expression not seen for that fact. The primary independent unit is a training seed or generated world, never an individual probe.

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.

This design yields a large controlled synthetic Forward effect across training seeds, independently generated worlds from the same programmed generator, two Qwen sizes, Llama-3.2-1B, and held-out natural wording over the same synthetic facts. We then test rather than assume its generality. A five-domain natural-source extension at two Qwen sizes shows positive generated containment in four domains under both sizes, but held-out target-answer NLL is lower under DIVERSE in only seven of ten balanced domain–model strata and in two domains under both sizes. Strict generated exact match remains at floor. A source-linked Natural Questions extension terminates at a preregistered semantic-expression criterion before model inference; Reverse transfer is unsupported; and two mechanism stress tests reject a simple universal relation-preservation explanation.

The contribution is thus a controlled empirical finding plus calibrated boundaries, not a new augmentation recipe or complete document-to-model account. Closely related work establishes that biographies, rephrased documents, quizzes, formats, or restatements can help (Yang et al., 2024; Kim et al., 2026; Zhou et al., 2026; Niklaus et al., 2026); we identify what remains after frequency, answer occurrence, sample count, tokenizer budget, and global template frequency are matched within facts, and report both replications and failed generalizations.

Our contributions are:

- an estimand that isolates within-fact expression allocation while matching fact identity, exposure, answer occurrence, training/token budget, and global template marginals;
- large held-out Forward callability differences replicated across seeds, independently generated worlds from the same programmed generator, Qwen sizes, Llama, held-out wording, and a same-family Qwen3 240-fact persistence test through the largest tested 8B checkpoint;
- a five-domain, two-size natural-source extension showing model- and domain-dependent generated-answer containment and target-answer likelihood; and
- diagnostics and explicit natural-data, Reverse, mechanism, and beyond-tested-capacity boundaries that prevent the controlled effect from being promoted into a universal law or data recipe.

2 RELATED WORK

Knowledge storage and expression diversity. Synthetic biographies show that repeated biographies, permutation, and augmentation can improve extraction of stored facts (Allen-Zhu & Li, 2023). EntiGraph synthesizes diverse text from source documents for continued pretraining (Yang et al., 2024), while broad study items improve later use and retention during continual fact writing (O'Neill, 2026). Behnia et al. (2025) independently manipulate abstract contextual structure and diversity, finding training-time-dependent factual generalization. These studies establish that expression or context matters. They do not provide our combination of fixed per-fact exposure, answer exposure, tokenizer budget, global template marginals, world-level replication, and explicit boundary tests.

Table 1 makes the closest construct-level distinctions explicit. It is intentionally not a checklist of claimed superiority: each neighboring study controls a different alternative explanation, and "not reported" means only that the cited paper does not establish that property. The closest abstract testbed, *Facts in Stats*, approximately matches fact and template frequencies while changing their pairing support; the closest item-aligned control, FormatMix, matches total training instances while reallocating formats across source items. Neither comparison makes the present result automatic because the estimands, linguistic units, and acquisition endpoints differ.

Repeated data are also not a cost-free substitute for new data: deduplication changes memorization and language-model quality (Lee et al., 2022), while data-constrained scaling studies characterize when repeated epochs retain value (Muennighoff et al., 2023). These results motivate SINGLE-REPEAT as a scientifically meaningful control. Our contrast does not ask whether more data are useful; both conditions contain the same number of fact and answer exposures.

Synthetic-data diversity and cross-format robustness. Dataset-level diversity correlates with downstream utility in synthetic pretraining data (Chen et al., 2024); structured web rephrasing, quiz-enhanced pretraining, cross-document QA synthesis, and multi-format supervision also improve downstream behavior (Niklaus et al., 2026; Kim et al., 2026; Zhou et al., 2026; Liu et al., 2026). Their

Work	Analysis unit / intervention	Reported matched control	Difference from our estimand
TELLME (Kim et al., 2026)	Domain corpus; plain text mixed with generated explanatory QA	TELLME and CPT+IT see the same total plain-text and QA tokens	No stable fact-level allocation or per-fact answer-exposure control; evaluates domain benchmarks and retention
WRAP++ (Zhou et al., 2026)	Documents and cross-document entity relations; relational QA synthesis	Selected ablations use an approximately 8B-token budget	Source relations and associative coverage change; no fixed per-fact exposure or answer occurrence
FormatMix (Liu et al., 2026)	Stable source QA item; one versus multiple answer formats	Same source-item pool and total training instances for 30% FormatMix and MCQ-Matched	Allocation is across a selected subset of items; per-item exposure and answer realization are not fixed across arms
Facts in Stats (Behnia et al., 2025)	Atomic source–target pair embedded in abstract statistical templates	Fixed-length training; facts and templates have approximately equal marginal frequency at a given diversity level	Diversity changes the template–fact exposure mask and ID/OOD support in an abstract token stream, not matched natural-language QAS
O'Neill (O'Neill, 2026)	Invented fact; narrow statements versus 24 broad study or recitation items	Same facts, model, seeds, and optimization steps in the causal factorization	Prompt breadth and token throughput differ; endpoint is recitation-to-use and continual retention rather than fixed-exposure Forward callability
Ours	Stable synthetic fact; four training rows allocated to one or four faithful expressions	Fact, rows, serialized and native-token answer exposure, optimizer and token budgets, and global template marginals	Held-out per-fact Forward callability with seed/world inference and separately reported natural, Reverse, and mechanism boundaries

Table 1: Closest comparisons to the matched within-fact expression-allocation estimand, based on full-text verification. The appendix provides a broader comparison and evidence qualifications.

comparison unit is a dataset, source, sample, or format allocation, and data amount or associative coverage may change. We instead maintain stable fact identities and equal per-fact exposure, asking whether allocation alone changes callability.

Data transformation and curation. Question generation itself can select evidence unevenly and distort supervision (Alimaskina et al., 2026). DataPrep-Bench evaluates end-to-end data preparation by downstream utility, while DataOrchestra learns per-example choices among dropping, cleaning, rewriting, and retaining data (Liang et al., 2026; Huang et al., 2026). These works motivate source-aware evaluation but do not isolate within-fact allocation under the same matching constraints. Our focus is diagnostic rather than a curation policy. The natural-source extensions use source-QA factual content, but their controlled silver rewrites and weaker matching do not establish a complete document-to-evidence-span-to-fact-to-QA-to-model chain.

3 Experimental Design

3.1 Estimand and matched conditions

For world w, training seed s, condition $c \in \{S, D\}$, and endpoint-selected paired probe IDs $P_{w,s}$, let $Y_{w,s,i}(c)$ be the stored outcome for probe i. The controlled synthetic Forward-EM estimand is implemented as

$$\bar{Y}_{w,s}(c) = |P_{w,s}|^{-1} \sum_{i \in P_{w,s}} Y_{w,s,i}(c), \qquad d_{w,s} = \bar{Y}_{w,s}(D) - \bar{Y}_{w,s}(S),$$
$$\Delta_{\text{seed}} = |S|^{-1} \sum_{s \in S} d_s \ (w \text{ suppressed}), \qquad d_w = |S_w|^{-1} \sum_{s \in S_w} d_{w,s}, \quad \Delta_{\text{world}} = |W|^{-1} \sum_{w \in W} d_w. \tag{1}$$

Here S, S_w, and W are the fixed-world seed set, the seeds observed in world w, and the world set. The top-level unit is a training seed for fixed-world replication and a generated synthetic world for cross-world studies; probes and facts are aggregated within units and never treated as IID top-level observations.

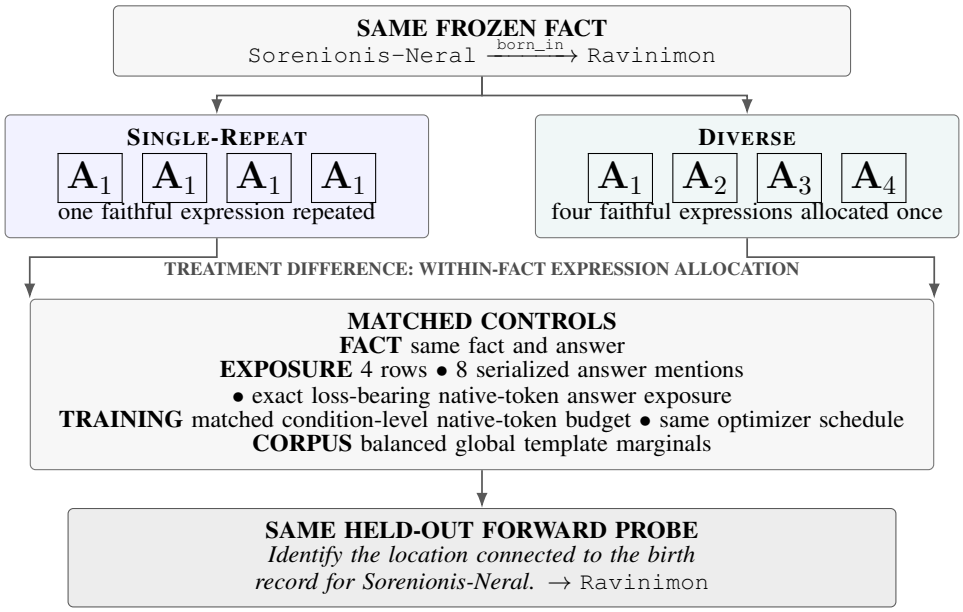

Figure 1: Fixed-exposure within-fact allocation. Here $A_1, \ldots, A_4$ denote different faithful expressions of the same frozen fact, not different facts. SINGLE-REPEAT assigns four exposures to A_1; DIVERSE allocates the same four exposures across $A_1, \ldots, A_4$. Fact identity and answer, answer exposure, condition-level native-token budget, optimizer schedule, global template marginals, and the held-out Forward probe are matched; the treatment contrast is within-fact expression allocation under these controls.

Both conditions contain four training rows per fact: SINGLE-REPEAT repeats one faithful expression and DIVERSE allocates the four rows across four faithful expressions. Under the frozen Question–Answer–Solution serialization, each row carries the answer in Answer and Solution, yielding eight serialized mentions per fact; a full-context native-tokenizer audit confirms matched loss-bearing answer-token exposure.

Figure 1 visualizes the contrast; Appendix A preserves the registered strings. Synthetic worlds contain 24 facts/96 rows per condition; the confirmatory source-QA study retains 43 parents/172 rows. Paired conditions share facts, answer exposure, initialization, optimization, and probes; synthetic totals match exactly and the natural per-parent ceiling is 5% (maximum 2.63%).

Paired synthetic conditions exactly match per-fact loss-bearing answer-token exposure and condition-level total native-token budget. Full-row token counts need not be identical for every individual fact because non-answer linguistic realization is part of the expression-allocation intervention. Synthetic corpora additionally balance corpus-level template marginals, preventing a global template-frequency difference from identifying the condition. Training and evaluation text for a fact are disjoint.

3.2 SYNTHETIC WORLDS AND PROBES

Each synthetic world contains three relation types with eight facts per relation (24 facts total). Five world seeds (101, 203, 307, 401, 509) generate disjoint fact, subject, and object identities. Three training seeds (157, 173, 191) are crossed with two conditions, producing 30 runs per model size. Train and probe registries are frozen separately: for a given fact, no training question is reused as an evaluation question. The primary endpoint is exact match on held-out Forward cross-interface probes. Forward asks for the trained object from a held-out expression of the subject–relation pair. Reverse, which asks for the subject from an object, is analyzed separately and does not contribute to the Forward primary inference.

The held-out wording evaluation reuses the 30 Qwen3-1.7B checkpoints without additional training. It replaces generator-style Forward probes with 12 manually verified natural templates, yielding 96

probes per world and 480 base-probe rows. This changes wording and interface while retaining the same synthetic facts, relation ontology, model, and checkpoints.

3.3 NATURAL-SOURCE TRANSFER AND BOUNDARIES

The natural-source experiments use factual propositions derived from read-only Finance, Earth-science, Materials, Life science, and Medical source-QA artifacts. Training expressions are controlled, pipeline-verified faithful rewrites, not naturally occurring diversity. Each retained parent has four SINGLE-REPEAT or four DIVERSE training rows and two held-out source-QA probes. These weaker-control natural-source artifacts provide less control than the synthetic worlds: per-fact native-token totals are approximate and expressions are silver rather than independently human-verified. The Finance artifact contains 300 facts; each other artifact contains 200 facts.

Natural-source endpoints. Strict normalized exact match is a stringent generated-behavior endpoint: it is one only when the fully normalized response equals the fully normalized canonical answer. Normalized containment is a permissive generated-behavior endpoint: it is one when the fully normalized canonical answer occurs as a contiguous character substring of the fully normalized response. For the natural-source evaluator, normalization applies Unicode case folding, removes ASCII punctuation, deletes all whitespace, and applies no Unicode normalization, article removal, alias handling, or token-boundary rule; empty normalized answers or responses score zero. The earlier confirmatory source-QA evaluator uses the same nonempty substring/equality rules but removes Unicode-category punctuation and collapses remaining whitespace to single spaces. Held-out answer-token NLL is instead a teacher-forced likelihood diagnostic: the model is conditioned on the held-out probe followed by the evaluator's literal answer-prefix marker and the canonical answer, and the loss is averaged over only the native-token positions overlapping the answer span, with equal token weighting. Lower NLL is better; it is not a generated-behavior endpoint and does not by itself establish retrieval, storage, or mediation. Reported seed values are token-weighted within seed, and table means average those seed-level diagnostics without pooling domains.

The diagnostic study uses 27 parents with strict exact match and is retained as a boundary result rather than confirmatory evidence. The confirmatory study uses 44 independently reviewed candidate parents with no diagnostic-set overlap. A condition-free base evaluation is followed by a prespecified ceiling rule, excluding one parent and retaining 43. The five-seed comparison crosses two conditions (10 runs and 10 final-checkpoint evaluations) over 86 probes. Its amended primary endpoint is normalized answer containment; strict normalized exact match is secondary. The endpoint amendment occurred after condition-free base outputs were visible but before condition training or results, and is reported as a protocol deviation rather than hidden.

The subsequent natural breadth–capacity extension uses the five-domain panel at Qwen3-0.6B and 1.7B with the balanced three-seed comparison, reusing earlier Qwen3-0.6B Finance and Earth-science results. We report all domains as natural-source boundary evidence, keep domain and model strata separate, and do not treat the weaker controls as a second causal estimate.

3.4 MODELS AND TRAINING

The main Forward analyses train Qwen3 base models (Yang et al., 2025) at 0.6B and 1.7B parameters and Llama-3.2-1B; the first mechanism stress test additionally includes Llama-3.2-3B, and the larger-inventory/8B robustness matrix adds Llama-3.1-8B. Each snapshot is frozen before condition results. Artifacts are reconstructed for the model's native tokenizer when required while preserving identical semantic payload; Qwen3-0.6B/1.7B share exact token IDs on all frozen inputs, whereas Llama receives tokenizer-specific matched artifacts. All main Forward replication runs use full causal language modeling for 10 epochs, learning rate 5×10^{-5}, no weight decay, 5% warmup, maximum length 128, bfloat16 precision, batch size 2, and gradient accumulation 8. Mechanism and final robustness experiments retain separately frozen model-specific resource settings; the Llama-3.1-8B cells use four-rank FSDP and the three cells are analyzed independently. The natural confirmatory evaluation permits 64 generated tokens; synthetic evaluation permits 16. The five-seed Finance extension uses the same Qwen3-0.6B natural-source recipe and is interpreted as natural-source boundary evidence. No automatic retries, seed replacement, or selective run exclusion are allowed.

Domain	Containment D−S (pp)		Held-out NLL D−S		Direction (positive/lower)		Strict EM	Evidence role
	0.6B	1.7B	0.6B	1.7B	Cont.	NLL		
Finance	+3.56	+3.44	−0.2930	+0.0336	3/3,3/3	3/3,1/3	$\leq 0.08\%$	Model-conditioned boundary
Earth-science	−4.25	+0.17	−0.2362	−0.0239	1/3,1/3	3/3,3/3	$\leq 0.08\%$	Domain-conditioned boundary
Materials	+5.00	+3.42	−0.1663	−0.2909	3/3,3/3	3/3,3/3	$\leq 0.08\%$	Natural-source boundary evidence
Life science	+2.33	+2.92	+0.0310	−0.1281	2/3,3/3	1/3,2/3	$\leq 0.08\%$	Natural-source boundary evidence
Medical	+0.08	+7.33	+0.0830	−0.0511	2/3,3/3	1/3,3/3	$\leq 0.08\%$	Natural-source boundary evidence

Table 3: Balanced three-seed natural-source results for Qwen3-0.6B and Qwen3-1.7B, kept as separate domain/model strata. Containment is generated behavior (percentage-point DIVERSE−SINGLE); held-out NLL is a teacher-forced diagnostic (lower is better); direction columns report positive-containment and lower-NLL seed counts in 0.6B,1.7B order. Strict normalized EM remains at floor. Rows are not pooled with synthetic effects; the five-seed Finance extension remains in Appendix H.

P43 is a separate formal same-family persistence matrix under the frozen `large240-v2` artifact: Qwen3-1.7B and Qwen3-8B Base, worlds 101/203/307/401/509, seeds 157/173/191, four exposures per fact, and the held-out Forward cross-interface exact-match endpoint. Both cells use full-parameter causal-LM training for 10 epochs and select the final epoch-10 checkpoint; the 8B implementation uses the frozen FSDP configuration. Worlds are the top-level inference units with training seeds nested, and model strata are never pooled.

3.5 Inference and decision criteria

Effects are reported as DIVERSE minus SINGLE-REPEAT; the independent unit is a preregistered training seed or synthetic world, never an individual probe. For paired synthetic replications we report the mean unit effect, a descriptive bootstrap interval, and an exact test. **Wins/ties/losses count top-level paired units with DIVERSE−SINGLE effect greater than, equal to, or less than zero, respectively.** The exact one-sided sign-flip test assumes sign symmetry, uses the mean statistic, enumerates all 2^n assignments for $n = 5, 10$, and tests the `greater` alternative. With five world-level units, resolution is coarse; results are interpreted jointly through effect magnitude and directional consistency. These equations describe synthetic Forward EM; natural-source containment, strict EM, and answer-token NLL use endpoint-specific summaries in Sec. 3.3. Detailed criteria and pooling boundaries are in the appendix.

4 Results

We organize the evidence by scientific role rather than experiment number. **RQ1** asks whether the controlled effect survives training randomness, independently generated worlds, model changes, and larger fact inventories. **RQ2** asks whether it survives more natural held-out wording over the same synthetic facts. **RQ3** asks whether the pattern remains detectable when the factual content comes from real source-QA under weaker control. **RQ4** asks whether a simple relation-preservation account explains the controlled effect across model strata. Figure 2 summarizes comparable synthetic Forward replications; Appendix D keeps non-comparable boundary endpoints and inference units separate.

Earlier source-QA, constructibility, Reverse, and mechanism studies provide complementary boundary evidence. A matched source-QA confirmation missed its prespecified criterion, stricter natural constructions stopped before treatment-model execution, Reverse transfer is unsupported, and relation stress tests do not support a universal mechanism; full endpoint-specific outcomes and decision criteria are reported in Appendix D.

Table 3 summarizes the five-domain panel rather than a single pooled treatment effect. The reader should compare containment and teacher-forced answer-token NLL within each domain–model stratum; these endpoints answer different questions and are not interchangeable. The panel is included to locate the synthetic result's external-validity boundary, not to establish a second causal replication.

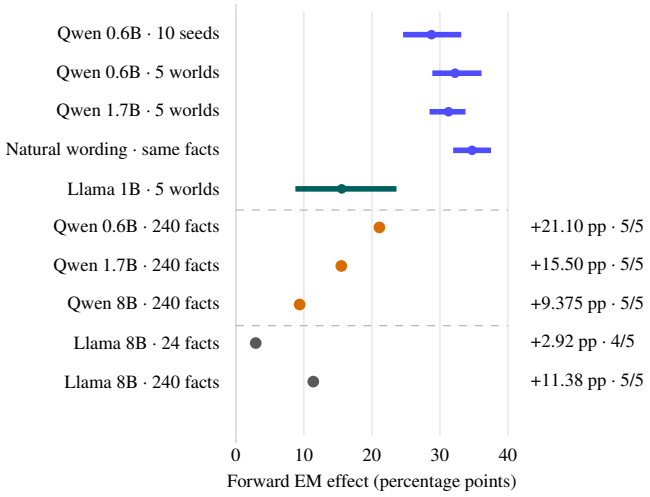

Figure 2: Held-out Forward exact-match effects grouped by scientific role. Blue intervals summarize core replications; orange points are the same-family Qwen3 240-fact persistence cells, and gray points are separate cross-family/large-regime cells. The Qwen3 persistence points are not connected, pooled, or fit with a trend; they show positive effects through the largest tested 8B checkpoint, not a scaling law.

4.1 Controlled synthetic replication

RQ1: Does the effect survive seeds, worlds, model family, and larger regimes? The one-world pilot produces a +29.17-point effect. The preregistered ten-seed replication estimates +28.75 points (95% bootstrap CI [24.58, 33.13]), with 10/0/0 wins/ties/losses and exact $p = 0.00098$. Thus the pilot direction is not explained by one favorable initialization.

The five-world Qwen3-0.6B experiment estimates +32.22 points [28.88, 36.11], with all five worlds positive and exact $p = 0.03125$. The same independently generated worlds and tokenizer-compatible artifacts at Qwen3-1.7B yield +31.25 points [28.47, 33.75], again 5/0/0 and $p = 0.03125$. The -0.97-point difference between model sizes is descriptive only; two sizes do not identify a scaling law. P43 supplies a separate same-family persistence check under the larger `large240-v2` inventory, reported below rather than pooled with these earlier cells.

The cross-family replication freezes the five semantic worlds, three training seeds, matched conditions, endpoint, and world-level inference while changing the model family. Llama-3.2-1B yields per-world effects of +14.58, +14.58, +13.89, +4.86, and +29.86 points. The mean is +15.56 [8.75, 23.61], with a hierarchical interval of [7.08, 24.58], 5/0/0, and exact $p = 0.03125$. All five independently generated worlds favor DIVERSE under the prespecified replication criterion. Its smaller magnitude than the Qwen estimates is reported descriptively; three model checkpoints do not identify a scale or family effect.

4.2 Large-regime and 8B robustness

RQ1 extension: Does the effect survive a larger inventory and an 8B model? The final large-regime robustness matrix extends the controlled Forward endpoint beyond the original sub-2B, 24-fact setting. Qwen3-0.6B with 240 facts per world yields a +21.10-point mean effect across five worlds (5/5 positive, strong robustness). Llama-3.1-8B with 24 facts yields +2.92 points (4/5 positive, directional robustness), while the 240-fact Llama-3.1-8B cell yields +11.38 points (5/5 positive, strong robustness). These historical cells are analyzed separately at the world level and do not identify a scaling, family, architecture, tokenizer, or interaction effect.

P43 then holds the `large240-v2` artifact, worlds, seeds, conditions, endpoint, and analysis unit fixed within Qwen3 while adding the 1.7B and 8B Base strata. Under this matched 240-fact design, the mean Forward effects are +15.50 points at 1.7B and +9.375 points at 8B; all five worlds are

positive in each stratum and the exact one-sided sign-flip test is $p = .03125$ in both. Both strata pass the frozen strong-persistence gate, forming a same-family persistence ladder rather than a fitted scaling trajectory.

4.3 NATURAL WORDING OVER SYNTHETIC FACTS

RQ2: Is the result tied to generator-style probes? With manually verified natural held-out templates, Qwen3-1.7B produces +34.72 points [31.94, 37.50], 5/0/0, and $p = 0.03125$. The result rules out a narrow explanation in which diversity only fits the original generator's probe phrasing. It does not change the underlying facts, relations, checkpoints, or model family, and therefore tests wording/interface robustness rather than real-data validity.

4.4 NATURAL-SOURCE TRANSFER AND BOUNDARIES

RQ3: Does the pattern remain detectable when facts come from real source-QA? These experiments replace programmed synthetic facts with real source-QA factual content while retaining the controlled SINGLE-REPEAT/DIVERSE intervention. They use weaker-control natural-source artifacts: expressions are pipeline-verified silver and per-fact native-token totals are approximate. They therefore provide external-validity evidence rather than a second causal estimate.

The five-domain natural-source panel compares two Qwen sizes in a balanced three-seed analysis. Finance containment is $+3.56$ and $+3.44$ points for 0.6B and 1.7B, respectively, while held-out answer-token NLL changes by -0.2930 and $+0.0336$. Materials, Life science, and Medical show positive containment under both sizes, but their NLL directions differ by size in Life science and Medical. Strict normalized EM remains at or effectively at floor. These are bounded natural-source observations rather than a pooled natural treatment estimate.

The Earth-science extension yields containment changes of -4.25 and $+0.17$ points for 0.6B and 1.7B, respectively, with held-out NLL changes of -0.2362 and -0.0239. Across all five domains, generated containment is positive under both sizes in four domains, whereas NLL is lower under DIVERSE in seven of ten domain–model strata and in both sizes for only Earth-science and Materials. Strict normalized EM remains at floor. Thus, the broader panel shows model- and domain-conditioned natural behavior rather than a uniform transfer pattern.

Earlier source-QA and constructibility studies provide complementary boundaries. A matched source-QA confirmation missed its prespecified criterion, while stricter Earth-science and Natural Questions constructions stopped before treatment-model execution (Appendix D; Appendix C). These results are boundary evidence rather than natural treatment replications.

4.5 MECHANISM STRESS TESTS

RQ4: Does the Forward effect reduce to relation preservation? Two mechanism stress tests examine this question separately from the main DIVERSE–SINGLE-REPEAT estimand. The first test's frozen relation-completeness contrast is mixed or negative across four model strata, while its breadth bridge is not a relation-preservation test. The second uses a native-tokenizer, relation-competitive 2×2 factorial design with two model strata, five worlds, and three training seeds. Its primary endpoint is a relation-discrimination margin, with world as the inference unit within each model stratum. At the final checkpoint, the complete-minus-partial margin is positive in all five Llama worlds (descriptive mean $+7.99$ points) but positive in only two of five Qwen worlds and negative on average (-1.28 points; world 709 -10.09). The result is therefore model/world-conditioned under this synthetic design and does not support a pooled or universal relation-preservation mechanism. It also does not identify whether the heterogeneity arises from architecture, tokenizer, pretraining data, optimization, or another model-stratum difference.

This stress test narrows rather than overturns the main result. The robust matched expression-allocation effect remains supported for the frozen Forward endpoint; what is rejected is the stronger interpretation that a single relation-preservation account explains it across model strata. No unit was retried, replaced, excluded, or pooled.

4.6 REVERSE TRANSFER

Reverse transfer is unsupported in the tested synthetic Llama setting: effects are near zero or negative and remain outside the Forward decision criterion (Appendix D).

5 DISCUSSION

5.1 WHAT THE CONTROLLED EFFECT ESTABLISHES

Within the controlled synthetic construction, the experiments isolate an effect that most prior natural-language and data-level comparisons do not. Facts, four rows, eight serialized answer mentions, loss-bearing answer-token exposure, tokenizer budgets, and global template frequencies are matched; only expression allocation changes. Headline replications yield 15.56–34.72-point Forward effects, while larger-inventory cells add +21.10, +2.92, and +11.38 points. P43 further shows +15.50 and +9.375 points at Qwen3-1.7B and Qwen3-8B under the matched 240-fact design, with five positive worlds in each stratum. The bridge and P43 ladder are directional/descriptive rather than a scaling trajectory; descriptively smaller P43 means cannot establish attenuation because ceiling effects and improved SINGLE confound that comparison, but do establish persistence through tested 8B.

This is stronger than a dataset-level association but narrower than a mechanism claim: the bundled intervention does not isolate lexical, syntactic, interface, or interaction components, and broader surface-form coverage of the held-out interface may mediate it. Plausible accounts include reduced dependence on one training interface and model-specific interactions with pretrained representations; the relation-preservation stress test does not support a universal explanation. The intermediate-breadth test likewise does not support a monotonic $K = 1 < K = 2 < K = 4$ law, so expression count alone is not a sufficient scalar summary and the governing mechanism remains unresolved (Appendix F).

5.2 NATURAL-SOURCE EVIDENCE REVEALS AN ENDPOINT BOUNDARY

Natural-source evidence exposes an endpoint and heterogeneity boundary. Across domains and model sizes, generated containment and teacher-forced answer-token NLL are heterogeneous, while strict exact match remains at floor; these studies are therefore not a uniform behavioral or likelihood replication, and NLL does not establish retrieval or free generation. Contrasts are informative rather than pooled estimates, and auxiliary familiarity diagnostics in Appendix D are not mediation tests.

5.3 IMPLICATIONS FOR MEASURING TRAINING EXPOSURE

Fact frequency and expression count are incomplete summaries of factual training exposure: fixed within-fact exposure allocation can change held-out callability, but not uniformly across endpoints or domains. Future analyses should track source identity, per-fact allocation, and endpoint; this is a measurement implication, not a validated production recipe.

6 CONCLUSION

In the controlled synthetic setting, reallocating fixed exposure across faithful expressions changes held-out Forward callability, with P43 showing persistence through the largest tested Qwen3-8B checkpoint under the matched 240-fact design. The five-domain natural panel remains heterogeneous across endpoints, models, and domains. The evidence therefore establishes the controlled effect while bounding natural-source, beyond-tested-capacity, scaling, mechanism, and bidirectional claims.

LIMITATIONS

We cover two model families, one programmed generator, and synthetic factual worlds. P43 compares Qwen3 1.7B/8B within family, not randomized parameter counts; beyond-8B persistence and causal attenuation with scale are untested. Natural silver rewrites have approximate token totals; containment is not recall, strict EM is at floor, and endpoints vary. Reverse, internal mechanism, full document-to-model survival, production recipe remain unsupported.

AI USE STATEMENT

Generative AI tools assisted with research planning and hypothesis refinement, literature search and citation checking, protocol and experiment-design feedback, software engineering and debugging, result-interpretation support, manuscript editing, figure/layout assistance, and code/document inspection. All AI-assisted suggestions and code changes were reviewed by the authors and validated against frozen configurations, machine-generated artifacts, tests, and source references. Result-bearing experiments and statistical outputs were generated by frozen code and were not selectively modified to obtain desired outcomes. The authors take responsibility for the final paper and artifacts.

REFERENCES

Ekaterina Alimaskina, Denis Shveykin, Gleb Molodtsov, Igor Shalygin, Alexey Kadeishvili, and Aleksandr Beznosikov. Self-study reconsidered: The hidden fragility of learning from self-generated QA. *arXiv preprint arXiv:2606.32002*, 2026. URL https://arxiv.org/abs/2606.32002.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 2023. URL https://arxiv.org/abs/2309.14316.

Tina Behnia, Puneesh Deora, and Christos Thrampoulidis. Facts in stats: Impacts of pretraining diversity on language model generalization. *arXiv preprint arXiv:2510.16096*, 2025. URL https://arxiv.org/abs/2510.16096.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*, 2024. URL https://arxiv.org/abs/2410.15226.

Zhen Huang, Yikun Wang, Shijie Xia, and Pengfei Liu. DataOrchestra: Learning to orchestrate per-example curation of pretraining data. *arXiv preprint arXiv:2607.24717*, 2026. URL https://arxiv.org/abs/2607.24717.

Minjun Kim, Inho Won, Hyeonseok Lim, MinKyu Kim, Junghun Yuk, Wooyoung Go, Jongyoul Park, Jungyeul Park, and KyungTae Lim. TELLME: Test-enhanced learning for language model enrichment. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1655–1677, 2026. doi: 10.18653/v1/2026.findings-eacl.84. URL https://arxiv.org/abs/2608.11788.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577.

Hao Liang, Qifeng Cai, Yibo Lin, Jianzhuo Du, Qifeng Xia, Sizhe Qiu, Linzhuang Sun, Meiyi Qiang, Zhaoyang Han, Xiaochen Ma, Bohan Zeng, Ruichuan An, Conghui He, and Wentao Zhang. DataPrep-Bench: Benchmarking LLMs as training data preparators. *arXiv preprint arXiv:2607.20465*, 2026. URL https://arxiv.org/abs/2607.20465.

June M. Liu, Shaomian Zheng, He Cao, Dingnan Jin, Qing Cui, and Jun Zhou. Improving cross-format robustness in language models with multi-format training. *arXiv preprint arXiv:2606.11643*, 2026. URL https://arxiv.org/abs/2606.11643.

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. *arXiv preprint arXiv:2305.16264*, 2023. URL https://arxiv.org/abs/2305.16264.

Joel Niklaus, Atsuki Yamaguchi, Michal Štefánik, Guilherme Penedo, Hynek Kydlíček, Elie Bakouch, Lewis Tunstall, Edward Emanuel Beeching, Thibaud Frere, Colin Raffel, Leandro von Werra, and Thomas Wolf. How can we synthesize high-quality pretraining data? a systematic study of prompt design, generator model, and source data. *arXiv preprint arXiv:2604.13977*, 2026. URL https://arxiv.org/abs/2604.13977. Accepted to COLM 2026.

Charles O'Neill. Can a language model learn facts continually in its weights? *arXiv preprint arXiv:2607.11020*, 2026. URL https://arxiv.org/abs/2607.11020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*, 2024. URL https://arxiv.org/abs/2409.07431.

Jiang Zhou, Yunhao Wang, Xing Wu, Tinghao Yu, and Feng Zhang. WRAP++: Web discovery amplified pretraining. *arXiv preprint arXiv:2604.06829*, 2026. URL https://arxiv.org/abs/2604.06829.

The following registered example makes the treatment contrast inspectable without adding a new construction. It answers what changes within a fact, what remains matched, and which held-out probe is evaluated; it is not an additional result-bearing example.

A VERBATIM FROZEN TREATMENT EXAMPLE

Frozen fact	SINGLE-REPEAT	DIVERSE
`Sorenionis-Neral born_in Ravinimon`	*Which birthplace does the archive record for Sorenionis-Neral?* $\times 4$	*Which birthplace does the archive record for Sorenionis-Neral?* *Which location does the registry link to the birth of Sorenionis-Neral?* *What birthplace location is assigned to Sorenionis-Neral in the archive?* *What location is associated with Sorenionis-Neral's birth in the registry?*
Shared fields	Every row includes `Answer: Ravinimon` and `Solution: The archive links Sorenionis-Neral to Ravinimon.`	
Held-out Forward	*Identify the location connected to the birth record for Sorenionis-Neral.* → `Ravinimon`	

Table 4: Exact registered world-101 strings underlying the schematic in Figure 1. No treatment or probe wording was redesigned for presentation.

B FROZEN EXPERIMENTAL MATRIX

The matrix records the evidence units and constructibility stops used throughout the paper. P43 is listed separately from historical robustness cells because its two Qwen3 strata form a same-family persistence comparison, not a pooled scaling analysis.

Exp.	Runs	Evals	Unit
EXP-010	20	20	10 seeds
EXP-011	30	30	5 worlds
EXP-012	30	30	5 worlds
EXP-013	0	30	5 worlds
EXP-014	2	2	27 parents
EXP-015	10	10	5 seeds
EXP-016	0	0	96-parent CPU gate
EXP-017	30	30	5 worlds
EXP-019	180	180	4 strata, 5 worlds
EXP-020	120	480	2 strata, 5 worlds
P37-C	90	90	3 cells, 5 worlds
P43	60	60	2 Qwen3 strata, 5 worlds
P32	0	0	16-unit CPU Gate

Table 5: Completed matrices and constructibility Gates. EXP-013 is evaluation-only and reuses EXP-012 checkpoints. EXP-017 also includes five base evaluations. EXP-019 includes 180 training runs and checkpoint evaluations across four separate model strata; EXP-020 includes 120 training runs and 480 checkpoint evaluations across two separate model strata; P37-C includes 90 training checkpoints and 90 checkpoint evaluations across three independently analyzed cells. P32 performs no target-model inference, training, or evaluation.

C NATURAL QUESTIONS CONSTRUCTIBILITY FUNNEL

P32 freezes 16 source-linked Natural Questions units and requests eight candidates for each of three expression intents. All 48 requests complete and yield 176 parsed candidates; 51 pass deterministic filters. Intent-level pass counts are 9, 18, and 24, but no unit has at least one pass for all three intents, making the all-intent upper

bound 0/16 against the preregistered 12/16 threshold. The futility rule therefore fires before semantic judges, native-token tuple search, full-data expansion, target Base screening, treatment training, or outcome analysis. No unit is replaced or manually rescued.

D BOUNDARY EVIDENCE

Evidence	Setting	Endpoint / unit	Observation	Decision	What it does not show
Finance natural-source extension	Finance source-QA	containment / seed	+1.9 pp; 4/5 positive; NLL 5/5 lower	Natural-source boundary evidence	natural causal law
Earth-science extension	Earth-science source-QA	containment / seed	-4.25 pp; 1/3 positive; NLL 3/3 lower	Cross-domain heterogeneity	natural causal law
Source-QA diagnostic	Real source-QA	strict EM / parent	-1.85 pp; 0/26/1	Diagnostic boundary	harmful diversity
Matched source-QA confirmation	Real source-QA	containment / seed	+5.12 pp; 4/0/1; $p = .0625$	Criteria not met	natural confirmation
Earth-science constructibility	Earth science	constructibility / parent	46/96 pass	Data boundary	model effect
Natural Questions constructibility	Natural Questions	constructibility / fixed unit	0/16 all-intent upper bound	Preregistered stop	natural treatment effect
Reverse transfer	Synthetic Llama	Reverse EM / world	-1.67 pp; 0/1/4	Unsupported	bidirectional transfer
Relation-completeness stress test	Four model strata	completeness / world-stratum	mixed or negative	Mechanism boundary	pooled mechanism
Relation-discrimination stress test	Qwen / Llama	discrimination / world-stratum	-1.28 / +7.99	Model-conditioned	architecture cause

Table 6: Boundary evidence, separated by endpoint and inference unit. Rows are not pooled or plotted on a common effect axis. Natural-source rows use weaker-control artifacts; constructibility rows stop before treatment-model execution; relation stress tests keep model strata separate.

The matched source-QA confirmation did not meet its prespecified confirmation criterion: its mean was $+5.12$ points (below $+10$), four of five seed effects were positive, and exact $p = .0625$ rather than the criterion's $p = .03125$.

P43 uses a separate frozen strong-persistence criterion: mean world effect $> +5$ points, five positive worlds, and exact one-sided sign-flip $p = .03125$ within each model stratum. It is not evaluated with the older $+10$-point criteria, and it does not support a scaling law or a pooled model effect.

Possible natural-source explanations include answer length, irregular relations, rewrite utility, and initial familiarity. Finance Base familiarity slightly favors DIVERSE (-0.0236 NLL), whereas auxiliary gpt-oss base-only evidence favors SINGLE-REPEAT ($+0.00396$); these descriptive contrasts are not mediation tests.

E STUDY-SPECIFIC DECISION CRITERIA

For the Qwen five-world synthetic tests, the frozen decision criteria require a minimum mean effect of +10 percentage points, a bootstrap lower bound above zero, and an exact one-sided sign-flip $p \leq .05$. EXP-017 deliberately has no minimum-effect threshold: its strong tier requires a positive mean, five positive worlds, and exact $p = .03125$; directional support requires a positive mean and at least four positive worlds. EXP-015 requires mean effect ≥ 10 points, five strictly positive seed effects, exact $p = .03125$, a complete 5×2 matrix, and no seed replacement or selective exclusion. These criteria were fixed before the corresponding result-bearing runs and are not applied across heterogeneous endpoints or model/domain strata.

F P39 DOSE-RESPONSE BOUNDARY

P39 added only the frozen $K = 2$ condition to the existing $K = 1$ (SINGLE-REPEAT) and $K = 4$ (DIVERSE) checkpoints. The table reports world-level held-out Forward exact-match means, with three training seeds nested within each world and model stratum. All 30/30 new training runs and 30/30 final evaluations completed without result-bearing retries, replacements, or exclusions. The prospective interpretation rule was applied before

treating the adjacent effects as evidence for a graded response. Here, "Strict order" is "yes" exactly when the within-world means satisfy $K_1 < K_2 < K_4$; the column is reported per world and is not a pooled model-level count.

Model	World	$K = 1$	$K = 2$	$K = 4$	$K2 - K1$	$K4 - K2$	Strict order
Qwen3-0.6B	101	82.36	83.06	99.93	+0.69	+16.88	yes
	203	74.72	79.51	99.93	+4.79	+20.42	yes
	307	80.35	77.78	97.92	-2.57	+20.14	no
	401	76.81	81.60	100.00	+4.79	+18.40	yes
	509	78.06	49.51	100.00	-28.54	+50.49	no
Llama-3.2-1B	101	85.42	86.53	100.00	+1.11	+13.47	yes
	203	85.42	84.79	100.00	-0.62	+15.21	no
	307	86.11	90.83	100.00	+4.72	+9.17	yes
	401	79.86	94.17	84.72	+14.31	-9.44	no
	509	68.75	66.81	98.61	-1.94	+31.81	no

Table 7: P39 world-level Forward exact-match percentages for the frozen $K = 1/K = 2/K = 4$ comparison. Adjacent effects are percentage points; rows are descriptive within each model stratum and are not pooled across models.

G EXP-015 SEED-LEVEL RESULTS

This table answers whether the matched source-QA containment direction is consistent across the five preregistered seeds. It reports the complete seed-level matrix; it is descriptive boundary evidence and is not pooled with synthetic Forward effects.

Seed	SINGLE-REPEAT	DIVERSE	Effect
20260816	69.77	79.07	+9.30
20260817	70.93	80.23	+9.30
20260818	74.42	79.07	+4.65
20260819	75.58	81.40	+5.81
20260820	80.23	76.74	−3.49

Table 8: Normalized answer containment (percent) in EXP-015.

H P38R NATURAL-SOURCE SEED RESULTS

The natural-source extensions use separate domain strata and the same two-condition training recipe. The entries below are copied from the frozen aggregate reports; containment is parent-macro and NLL is held-out answer-token weighted. Strict normalized exact match is zero in every listed cell.

Finance N2 (300 facts; Qwen3-0.6B)						
Seed	SINGLE-REPEAT	DIVERSE	Cont. D−S	NLL_S	NLL_D	NLL D−S
271	57.00	59.50	+2.50	1.4410	0.9870	-0.4540
277	63.17	63.83	+0.67	1.3669	1.1666	-0.2003
281	54.00	61.50	+7.50	1.3006	1.0759	-0.2247
283	64.83	62.17	-2.67	1.4647	1.0330	-0.4316
293	58.00	59.50	+1.50	1.4430	1.0868	-0.3562
Mean	59.40	61.30	+1.90	1.4032	1.0699	-0.3334

Table 9: Finance N2 seed-level containment percentages and held-out answer-token NLL.

Earth-science N2 (200 facts; Qwen3-0.6B)						
Seed	SINGLE-REPEAT	DIVERSE	Cont. D−S	NLL_S	NLL_D	NLL D−S
271	76.00	66.50	-9.50	1.5193	1.0562	-0.4631
277	68.75	64.25	-4.50	1.1748	1.1097	-0.0651
281	67.00	68.25	+1.25	1.4290	1.2487	-0.1804
Mean	70.58	66.33	-4.25	1.3744	1.1382	-0.2362

Table 10: Earth-science N2 seed-level containment percentages and held-out answer-token NLL.

I P40 NATURAL-SOURCE PANEL SEED RESULTS

P40 adds the three new domains and the larger Qwen3-1.7B stratum for the existing domains. The table reports the balanced three-seed comparisons used for the main-text natural-source summary; containment differences are percentage points and NLL differences are DIVERSE−SINGLE. Strict normalized EM is at or below 0.08% in these cells.

Domain	Model	Seed	Cont. D−S	NLL D−S	NLL lower?
Earth-science	Qwen3-1.7B	271	-1.75	-0.0496	yes
Earth-science	Qwen3-1.7B	277	-1.00	-0.0059	yes
Earth-science	Qwen3-1.7B	281	+3.25	-0.0164	yes
Finance	Qwen3-1.7B	271	+5.00	+0.0525	no
Finance	Qwen3-1.7B	277	+5.00	-0.0289	yes
Finance	Qwen3-1.7B	281	+0.33	+0.0772	no
Life science	Qwen3-0.6B	271	-3.50	+0.1276	no
Life science	Qwen3-0.6B	277	+6.50	-0.1588	yes
Life science	Qwen3-0.6B	281	+4.00	+0.1242	no
Life science	Qwen3-1.7B	271	+4.75	-0.2196	yes
Life science	Qwen3-1.7B	277	+3.50	+0.0043	no
Life science	Qwen3-1.7B	281	+0.50	-0.1688	yes
Materials	Qwen3-0.6B	271	+8.25	-0.0670	yes
Materials	Qwen3-0.6B	277	+3.25	-0.2351	yes
Materials	Qwen3-0.6B	281	+3.50	-0.1969	yes
Materials	Qwen3-1.7B	271	+4.25	-0.1530	yes
Materials	Qwen3-1.7B	277	+2.50	-0.3349	yes
Materials	Qwen3-1.7B	281	+3.50	-0.3850	yes
Medical	Qwen3-0.6B	271	-5.25	+0.1702	no
Medical	Qwen3-0.6B	277	+2.75	+0.0046	no
Medical	Qwen3-0.6B	281	+2.75	+0.0743	no
Medical	Qwen3-1.7B	271	+10.50	-0.0472	yes
Medical	Qwen3-1.7B	277	+4.50	-0.0418	yes
Medical	Qwen3-1.7B	281	+7.00	-0.0644	yes

Table 11: P40 seed-level natural-source results for the added model/domain strata. No rows are pooled across domains or models; the table is descriptive and reports the complete added matrix.

J P43 CAPACITY-PERSISTENCE PER-WORLD RESULTS

The table below reports the complete P43 world-level effects, with the three training seeds nested within each world. The two columns are separate Qwen3 model strata under the same frozen `large240-v2` 240-fact design; the table is intended to inspect directional consistency, not to fit or imply a capacity trend. Both strata have five positive worlds and exact one-sided sign-flip $p = .03125$; paired bootstrap intervals are descriptive.

World	Qwen3-1.7B	Qwen3-8B
101	+15.97	+6.53
203	+18.26	+8.96
307	+9.86	+3.26
401	+22.15	+17.71
509	+11.25	+10.42
Mean	+15.50	+9.375
Bootstrap 95% (descriptive)	[+11.64,+19.36]	[+5.35,+14.01]

Table 12: P43 complete world-level held-out Forward cross-interface exact-match effects in percentage points (DIVERSE−SINGLE). Each world averages three training seeds; model strata are analyzed separately.

K Extended Nearest-Work Comparison

Table 13 extends the compact main-text comparison. Cells record only controls or analysis units explicitly supported by the cited paper; NR means not reported for that comparison. A token- or instance-matched dataset comparison is not automatically a fixed per-fact comparison.

Work	Stable unit	Fixed exposure	Answer matched	Budget matched	Held-out callability	Source linked	Primary scope
Allen-Zhu–Li (Allen-Zhu & Li, 2023)	yes	no	no	no	yes	synthetic; near-real	Biography augmentation, extraction, and QA transfer
EntiGraph (Yang et al., 2024)	source corpus	no	NR	selected token ablation	source QA	yes	Document synthesis for continued pretraining
TELLME (Kim et al., 2026)	no	no	no	total tokens for one baseline	benchmark	yes	Domain CPT with explanatory QA
WRAP++ (Zhou et al., 2026)	relation; source	no	NR	token ablations	benchmark	yes	Cross-document relational QA synthesis
FinePhrase (Niklaus et al., 2026)	document	no	NR	dataset-scale analyses	benchmark	yes	Generator, prompt, and source factors in rephrasing
FormatMix (Liu et al., 2026)	source item	no	no	training instances	yes	natural bench-marks	Cross-format robustness
Facts in Stats (Behnia et al., 2025)	atomic pair	no	implicit target occurrence	fixed steps/length	yes	no	Abstract template–fact OOD generalization
O'Neill (O'Neill, 2026)	invented fact	no	no	opt. steps	yes	no	Fact use and continual retention
Self-Study (Alimaskina et al., 2026)	evidence span	no	no	generation budget	pre-training audit	yes	Evidence selection and QA-generation fragility; no downstream treatment comparison
DataPrep-Bench (Liang et al., 2026)	dataset; document	no	NR	task-specific	benchmark	yes	End-to-end data-preparation evaluation
DataOrchestra (Huang et al., 2026)	example; chunk	no	NR	corpus tokens	benchmark	yes	Learned per-example curation policy
Ours	fact	yes	yes	rows, optimizer, tokens	yes	tested; not established	Matched within-fact allocation and boundaries

Table 13: Extended construct comparison. "Yes" denotes the unit used by that paper, not semantic equivalence with our fact definition.

L Synthetic Construction and Token Audit

The model-free generator creates three programmed relations (`born_in`, `made_of`, and `founded_in`), eight facts per relation, and nonce subject/object identities from a frozen seeded vocabulary. Each fact has a stable identifier independent of its wording. The frozen registry contains four training templates plus disjoint direct, paraphrase, cloze, and Reverse probes per relation. All training rows use the same three-field serialization (Question, Answer, Solution). The canonical answer appears once in Answer and once in Solution; hence

four rows produce eight serialized answer mentions per fact. Full causal language modeling leaves both spans loss-bearing. The renderer rotates the canonical SINGLE-REPEAT template over facts within each relation; consequently both conditions have identical corpus-level relation–template marginals even though one fact sees one template four times and the other sees all four once.

Token audits use each model's native tokenizer with special-token insertion disabled. They sum the full serialized training text, not whitespace proxies. A full-context character-offset audit additionally counts native-token positions overlapping every answer span; all 384 headline fact–artifact pairs match between conditions. The five Qwen artifacts have zero total condition imbalance, and Qwen3-0.6B and Qwen3-1.7B produce identical token IDs on all 1,440 frozen inputs. The Llama-specific semantic copies also match exactly within condition: for worlds 101, 203, 307, 401, and 509, SINGLE-REPEAT/DIVERSE totals are respectively 3904/3904, 3848/3848, 3856/3856, 3872/3872, and 3840/3840. Tokenizer-specific reconstruction changes token accounting, not facts, templates, worlds, seeds, conditions, or endpoints.

M STATISTICAL PROCEDURE

M.1 OBSERVED POINT ESTIMATES

The frozen evaluator first builds condition-specific maps keyed by `probe_id`, requires identical paired key sets, sorts those IDs, and selects the probe types belonging to the endpoint. For the selected rows it computes a flat arithmetic mean within each run, then forms the paired DIVERSE−SINGLE effect. For world w, seed s, and condition c, with selected probe set $P_{w,s}$,

$$\bar{Y}_{w,s}(c) = \frac{1}{|P_{w,s}|} \sum_{i \in P_{w,s}} Y_{w,s,i}(c), \qquad d_{w,s} = \bar{Y}_{w,s}(D) - \bar{Y}_{w,s}(S), \tag{2}$$

$$d_w = \frac{1}{|S_w|} \sum_{s \in S_w} d_{w,s}, \qquad \Delta_{\text{world}} = \frac{1}{|W|} \sum_{w \in W} d_w. \tag{3}$$

For the fixed-world seed replication, w is suppressed and $\Delta_{\text{seed}} = |S|^{-1} \sum_{s \in S} d_s$. Thus probes and facts are aggregated within the preregistered seed/world units and are never treated as IID top-level observations. In the balanced synthetic Forward endpoint, three relations, eight facts per relation, and two selected probes per fact make this flat mean numerically equal to a fact-within-relation mean followed by a relation mean; the equality is a property of the balanced counts, not the executed point-estimator hierarchy.

M.2 EXACT SIGN-FLIP TEST

For top-level paired effects $d_1, \ldots, d_n$, the observed statistic is $T_{\text{obs}} = n^{-1} \sum_j d_j$. For $n \leq 20$, the frozen implementation enumerates all 2^n sign assignments under the sign-symmetry null and reports the one-sided `greater` p-value

$$p = \frac{\#\{b : T_b \geq T_{\text{obs}} - 10^{-15}\}}{2^n}, \qquad T_b = \frac{1}{n} \sum_j b_j d_j,$$

where $b_j \in \{-1, +1\}$. For $n > 20$, the unused fallback draws 100,000 assignments with the fixed random seed 20260806 and uses the add-one correction. With five world-level units, the exact sign-flip test has coarse resolution; these replications are therefore interpreted jointly through effect magnitude and directional consistency rather than as high-precision population estimates.

M.3 BOOTSTRAP INTERVALS

Conditions are paired before resampling. The fact-cluster bootstrap groups paired probe differences by (relation_type, fact_id), averages probes within fact, resamples the observed fact means with replacement separately within each relation, concatenates the strata, and takes an arithmetic mean. The seed bootstrap resamples the observed seed-effect vector with replacement and preserves its length. The world bootstrap does the same for the observed world-effect vector. The hierarchical bootstrap resamples worlds, then training seeds within each sampled world, then observed fact means within each fixed relation label; it averages facts within relation, relations, seeds, and sampled worlds in that order. Pairing is preserved because resampling acts on precomputed DIVERSE−SINGLE differences, and relation labels are fixed rather than resampled.

All shared intervals use the 2.5th and 97.5th percentiles. The shared `_percentile` helper sorts draws and linearly interpolates at $(n-1)p$. These intervals are descriptive wherever the manuscript labels them descriptive. EXP-015 uses its separate discrete order-statistic implementation rather than the shared interpolated helper.

Analysis	Fact-cluster B/seed	Seed B/seed	World B/seed	Hierarchical B/seed
EXP-010	10,000 / 20260803	100,000 / 20260806	–	–
EXP-011	10,000 / 20260803 + w	10,000 / 20260809 + w	100,000 / 20260807	100,000 / 20260808
EXP-012	10,000 / 20260803 + w	10,000 / 20260809 + w	100,000 / 20260810	100,000 / 20260811
EXP-013	not emitted	not emitted	100,000 / 20260807	100,000 / 20260808
EXP-017	10,000 / 20260817 + w	10,000 / 20260816 + w	100,000 / 20260814	100,000 / 20260815
EXP-019	10,000 / 20260803 + w	10,000 / 20260809 + w	10,000 / 20260827	10,000 / 20260828
EXP-015	–	100,000 / 20260813	–	–
P40 natural panel	–	–	–	–

Table 14: Frozen bootstrap constants. Here w denotes the deterministic world offset added by the corresponding helper. The P40 aggregator performs no inferential bootstrap or sign-flip test; it reports descriptive arithmetic means over three seed-level values per domain/model stratum.

M.4 Natural answer-token likelihood

For a held-out probe, the evaluator forms x = prompt + `"\n"` + answer-prefix + answer, where `answer-prefix` is the evaluator's literal answer-prefix marker; for a serialized training row it locates the final literal answer marker with the canonical answer. Native tokenization uses `add_special_tokens=False`. Character-offset overlap selects the native-token positions T_r that intersect the canonical answer span, and the answer-token loss is

$$L_r^{\text{ans}} = -\frac{1}{|T_r|} \sum_{t \in T_r} \log p_\theta(y_t \mid x, y_{<t}). \tag{4}$$

The loss is token-weighted within a run; P40 then takes arithmetic means across the three seed-level values within each domain/model stratum. Lower is better. This is a teacher-forced target-answer likelihood diagnostic, not a generated-behavior endpoint; the evaluator also computes `fact_macro_answer_nll`, but that field is not the P40 headline quantity.

M.5 Endpoint-specific boundaries

The equations above describe the controlled synthetic Forward-EM estimand. Natural-source containment, strict EM, and answer-token NLL use their endpoint-specific DIVERSE−SINGLE summaries. Model strata and Forward, Reverse, containment, constructibility, and discrimination endpoints are analyzed separately and are never pooled.