Facts in Stats: Impacts of Pretraining Diversity on Language Model Generalization

Tina Behnia [1] **Puneesh Deora** [1] **Christos Thrampoulidis** [1]

Abstract

Language models are pretrained on sequences that blend statistical regularities, which make text fluent, with factual associations between specific tokens, which encode knowledge of facts. While recent works suggest that generalization depends critically on the interaction of these two streams, such as the diversity of the contexts in which facts appear, these effects remain difficult to study systematically. This paper introduces a flexible synthetic testbed that combines a statistical stream of generic tokens with an abstract factual stream of source-target token pairs, enabling fine-grained control over their interaction, such as their composition into a sequence (contextual structure) or the level of context diversity carrying the facts at training time. Through controlled experiments, we find that higher contextual diversity delays in-distribution factual learning, while low diversity can harm out-of-distribution generalization in ways that depend on the contextual structure. As a result, the optimal diversity level depends on the training budget. Beyond factual recall failures, we also identify failures in statistical generalization as we study how the interplay between contextual design and diversity level impacts different aspects of generalization. Furthermore, through a series of controlled interventions on the model components, we trace failure in different aspects of generalization to distinct optimization bottlenecks, highlighting the importance of the embedding and unembedding layers. Overall, our synthetic framework allows us to isolate effects that would be confounded in large-scale studies, offering a controlled testbed for future investigations.

[1]Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada. Correspondence to: Tina Behnia <tina.behnia@ece.ubc.ca>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Modern transformer-based language models (LMs), trained on massive corpora of natural-text sequences, simultaneously learn to generate contextually plausible sequences that follow statistical and linguistic patterns, while also learning significant amount of real-world knowledge. Their parameters effectively become an implicit knowledge base that can rival structured knowledge bases in question-answering and recall tasks (Petroni et al., 2019; Roberts et al., 2020; Dai et al., 2022; Allen-Zhu & Li, 2024; Akyurek et al., 2022; Yang et al., 2025; Mallen et al., 2023; Meng et al., 2022). This dual capability is remarkable precisely because the LM receives no explicit fact supervision during training: it never encounters knowledge graphs, labeled facts, or any explicit differentiation between linguistic and factual content. Yet somehow, it develops the ability to both follow linguistic patterns and encode world knowledge. *How does the conceptually-simple next-token prediction objective enable a language model to implicitly learn both linguistic structure and factual knowledge? Are there inherent trade-offs in acquiring these two types of information?*

Intuitively, the repetition of factual information across varying and diverse linguistic contexts during training, essentially *paraphrases* of the same underlying facts, likely plays a key role in facilitating factual recall from pure next-token prediction supervision. But, this raises several questions:

How frequently must a model encounter a fact during training to reliably learn it, and how does this depend on the diversity *of linguistic contexts (paraphrases) in which the fact appears? Under what training conditions is it possible to isolate factual associations from linguistic structures?* Moreover, *is diversity only possibly impacting factual recall, or can it also impact the model's ability to learn statistical patterns?*

The intriguing factual-recall capability of LLMs has motivated several recent studies probing how models acquire, represent, and recall facts using controlled setups. Allen-Zhu & Li (2024) and very recently Zucchet et al. (2025) study synthetic biography corpora with fixed templates and varying factual associations (e.g., name, age, occupation) to examine how paraphrastic diversity affects factual re-

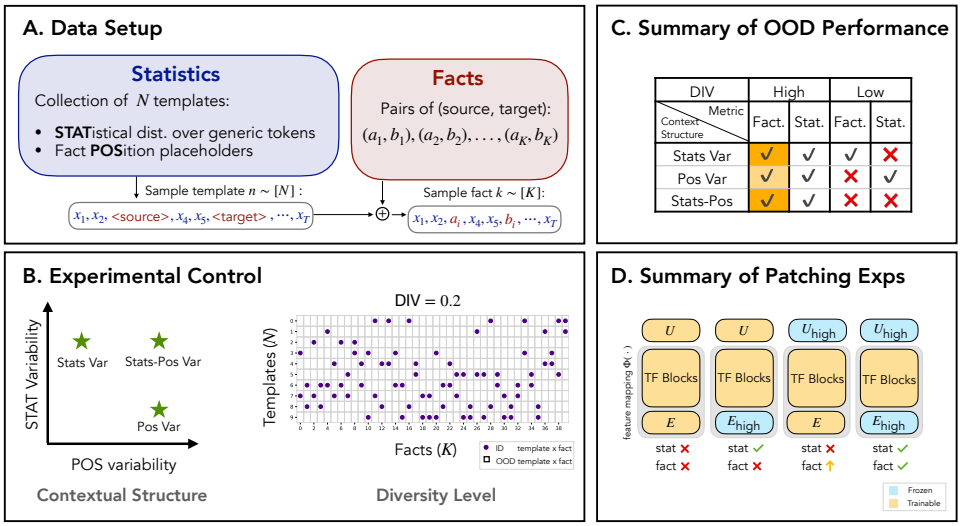


Figure 1. Proposed testbed and summary of key findings. (A) Sequences are generated by combining a **factual stream**, a set of atomic pairs, and a **statistical stream** of templates, each a statistical distribution over sequences with two fact placeholder positions. Final sequences insert a fact into a sequence sampled from a template (Sec. 2). **(B)** We control two key data properties. **(Left) Contextual structure:** The degree of statistical and (fact) positional variation across templates. We explore three structures: **Stats Var** (varied statistics, fixed positions), **Pos Var** (fixed statistics, varied positions), and **Stats-Pos Var** (varied on both axes). **(Right) Diversity level:** The number of unique templates each fact is paired with in the training corpus. The table shows an *exposure matrix*, where dots indicate template-fact pairs used for training (Sec. 2.1-2.2). **(C)** Low diversity impairs out-of-distribution generalization, but the specific aspects affected (factual recall vs. statistical pattern generation) depend on contextual structure. At high diversity, all structures enable both types of generalization, though with varying efficiency (colored boxes: darker means more training iterations needed with higher `DIV`) (Sec. 3). **(D)** While training on low-diversity data impairs statistical and factual generalization, when the embeddings or unembeddings from a high-diversity model are *patched* in, the remaining modules can still learn a generalizing solution even when trained on that same low-diversity data (Sec. 4).

call. While their data setup provides more control than raw natural-text corpora, where linguistic patterns and facts are intricately interleaved, they still operate at considerable scale, requiring transformers with 8-12 layers/8-12 heads/512-768 dimension for empirical analysis. This limits accessibility and makes comprehensive parameter sweeps over controllable variables computationally challenging. Additionally, by employing fixed templates, these setups isolate focus on factual learning while ignoring the model's ability to simultaneously acquire linguistic structure and leaving fine-grained impacts of diversity on learning of both components underexplored.

Motivated by these, we develop a small-scale synthetic testbed for studying at a fine-grained level the interactions between statistical and factual learning streams, specifically as captured in terms of context *diversity*. Our framework offers three key advantages. *Tractability:* Its minimal scale enables comprehensive parameter sweeps and in-depth analysis that would be computationally challenging in large-scale setups. All our experiments use 4-layer/4-head/32-dim transformers, and phenomena can be reproduced even with 1-layer transformers. All experiments are performed on on single Tesla V100-SXM2-GPU(16GB) and can be replicated on the free-version of Google-Colab. *Reproducibility:* Despite its minimalism, our framework reproduces empirical phenomena from large-scale studies, including stage-wise factual learning (Zucchet et al., 2025) and improved factual recall with larger diversity (Allen-Zhu & Li, 2024). *Experimental control:* The explicit statistical stream enables exhaustive exploration of how pretraining data diversity, both in terms of exposure level and structural variety, impacts transformer out-of-distribution generalization under next-token prediction training.

In broader perspective, our methodology mirrors research turning to deliberately small-scale tasks and transformers to study other next-token prediction phenomena, such as in-context learning (Garg et al., 2022; Edelman et al., 2024), grokking (Power et al., 2022; Nanda et al., 2023), and arithmetic reasoning (Lee et al., 2024). Within factual recall research, our data setup bears some abstract similarity to Nichani et al. (2025), who use synthetic tasks to analyze scaling laws and dynamics of factual recall. Yet, their setup lacks statistical learning, any account of diversity effects, and operates in non-autoregressive setting. See App. A for detailed comparison to related works.

1.1. Summary of Methodology and Findings

Our data framework factorizes sequences into a *statistical* stream and a *factual* stream. We model the statistical stream as a mixture of N *template* distributions over token

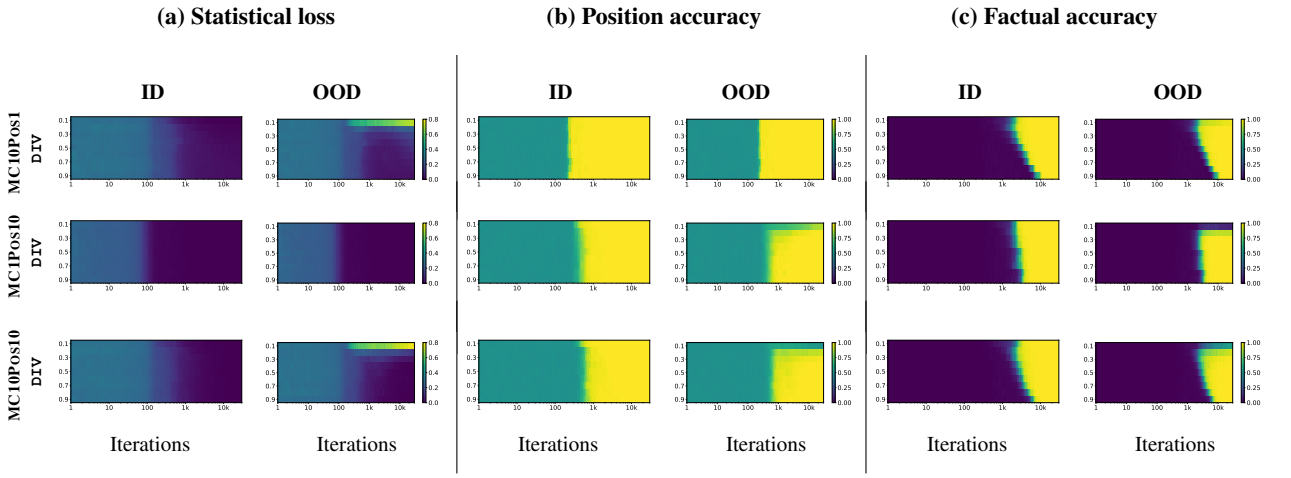


Figure 2. Generalization performance for varying diversity levels and context structures. Heatmaps show **(a)** $\texttt{Loss}_{\text{stat}}$, **(b)** $\texttt{Acc}_{\text{pos}}$ and **(c)** $\texttt{Acc}_{\text{fact}}$ over training iterations (horizontal axis) and diversity level (vertical axis) for the three contextual structures defined in Sec. 2.2: (top) MC10Pos1 with only statistical variations, (middle) MC1Pos10 with only positional variations and (bottom) MC10Pos10 with both types of variations across templates. Each metric is displayed for in-distribution (ID, left column) and out-of-distribution (OOD, right column) fact–template pairs. Note that better statistical performance is marked by lower $\texttt{Loss}_{\text{stat}}$ (darker colors) and better positional and factual accuracy is marked by higher metric values (brighter colors). Increasing diversity can slow down convergence, whereas very low diversity leads to OOD failure on one or multiple metrics. See Sec. 2.3 for metrics and Sec. 3 for details.

sequences. Each template generates a sequence from a template-dependent Markov Chain (MC) distribution and reserves two designated positions for fact insertion; these positions may vary across templates to capture different structural patterns. The factual stream consists of *source–target* token pairs drawn from distinct vocabularies that are inserted in a template's designated positions to generate complete sequences. This gives precise and independent control along two axes: (a) **Diversity level**—how many different templates each fact appears in during training, and (b) **Contextual structure**—how template formats vary in their statistical and positional properties. See Fig. 1-(A,B).

We train small transformers on this data and evaluate generalization through carefully designed prompts, testing three distinct aspects: **(1) Factual recall accuracy**—the model's ability to produce the correct fact target for a given source in the prompt; **(2) Position accuracy**—adherence to the specific positional patterns of the prompt template; and **(3) Statistical accuracy**— adherence to the statistical distribution of the prompt template. We evaluate performance on both **in-distribution (ID)** data, where fact-template combinations match training data, and **out-of-distribution (OOD)** data, where facts appear in novel templates. See Sec. 2 for framework details. We summarize our key findings (also see Fig. 1-(C,D)):

Impact of Diversity. The diversity of the training data has several effects on generalization (Figs. 2-4 and Sec. 3):

1. As diversity increases, achieving ID and OOD factual accuracy becomes slower.
2. Low diversity degrades OOD generalization in factual recall, statistical accuracy, or both, depending on the contextual structure.
3. Data diversity introduces a cost-benefit dynamic: Higher diversity guarantees better performance but only when provided with sufficient training budget. With limited training time, intermediate diversity levels are optimal.
4. The learning dynamics of the two streams (statistical and factual) are coupled: Increased complexity in one stream, delays learning of both streams.

Optimization Bottlenecks. Parameter values that achieve perfect OOD generalization exist regardless of diversity level. However, Diversity shapes the optimization path, determining whether and how quickly training reaches generalizing solutions. Through targeted interventions, we localize the bottlenecks for finding generalizing solutions to specific modules in the transformer architecture (Figs. 5-6 and Sec. 4).

5. We trace back the statistical and positional generalization failures to poor optimization of the attention and embedding modules, while factual-recall failures trace back to the joint optimization of the embedding and unembedding layers.
6. The optimization of the unembedding layer causes the slowdown in factual recall at high diversity; training features alone with fixed optimal unembeddings shows no such slowdown.

2. Data Setup

We consider sequences $\mathbf{x} = (x_1, ..., x_T)$ of length T, produced by mixing two independent sources: a *factual* and a *statistical* stream, with each token x_t belonging to a global vocabulary $\mathcal{V} = [V]$.

Factual stream. We specify K atomic facts $\mathcal{K} := \{(a_k, b_k)\}_{k\in[K]}$ each given as an ordered *source-target* pair of distinct tokens a_k, $b_k \in \mathcal{V}$. The vocabulary of facts is $\mathcal{V}_\mathcal{K} = \{a_k\}_{k\in[K]} \bigcup \{b_k\}_{k\in[K]}$.

Statistical stream (Templates). Background sequences $\mathbf{z} = (z_1, \ldots, z_T)$ to which facts are later embedded are sampled from a mixture of N template distributions: $\mathbf{z} \sim \mathcal{D} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{D}_n$. Formally, a template $\mathcal{D}_n$ is a probability distribution over length-T sequences drawn from the *generic vocabulary* $\mathcal{V}_\mathcal{D} := \mathcal{V} \setminus \mathcal{V}_\mathcal{K}$ of size $V_\mathcal{D} = |\mathcal{V}_\mathcal{D}|$, with two marked positions $\mathbf{I} = (i, j) \in [T] \times [T]$, $i \leq T/2 < j$ reserved for the source/target fact tokens. For concreteness, unless otherwise stated, we assume the generic tokens are produced by a first-order Markov chain (MC). Thus, a template $\mathcal{D}_n$ is defined by: **(i)** a transition matrix $\mathbf{P}_n \in \Delta^{V_\mathcal{D}\times V_\mathcal{D}}$, and **(ii)** a placeholder position pair $\mathbf{I}_n$. We sample from $\mathcal{D}_n$ by generating a length $T-2$ MC sequence from $\mathbf{P}_n$, and reserving the positions $\mathbf{I}_n$ for facts.

Final sequence generation. We draw sequences $\mathbf{x}$ that combine facts with templates. First, we select a template n with (say) position $\mathbf{I} = (i, j)$, and sample $\mathbf{z} \sim \mathcal{D}_n$. Then, we insert a chosen fact (a, b) in positions (i, j) to form the final sequence $\mathbf{x}$, i.e., $x_i \leftarrow a$, $x_j \leftarrow b$ and $x_t \leftarrow z_t$ for $t \notin \{i, j\}$.

2.1. Level of Diversity

The **diversity-level parameter `DIV`** $:= c/N \in (0, 1)$, $c \in [N]$ is such that every fact-pair (a_k, b_k), $k \in [K]$ appears embedded in $\texttt{DIV} \cdot N$ different templates during training. Larger `DIV` indicates a higher diversity level. To describe template-fact pairings during training, we define binary *in-distribution (ID) exposure mask* $\mathbf{M}_{\text{in}} \in \{0, 1\}^{N\times K}$ with $\mathbf{M}_{\text{in}}[n, k] = 1$ only if (a_k, b_k) appears in training in a sequence drawn from $\mathcal{D}_n$. Thus, the k-th column lists all templates in which the k-th fact appears during training. We set $\texttt{DIV}\cdot N$ entries of the k-th column of $\mathbf{M}_{\text{in}}$ to 1 uniformly at random. The rest $(1 - \texttt{DIV})N$ templates are unseen at training for the specific k-th fact, forming its OOD set. See Fig. 1-B for an example with $N = 10$ and `DIV` = 0.2.

2.2. Contextual Structure

We define contextual structure as the specific configuration of templates, particularly how their two key constituents vary: fact placeholder and statistical contexts. We consider the following setups, all using exactly N templates, to allow their direct comparison across identical `DIV` values:

1) Position-varied Structure (MC1PosN). All templates use the same statistical distribution, i.e., $\mathbf{P}_n = \mathbf{P}$ for all $n \in [N]$, while fact placeholders $\mathbf{I}_n$ vary across templates n.

2) Statistical-varied Structure (MCNPos1). Each template has its own distinct transition matrix $\mathbf{P}_n$, while using the same fact placeholders ($\mathbf{I}_n = \mathbf{I}$ for all $n \in [N]$).

3) Mixed-varied Structure (MCNPosN). Each template has its own distinct transition matrix $\mathbf{P}_n$ and distinct position placeholder $\mathbf{I}_n$.

2.3. In- and Out-of-distribution Evaluation Metrics

To systematically evaluate the model, we measure its generalization using controlled probing. For sequence $\mathbf{x}$, drawn from template $n \in [N]$, that contains a given fact pair (a, b), we prompt the model with its first half, $\mathbf{x}_{1:T/2}$, containing the source token a, and have it auto-regressively complete the second half, i.e., $\hat{\mathbf{x}}_{T/2+1:T}$. A successful completion requires the model to simultaneously demonstrate: (1) **structural understanding**, by placing a fact token only at the designated position j_n and generic tokens elsewhere; (2) **factual recall**, by generating the correct target token b in the completion; and (3) **statistical consistency**, by ensuring the generated generic tokens follow the template's statistical rules.

We quantify each of these generalization aspects with the following metrics. For positional and factual evaluation, we use zero-one accuracy (see App. B for entropy-based counterparts). For each metric, we distinguish between in-distribution (ID) and out-of-distribution (OOD) performance: We evaluate the model on sequences from two disjoint sets of template-fact pairs (n, k): those seen during training (ID, where $\mathbf{M}_{\text{in}}[n, k] = 1$) and those held out during training (OOD, where $\mathbf{M}_{\text{in}}[n, k] = 0$).

1. **Position accuracy** assesses learning the composition of the two streams: placing a fact token from $\mathcal{V}_\mathcal{K}$ *only* at the designated target position and generic tokens $\mathcal{V}_\mathcal{D}$ everywhere else. Formally, let $\mathcal{T}_n = \{t \mid t > T/2, t \neq j_n\}$

$$\texttt{Acc}_{\text{pos}} := \tfrac{1}{2}\big(\mathbb{1}[\hat{x}_{j_n} \in \mathcal{V}_\mathcal{K}] + \tfrac{1}{|\mathcal{T}_n|}\sum_{t\in\mathcal{T}_n} \mathbb{1}[\hat{x}_t \in \mathcal{V}_\mathcal{D}]\big). \quad (1)$$

2. **Factual accuracy** measures factual recall, requiring that the only fact token in the completion be the correct target b. Concretely,

$$\texttt{Acc}_{\text{fact}} := \mathbb{1}\big[\{\hat{x}_t\}_{t>T/2} \cap \mathcal{V}_\mathcal{K} = \{b\}\big]. \quad (2)$$

3. **Statistical loss** evaluates statistical generalization. We compute the KL divergence between the model's predictions and the ground-truth, averaged over the set of positions

$\mathcal{G}$ where a generic token is generated. For each $t \in \mathcal{G}$, we compare the model's distribution over generic tokens (after masking the fact tokens), $\mathbf{p}_t \in \Delta^{V_\mathcal{D}}$, against the ground-truth $\mathbf{p}_t^*$, the row in the template's transition matrix $\mathbf{P}_n$ specified by the previous generic tokens in the sequence.

$$\texttt{Loss}_{\text{stat}} := \frac{1}{|\mathcal{G}|} \sum_{t \in \mathcal{G}} \mathrm{KL}\big(\mathbf{p}_t \,\|\, \mathbf{p}_t^*\big), \tag{3}$$

3. Impacts of Diversity

Experimental setup. We use a 4-layer decoder-only Transformer, (Radford et al., 2018) trained auto-regressively with the standard next-token prediction loss. Each training sequence has length $T = 50$. We use a template pool of size $N = 10$ and a fact set of $K = 100$ source-target pairs. For the MC, we choose a generic vocabulary set of size $V_\mathcal{D} = 3$. We sweep diversity DIV from $0.1 - 0.9$ and train the model for $30k$ iterations. Unless otherwise noted, metrics are averaged over three random initializations over both model initialization and data generation (including template definition and ID/OOD splits). For each template configuration of Sec.2.2, $\texttt{Loss}_{\text{stat}}$, $\texttt{Acc}_{\text{pos}}$ and $\texttt{Acc}_{\text{fact}}$ are reported in Fig. 2. Each figure shows the heatmap of the metrics across a grid of diversity levels DIV $\in (0, 1)$ and training iterations, for sequences randomly drawn from ID/OOD template-fact distributions.

Remark 3.1 (Variations of the setup). While here we focus on a fixed model size and a statistical stream with disjoint vocabulary and short-range dependencies, we provide experiments in App. C that the same observations hold more generally: we replicate the results of this section across various model depths, in the presence of long-range dependencies (e.g., HMMs), and with token ambiguity between the two streams. Furthermore, we verify that the choice of optimization hyperparameters, such as learning rate, weight decay, and schedulers, does not noticeably alter the impact of diversity on performance. We also briefly discuss the impact of having non-uniform diversity levels across facts at training time.

3.1. Impact of Diversity on ID Performance

Across contextual structures and full range of diversity levels, the model ultimately reaches perfect ID performance on all three generalization metrics. Yet, the rate varies. As Figs. 2-(a,b) show, both $\texttt{Loss}_{\text{stat}}$ and $\texttt{Acc}_{\text{pos}}$ converge to 0 and 1, respectively at nearly the same rate, largely unaffected by DIV. In contrast, $\texttt{Acc}_{\text{fact}}$ is sensitive to diversity: the more templates in which a fact appears in training, the longer the model requires to disentangle the correct source–target mapping from contextual patterns. Fig. 2-(c) illustrates this delay clearly: as DIV increases (lower rows in heat map), the yellow band marking perfect recall shifts rightward, signifying that additional training is needed to reach full accuracy in higher-diversity settings. However, the severity of this slowdown depends on the contextual structure, with the MC1POS10 setup showing only a minimal effect.

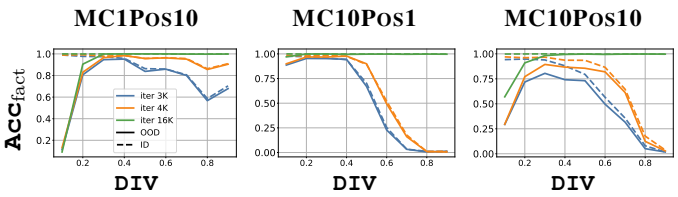


Figure 3. Diversity pays off with time. Factual accuracy versus diversity level after 3K, 4K and 16K iterations. Depending on the template structure, very low diversity can remain unrecoverable; no amount of additional training can restore OOD performance. High-diversity runs begin with lower accuracy but continue to improve until they surpass low-diversity models, demonstrating that diversity incurs an initial cost yet yields benefits in long training regimes. Each bold curve is the average over three individual runs for OOD (solid) and ID (dashed) data.

3.2. Impact of Diversity on OOD Performance

On unseen template-fact distributions, the effect of diversity level depends critically on contextual structure: how fact positions and token statistics vary across templates.

Position accuracy (Fig. 2-(b)). In the MC10POS1 setup, where all templates share the same placeholder positions, the model separates generic and fact positions at nearly the same rate for every diversity level DIV. In the MC1POS10 and MC10POS10 setups, where there are more variations across templates on the position assignments, diversity plays a crucial role: In extreme low-diversity settings (i.e., DIV $=$ 0.1 or 0.2), progress in $\texttt{Acc}_{\text{pos}}$ slows down noticeably or fails. However, once diversity is moderate or high, the learning speed differences across DIV become negligible.

Statistical loss (Fig. 2-(a)). In the MC1POS10 setup, where all templates share the same token statistics (i.e., identical transition matrices $\mathbf{P}_n$), the statistical loss is largely unaffected by the diversity level and converges at a consistent rate regardless of DIV. In contrast, for the MC10POS1 and MC10POS10 setups, which feature statistical variations across templates, diversity becomes a critical factor. At low diversity levels (DIV < 0.3), prolonged training causes the model's predicted distribution to diverge from the ground truth, as displayed by the light bars in the top-right corners of the OOD heatmaps in Fig. 2-(a)

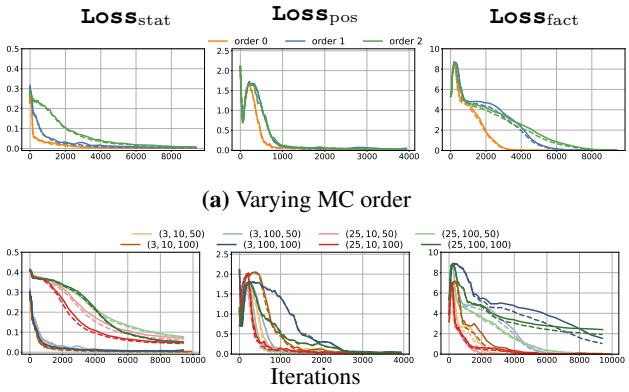
Factual accuracy (Fig. 2-(c)). The effect of diversity on factual recall is more nuanced. At DIV $= 0.9$, $\texttt{Acc}_{\text{fact}}$ improves more slowly, just as it does on ID data, whereas at DIV ≤ 0.2 the model suffers severe factual errors. The severity of this low-diversity failure, however, depends on

the contextual structure. In MC1POS10 setup, where all templates share the same transition matrix, the slowdown under high diversity is modest, yet the low-diversity failure is severe. In contrast, in MC10POS1, low diversity slightly delays reaching high $\mathbf{Acc}_{\text{fact}}$ rather than causing complete failure, as the fixed position signal across templates acts as a strong positional cue, making the model less sensitive to the lack of contextual variety. Lastly, the behavior for MC10POS10 is a hybrid of MC10POS1 and MC1POS10: It fails at low diversity and shows slower convergence at high diversity.

Low DIV training entangles facts and templates. When prompted with an unseen template-fact pair (n, k), a model trained on low-diversity data does not complete the sequence randomly. Instead, it closely follows one of the existing templates. However, this template is typically not the prompt template n, but a different template n' for which $\mathbf{M}_{\text{in}}[n', k] = 1$, i.e., the model follows a template-fact combination seen during training. We verify this in the MC10POS10 with $\texttt{DIV} = 0.1$, where we compute $\mathbf{Loss}_{\text{stat}}$ between the generated tokens and every template distribution $n' \in [N]$, not just the prompt template n. We observe that for OOD sequences, the average $\mathbf{Loss}_{\text{stat}}$ to the *best-matching* template is only 0.004, in contrast to the ground-truth prompt template (Fig. 2), which is around 0.8. Moreover, for 98% of OOD sequences, this best-matching template is one that was seen during training with the same fact. Thus, the model has learned to restrict its generation to valid statistical structures. However, under low diversity, it often fails to disentangle a fact from the contexts in which it was observed: Rather than adapting to the new template-fact combination specified by the prompt, it falls back to a familiar training-time template-fact pairing.

3.3. Trade-off of Diversity for Factual Accuracy

To understand how the impact of diversity on factual recall depends on training duration, in Fig. 3, we plot $\mathbf{Acc}_{\text{fact}}$ versus diversity level DIV at three training checkpoints – roughly 3K, 4K, and 16K iterations. With a short budget (3K steps), *intermediate* diversity is best: each fact appears in enough contexts to isolate the associations between fact tokens, yet the model is not overloaded with context variations that would stall learning. With more updates (4K steps), factual recall in the high-diversity regimes keeps improving, and in the later checkpoints (16K steps) the performance in high diversity setting recovers and the extra training time compensates for its slower start. Intermediate diversity also benefits from longer training, but extremely low diversity does not recover: except in the MC10POS1 setup, that has a strong position signal for the facts, low-diversity runs fail to make significant progress with additional training time. Thus, the sweet spot of diversity depends on both training length and how the statistical and factual streams interact in each template. Finally, when low diversity hurts OOD recall (solid lines), the ID recall (dashed lines) remains nearly perfect. However, in high-diversity runs, poor OOD performance is accompanied by similarly poor ID performance, suggesting that training length, not the distribution shift, is the limiting factor.



(a) Varying MC order

(b) Varying the streams' vocab size and the sequence length: $(V_{\mathcal{D}}, K, T)$

Figure 4. The complexity of one stream impacts learning of the other. Speed of learning the statistical, position, and factual metrics as data complexity is varied in a high-diversity setting, by varying **(top)** the MC order, and **(bottom)** the generic vocabulary size $V_{\mathcal{D}}$, number of facts K, and sequence length T. See Sec.3.4 for discussion and App.B for metrics.

3.4. Streams' Complexity and Learning Dynamics

Here, we study how learning dynamics are affected by systematically varying the complexity of the two streams. We conduct our analysis in a high-diversity regime ($\texttt{DIV} = 0.8$), where all three metrics are learnable with sufficient training. We track $\mathbf{Loss}_{\text{stat}}$ along with $\mathbf{Loss}_{\text{pos}}$ and $\mathbf{Loss}_{\text{fact}}$, the entropy-based analogues of position and factual accuracy (see App. B), while varying four key factors: the generic vocabulary size ($V_{\mathcal{D}} = 3, 25$), the number of facts ($K = 10, 100$), the sequence length ($T = 50, 100$), and the order of the MC (0 to 2). Results are shown in Fig. 4. We observe a strong dependence between how fast each metric can be learned. Specifically:

$\mathbf{Loss}_{\text{stat}}$. As expected, increasing the complexity of the statistical stream slows the decay of $\mathbf{Loss}_{\text{stat}}$. This effect is clear when either increasing the order of the MC (Fig. 4-(a)) or increasing the generic vocabulary size from $V_{\mathcal{D}} = 3$ to $V_{\mathcal{D}} = 25$ for both sequence lengths (Fig. 4-(b)). However, this is not the only factor, as the increased complexity of the factual stream (a larger K) also negatively impacts learning, albeit to a lesser degree. Interestingly, we observe a slight acceleration in statistical learning with increased sequence length (T). This effect is likely because longer sequences provide more training examples of the short-range dependencies due to the Markovian assumption.

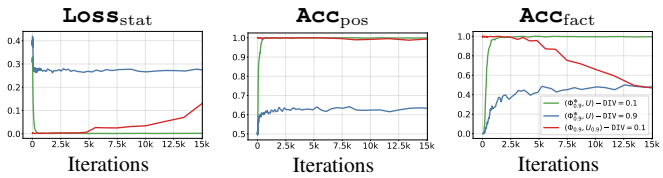


Figure 5. Probing the last-layer features. Learning curves for the unembedding layer ($\mathbf{U}$) retrained on last-layer features (Φ) from models pre-trained on high- or low-diversity data. When using features from a low-diversity model, retraining $\mathbf{U}$ even on high-diversity data shows little improvement over the baseline (blue). Conversely, retraining $\mathbf{U}$ on low-diversity data leads to successful generalization when using features from a high-diversity model (green). However, these high-quality features degrade if they are not frozen and are updated along with $\mathbf{U}$ on the low-diversity data (red). In the legend, ❄ denotes that frozen parameters, and the subscript indicates the diversity level (DIV) on which those parameters were originally trained. See Sec. 4.1 for details.

Loss$_{\text{pos}}$. The speed of positional learning is governed by two factors: sequence length (T) and the relative sizes of the vocabulary sets $\mathcal{V}_{\mathcal{D}}$ and $\mathcal{V}_{\mathcal{K}}$. Longer sequences generally delay the model's ability to identify the fact positions, an effect visible in Fig. 4-(b) by comparing the slower convergence of the darker curves (longer sequences) to the lighter ones. However, when the factual vocabulary is small and comparable to that of the statistical stream (e.g., $K = 10$ vs. $V_{\mathcal{D}} = 25$), the model learns the fact positions rapidly even with larger T, as shown by the red curve in Fig. 4-(b).

Loss$_{\text{fact}}$. The dynamics of factual learning are more nuanced, showing stronger dependencies on the other two learning aspects. First, factual learning is contingent on positional learning; the model learns *where* to place facts before it learns *what* facts to generate, consistent with observations in Zucchet et al. (2025). Consequently, any factor that slows positional learning, such as increasing the sequence length, slows down factual learning and often has an amplified negative impact. Second, factual learning is not independent of the statistical stream. Even when other factors are held constant, increasing only the statistical stream's complexity (larger $V_{\mathcal{D}}$ or higher order) can also slow down the model's ability to learn facts.

4. Localizing the Optimization Bottleneck

Previous sections showed that data diversity affects OOD generalization in two ways: sufficient diversity is needed to generalize to unseen template-fact pairs, while very high diversity can slow convergence to the generalizing solution. This impact is not about the amount of exposure to each of the streams themselves: all individual facts and templates appear with roughly the same frequency at training time for any DIV. Instead, DIV controls which *combinations* of facts and templates are seen during training, and thus sets the test-time distribution shift.

The effect of DIV is also clearly not about model expressivity. Success in mid– to high–diversity regimes shows that a generalizing solution is well within the model's capacity. The issue is one of *optimization*: the set of generalizing solutions remains a subset of the global minimizers of the training loss for every DIV. Diversity therefore, changes whether, and how quickly, training finds one of these solutions, rather than removing them from the landscape. We provide a more formal discussion in App. D, where we theoretically show in a linear setting how the lack of diversity can indeed lead to generalization failure by changing the implicit bias of the algorithm.

Drawing on this intuition, here we perform a series of architectural interventions to localize which specific transformer modules contribute to this optimization bottlenecks and if their impact varies across our three metrics of generalization.

4.1. Probing the Last-layer Features

Low diversity corrupts the learned features. Let $\boldsymbol{\theta}_{\text{low}}$ be the model trained with DIV $= 0.1$. We probe the learned features by freezing the whole network except the unembedding layer $\mathbf{U}$, then retrain $\mathbf{U}$ from a random initialization on diverse data with DIV $= 0.9$. This allows us to isolate the effect of diversity only on the last-layer features. As shown by the blue curves in Fig. 5, even after retraining $\mathbf{U}$, performance improves only marginally over $\boldsymbol{\theta}_{\text{low}}$. Thus, simple retraining of the last layer on high-quality data (Kirichenko et al., 2023) does not fix the problem, suggesting that $\boldsymbol{\theta}_{\text{low}}$ failed to learn the features required for generalization.

High diversity promotes good feature learning. Now let $\boldsymbol{\theta}_{\text{hi}}$ be the model trained with DIV $= 0.9$. We freeze all feature-producing modules and retrain $\mathbf{U}$ on low-diversity data with DIV $= 0.1$. The green curves in Fig. 5 reach perfect performance, which hints that features learned under high diversity disentangle template and fact effectively in the feature space such that $\mathbf{U}$ trained only on low-diversity data can still recover the right mapping. These features are, however, fragile: when we unfreeze the body of $\boldsymbol{\theta}_{\text{hi}}$ and continue training at DIV $= 0.1$ together with $\mathbf{U}$, performance drops (red curves), suggesting that continued exposure to low-diversity data corrupts the initially good features.

We further analyze how training diversity affects the clustering properties of the model's internal embeddings in App. E. We next investigate which modules introduce a bottleneck for effective feature learning.

4.2. Module-wise Interventions for Locating the Bottleneck

We now retrain models on different diversity levels, not end-to-end, but after applying controlled interventions on

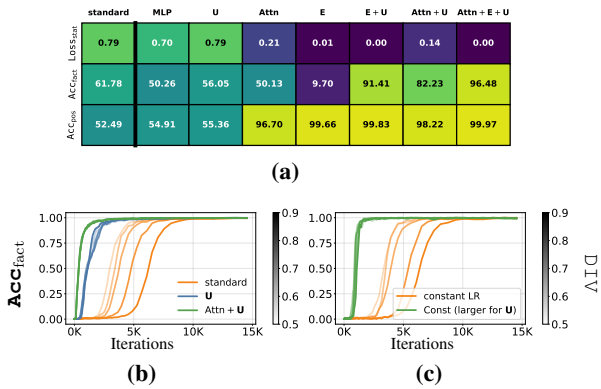


Figure 6. Tracing the optimization failure to individual model components. Models trained while a selected module (MLP, unembedding (**U**), token embedding (**E**), or attention (**Attn**)), is frozen to its counterpart from a well-trained model on high-diversity data. *Standard* refers to training the model, end-to-end with no interventions. All reported values are averaged over three independent runs. **(a)** Performance on all metrics after 30K iterations on low-diversity ($\texttt{DIV} = 0.1$) data. Note that for $\texttt{Loss}_{\text{stat}}$ darker and for $\texttt{Acc}_{\text{pos}}$ and $\texttt{Acc}_{\text{fact}}$ brighter colors are optimal. **(b)** Factual accuracy over time, for different diversity values (darker curves correspond to higher diversity), showing the convergence speed to perfect recall for various interventions. **(c)** Using larger learning rate for the unembedding layer **U** mitigates the training slowdown in higher diversity regimes. See Sec. 4.2 for details and App. C.2 for more results.

specific modules. For each intervention, we replace a module with its counterpart from a well-trained model ($\boldsymbol{\theta}_{\text{hi}}$, trained on $\texttt{DIV} = 0.9$). We target four modules: token embeddings (**E**), attention (**Attn**), MLP, and unembedding layer (**U**) (see App. C for more ablations). For all these modules, except **Attn**, we copy and freeze their corresponding weights at initialization. The special case where all modules except the unembedding layer are frozen is equivalent to the linear probing discussed in the previous section. Our intervention method differs for attention, where we patch the pre-computed attention patterns from $\boldsymbol{\theta}_{\text{hi}}$ during each forward pass (Zucchet et al., 2025).

Here, we focus our analysis on the MC10Pos10 setup trained at a low diversity of $\texttt{DIV} = 0.1$, as it fails on both factual and statistical learning, with results for this setting summarized in the table in Fig. 6. Comprehensive results across all diversity levels, along with corresponding experiments on a one-layer Transformer, are provided in App. C. Our key observations are as follows:

Stats and position fail due to poor Attn and/or E optimization. Transferring either the attention patterns (**Attn**) or the token embeddings (**E**) from $\boldsymbol{\theta}_{\text{hi}}$ allows the model to learn effective features for statistical and positional generalization, even when trained on low-diversity data. Using **Attn** patterns from $\boldsymbol{\theta}_{\text{hi}}$, essentially telling the model the correct position to focus in context, enables obtaining near-perfect $\texttt{Acc}_{\text{pos}}$ and a large drop in $\texttt{Loss}_{\text{stat}}$. Freezing **E** produces the same effect: near-perfect $\texttt{Acc}_{\text{pos}}$ and a statistical loss close to zero.

Factual recall needs a well-optimized U *and* good features. Intervening on feature-producing modules alone (any subset of component not including **U**) does not restore factual recall. However, intervening on **U** together with **Attn** or **E** boosts factual accuracy along with improved statistical generalization. Importantly, intervening on **U** alone, while features are learned from low-diversity data, does not help factual recall. Thus, the bottleneck for factual generalization is joint: **U** must be trained together with the features.

The optimization of U drives the high-diversity slowdown. We trace the cause of slower convergence in high-diversity regimes specifically to the unembedding layer **U**. As shown in Fig. 6-(b), while standard end-to-end training shows progressively slower convergence as diversity increases from $\texttt{DIV} = 0.5$ to 0.9, factual recall improves at a consistent pace, regardless of the diversity level, when we only train the features and keep the unembedding layer frozen to well-trained weights $\boldsymbol{\theta}_{\text{hi}}$. This suggests the optimization of the unembedding layer as the cause of the slow generalization with higher diversity. To test if this bottleneck can be mitigated in the end-to-end training, we train the full model this time with a larger learning rate (10^{-1}) for the unembedding layer, compared to the small learning rate (10^{-3}) for the body. As shown in Fig. 6-(c), with this adjusted learning rate, the number of iterations needed to reach the perfect accuracy becomes independent from the diversity level.

5. Beyond Small Scale: Synthetic Biographies

The experiments in the previous sections use a deliberately minimal testbed to achieve controllability. To test whether the main observations are specific to this minimal construction, we also run a larger-scale experiment closer to realistic settings. We use the synthetic-biography setup of Zucchet et al. (2025), discussed in Sec. 1: We generate 5-sentence biography entries for 10K fictional individuals, where each sentence describes one attribute of the person, such as occupation or place of birth. For each attribute, the sentence is drawn from a pool of 20 natural-language templates with placeholders for the name and attribute value. These templates are partitioned into ID and OOD sets for each individual-attribute fact.

Here, we slightly adapt this dataset to better match our setting: To introduce a simple statistical component to the task, we prepend each sequence with a random-length prefix of random tokens before feeding it to the model. This prevents the model from achieving 100% ID accuracy by simply memorizing a fully deterministic context.

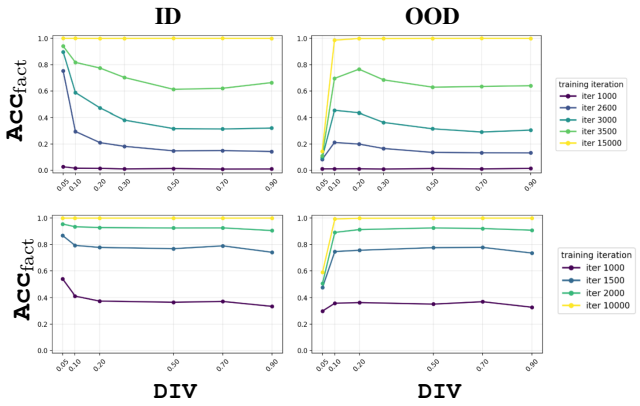


Figure 7. Experiments on synthetic-biography data with an 80M-parameter model. (Top) ID and OOD factual accuracy across diversity levels at several training checkpoints, for a model trained end-to-end. **(Bottom)** Same as (top), but with the embedding and unembedding layers initialized from, and frozen to, those of a model trained on high-diversity data, as described in Sec. 4.2. See Sec. 5 for details.

We train an 80M-parameter decoder-only Transformer with a GPT-2 tokenizer on this dataset with different diversity levels at training time. As shown in Figs. 7, in this larger-scale setting, we observe the same qualitative picture. (1) When diversity is low, the model suffers poor OOD factual recall, achieving only about 10% accuracy on unseen templates. (2) Higher diversity slows both ID and OOD factual learning; thus, under limited training budgets, the highest-diversity setting might achieve suboptimal performance. (3) Under low diversity, freezing the embedding and unembedding layers to those of a well-trained model, as described in Sec. 4.2, improves OOD accuracy from about 10% to 65%. Furthermore, the slowdown of factual accuracy in higher diversity regimes disappears with this intervention.

6. Conclusion and Limitations

We have introduced a synthetic framework to systematically investigate the generalization abilities of language models pre-trained on sequences that combine multiple streams of data. The simplicity of our testbed is intentional, offering several advantages: **(1)** it enables fine-grained control over diversity level and sequence structure, **(2)** it requires minimal computational overhead; and **(3)** it captures different aspects of language model generalization in ID and OOD settings. Using this, we have uncovered several illuminating behaviors regarding the interplay between *statistical* and *factual* generalization as functions of diversity and structure.

Our findings, derived from this controllable and inexpensive testbed, open several avenues for deeper investigation into the learning dynamics of language models often obscured in large-scale regimes. Building upon our results in Sec. 4, future works can formalize our observations, such as how data diversity shapes the optimization landscape and how different modules come to serve specialized roles. Also, our setup provides a controlled baseline for studying the model's behavior under more realistic and challenging data conditions, such as the noisy, long-tailed, and imbalanced distributions characteristic of web-scale corpora.

Finally, the minimalist nature of our testbed was a deliberate choice to enable controlled and low-cost analysis. At the same time, natural-language data is considerably more complex: factual and linguistic regularities are more entangled, facts are not always atomic or independent, and the same token can take on different roles across contexts. Our setup abstracts away many of these complications, while also testing several relaxations in the appendix and in the larger-scale experiment of Sec. 5. While the main observations often carry over, these more realistic variants also introduce additional nuances. For example, freezing the embedding and unembedding layers improves factual recall in Sec. 5, but does not fully recover performance as in Fig. 6. A key challenge for future work is therefore navigating the trade-off between analyzability and linguistic realism. Bottom-up refinements are necessary to better capture the complex structure of natural language, but it is also important to preserve enough control to isolate the mechanisms behind the observed behaviors.

Acknowledgements

This work was partially funded by the NSERC Discovery Grant No. 2021-03677, the Alliance Grant ALLRP 581098-22, and a gift from Google. TB and PD were also supported by the UBC 4YF Doctoral Fellowship. The authors also acknowledge the use of the Sockeye cluster by UBC Advanced Research Computing and thank the anonymous reviewers for their helpful feedback.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Akyurek, E., Bolukbasi, T., Liu, F., Xiong, B., Tenney, I., Andreas, J., and Guu, K. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 2429–2446, 2022. doi: 10.18653/v1/2022.findings-emnlp.180.

Allen-Zhu, Z. and Li, Y. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-*

first International Conference on Machine Learning, 2024. URL https://openreview.net/forum?id=5x788rqbcj.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36: 1560–1588, 2023.

Chang, H., Park, J., Ye, S., Yang, S., Seo, Y., Chang, D.-S., and Seo, M. How do large language models acquire factual knowledge during pretraining? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Chen, L., Bruna, J., and Bietti, A. Distributional associations vs in-context reasoning: A study of feed-forward and attention layers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WCVMqRHWW5.

Choe, M., Cho, H., Seo, C., and Kim, H. Do all autoregressive transformers remember facts the same way? a cross-architecture analysis of recall mechanisms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28482–28501, 2025.

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022.

Edelman, E., Tsilivis, N., Edelman, B., Malach, E., and Goel, S. The evolution of statistical induction heads: In-context learning markov chains. *Advances in Neural Information Processing Systems*, 37:64273–64311, 2024.

Elazar, Y., Kassner, N., Ravfogel, S., Feder, A., Ravichander, A., Mosbach, M., Belinkov, Y., Schütze, H., and Goldberg, Y. Measuring causal effects of data statistics on language model's 'factual' predictions. *arXiv preprint arXiv:2207.14251*, 2022.

Friederici, A. D. *Language in Our Brain: The Origins of a Uniquely Human Capacity*. The MIT Press, Cambridge, MA, 2017. ISBN 9780262035277.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Geva, M., Bastings, J., Filippova, K., and Globerson, A. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Hu, S., Tu, Y., Han, X., Cui, G., He, C., Zhao, W., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhang, X., Thai, Z. L., Wang, C., Yao, Y., Zhao, C., Zhou, J., Cai, J., Zhai, Z., Ding, N., Jia, C., Zeng, G., dahai li, Liu, Z., and Sun, M. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=3X2L2TFr0f.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pp. 15696–15707. PMLR, 2023.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Lee, N., Sreenivasan, K., Lee, J. D., Lee, K., and Papailiopoulos, D. Teaching arithmetic to small transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dsUB4bst9S.

Li, S., Li, X., Shang, L., Dong, Z., Sun, C.-J., Liu, B., Ji, Z., Jiang, X., and Liu, Q. How pre-trained language models capture factual knowledge? A causal-inspired analysis. In *Findings of the association for computational linguistics: ACL 2022*, pp. 1720–1732, 2022.

Li, Y., Huang, Y., Ildiz, M. E., Rawat, A. S., and Oymak, S. Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 685–693. PMLR, 2024.

Liu, L. Z., Wang, Y., Kasai, J., Hajishirzi, H., and Smith, N. A. Probing across time: What does roberta know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 820–842, 2021.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lv, A., Chen, Y., Zhang, K., Wang, Y., Liu, L., Wen, J.-R., Xie, J., and Yan, R. Interpreting key mechanisms of factual recall in transformer-based language models. *arXiv preprint arXiv:2403.19521*, 2024.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeLIgBKPS.

Makkuva, A. V., Bondaschi, M., Girish, A., Nagle, A., Jaggi, M., Kim, H., and Gastpar, M. Attention with Markov: A curious case of single-layer transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SqZ0KY4qBD.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546.

Matchin, W. and Hickok, G. The cortical organization of syntax. *Cerebral Cortex*, 30(3):1204–1221, 2020. doi: 10.1093/cercor/bhz163.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Nichani, E., Lee, J. D., and Bietti, A. Understanding factual recall in transformers via associative memories. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hwSmPOAmhk.

Omar, R., Mangukiya, O., Kalnis, P., and Mansour, E. ChatGPT versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466*, 2023.

Park, C. F., Lubana, E. S., and Tanaka, H. Competition dynamics shape algorithmic phases of in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XgH1wfHSX8.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Rajaraman, N., Bondaschi, M., Makkuva, A. V., Ramchandran, K., and Gastpar, M. Transformers on markov data: Constant depth suffices. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=5uG9tp3v2q.

Ren, Y., Wang, Z., and Lee, J. D. Learning and transferring sparse contextual bigrams with linear transformers. *Advances in Neural Information Processing Systems*, 37: 20304–20357, 2024.

Roberts, A., Raffel, C., and Shazeer, N. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 5418–5426, 2020.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

Sun, K., Xu, Y., Zha, H., Liu, Y., and Dong, X. L. Head-to-tail: How knowledgeable are large language models (LLMs)? AKA will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 311–325, 2024.

Tian, Y., Wang, Y., Chen, B., and Du, S. S. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in neural information processing systems*, 36:71911–71947, 2023.

Yang, Z., Band, N., Li, S., Candes, E., and Hashimoto, T. Synthetic continued pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=07yvxWDSla.

Zucchet, N., Bornschein, J., Chan, S. C., Lampinen, A. K., Pascanu, R., and De, S. How do language models learn facts? dynamics, curricula and hallucinations. In *Second*

Conference on Language Modeling, 2025. URL https://openreview.net/forum?id=vBcGnragkr.

Table 1. Table of notations.

Symbol	Meaning
T	Sequence length.
$\mathbf{x} = (x_1, \dots, x_T)$	Token sequence produced by mixing two streams.
$\mathcal{V} = [V]$	Global vocabulary.
Factual stream	
K	Number of atomic facts.
$\mathcal{K} = \{(a_k, b_k)\}_{k=1}^{K}$	Set of $(\text{source}, \text{target})$ pairs.
$\mathcal{A} = \{a_k\},\ \mathcal{B} = \{b_k\}$	Source and target vocabularies.
$\mathcal{V}_K = \mathcal{A} \cup \mathcal{B}$	Fact vocabulary.
Statistical stream (templates)	
N	Number of templates.
$\mathcal{D}_n$	Distribution associated with template n.
$\mathcal{V}_{\mathcal{D}} = \mathcal{V} \setminus \mathcal{V}_K$	Set of generic vocabularies (the MC vocabulary).
$\mathbf{I}_n = (i_n, j_n)$	Placeholder positions for fact pairs in template n ($i_n \le T/2 < j_n$).
$\mathbf{P}_n$	The MC transition matrix associated with template n.
Exposure and diversity	
$\mathbf{M}_{\text{in}} \in \{0,1\}^{N \times K}$	ID exposure mask: (n, k) entry is 1 if fact k occurs in template n during training.
$\texttt{DIV} \in (0, 1]$	Diversity level: fraction of templates in which each fact appears.
Metrics	
$\texttt{Acc}_{\text{pos}}$	Structural (position) accuracy (Eq. (1)).
$\texttt{Acc}_{\text{fact}}$	Factual accuracy (Eq. (2)).
$\texttt{Loss}_{\text{stat}}$	Statistical loss (Eq. (3)).
$\texttt{Loss}_{\text{pos}}$	Entropy analogue of $\texttt{Acc}_{\text{pos}}$ (Eq. (5)).
$\texttt{Loss}_{\text{fact}}$	Entropy analogue of $\texttt{Acc}_{\text{fact}}$ (Eq. (7)).

Other Notations. We denote matrices/vectors/scalars as $\mathbf{A}$/$\mathbf{a}$/a respectively. We view token-sequences as vectors and denote $\mathbf{a}_{t_1:t_2}$ as the subsequence from index t_1 to t_2. We denote $\mathbf{A}[i, j]$ the (i, j)-th entry of matrix $\mathbf{A}$, and respectively for vectors. We let $\mathbb{S}(\cdot)$ denote the softmax map, Δ the simplex, and $\mathrm{KL}(\mathbf{p}_1 \,\|\, \mathbf{p}_2)$ the Kullback–Leibler (KL) divergence between probability vectors $\mathbf{p}_1, \mathbf{p}_2$. We denote $[N] := \{1, \dots, N\}$ and use $\mathbb{1}[\mathcal{C}]$ for the indicator function (1 if condition $\mathcal{C}$ is satisfied, 0 otherwise).

Overview of Supplementary Material. Section A expands on the literature discussion in Section 1. Section B introduces additional metrics used in Section 3.4. Section C provides additional experimental details and results to complement Sections 3 and 4. Section D introduces a minimal toy setting – as a starting point for theoretical analysis. Finally, Section E examines the model's internal embeddings and how training diversity affects them. Code available at `https://github.com/TinaBehnia/FactStat`.

A. Additional Details on Related Works

LLMs have been observed to pack a substantial amount of knowledge in their parameters during pretraining, allowing them to answer real-world questions without consulting external resources (Petroni et al., 2019; Roberts et al., 2020), raising the question of whether they can replace conventional knowledge bases (Omar et al., 2023; Sun et al., 2024). By inspecting the internal activations of different modules and layers, a growing body of mechanistic interpretability work seeks to understand where LMs store knowledge (Meng et al., 2022; Dai et al., 2022; Geva et al., 2021) and how they recall it (Geva et al., 2023; Lv et al., 2024; Choe et al., 2025). Other studies trace each learned fact back to the pre-training data to investigate which corpus patterns enable its acquisition (Elazar et al., 2022; Akyurek et al., 2022; Li et al., 2022) and demonstrate that recall accuracy depends on the number of exposures to that fact in the pretraining corpus (Kandpal et al., 2023; Allen-Zhu & Li, 2024). Studies on learning dynamics similarly find that different knowledge types are learned at different rates (Liu et al., 2021). Chang et al. (2024) probe the factual recall dynamics by injecting fictional facts during pre-training and tracking

their probabilities over time, observing that knowledge accumulates through many small "micro-updates" that gradually decay unless the fact is periodically reinforced to avoid forgetting.

Recent works have initiated systematic exploration of factual recall in language models using *controlled synthetic setups*. Typically, these works model each fact as a triplet (source, *relation*, target) embedded in a context; the model is then probed with a context containing a (source, relation) and must produce the corresponding target (Allen-Zhu & Li, 2024; Nichani et al., 2025; Zucchet et al., 2025). We adopt a similar framework but omit the *relation* token. The simpler (source, target) setup suffices for our purpose of studying how models acquire and recall factual associations along with other aspects of generalization from first principles.

Allen-Zhu & Li (2024) use a controlled synthetic biography dataset: each biography entry is a multi-sentence paragraph about an individual, the *source*, where each sentence represents a (relation, target) chosen from a set of fixed-sentence templates, e.g., "*<Name> was born in <City>*". They specifically consider question-answer (QA) formats for probing knowledge, e.g., "*What is <Name>'s city of birth? <City>*", which are shown at training time for only a subset of individuals. We treat these QA forms as just another template family and evaluate by probing the model with any template that was *unseen* for a given fact during training. Focusing on factual recall performance, they show that factual recall after fine-tuning on question templates is only possible if facts are observed in several templates during pre-training and more paraphrase diversity markedly boosts recall. In other words, successful recall requires *varied* exposure to each fact, not mere repetition in a single template.

Our data-generation scheme extends this core insight by further abstracting the template–fact structure. This provides fine-grained control over the statistical and structural composition of the data, aspects that were fixed in the original work. We focus exclusively on pre-training experiments here, although the same abstract framework could also be used to study fine-tuning. Yet, we show that even without finetuning and even in minimal settings like those described in Sec. D low-diversity can impair generalization. Our analysis of the impacts of diversity on the statistical aspects of generalization as well as the systematic categorization of different contextual structures are also unique to our study compared to this prior work.

Recent work by Zucchet et al. (2025) analyzes the dynamics of factual recall in a synthetic biography setup, similar to that of Allen-Zhu & Li (2024), and reports a stage-wise dynamic. By inspecting the model's predictions at the target position across training checkpoints, the model first restricts its choice to the fact vocabulary and only later learns the correct mapping from the the specific (source, relation) present in the context to the correct fact token. We make similar observation between factual and position accuracy Sec. 3.4. Our evaluation, however, is broader: we probe the model with an incomplete prompt and grade the *entire* completion from different aspects – whether the correct fact token appears in the correct position, whether the remaining positions are filled with generic tokens, and whether those tokens follow the template's statistical pattern. As Zucchet et al. (2025) follow similar setup as Allen-Zhu & Li (2024), the unique and distinctive characteristics of our setup mentioned above—particularly the explicit statistical stream enabling a joint study of statistical and factual generalization aspects, alongside systematic control over contextual structure and diversity—apply equally as differentiators here.

Both of the above referenced studies also vary the data distribution, exploring how "celebrity" entries – individuals whose biographies appear in many templates during pre-training – affect factual recall for less-frequent entries and alter learning dynamics. While our current work focuses only on overall context diversity, our flexible framework can readily accommodate such experiments on the impact of data distribution by appropriately designing the template-fact exposure matrix $\mathbf{M}_{\text{in}}$ to vary fact frequencies across templates. We leave this as interesting future work.

Perhaps the most closely-related work in terms of model abstraction, although coming from differing motivations, appears in Nichani et al. (2025). While they focus on capacity and storage tradeoffs for factual recall, we investigate the impacts of diversity and tradeoffs between statistical and factual accuracy. In their synthetic data setup, they sample the (source, relation, target) mappings randomly by choosing them from a fact set. Each sequence is generated by placing the (source, relation) at two random position, appending the target at the end and filling the remaining positions with tokens uniformly drawn from a disjoint *noise* vocabulary (functionally identical to our generic tokens). Training then minimizes the loss at the target position, focusing on the fact storage capacity of the model. Similarly in our setup, we preserve the separation of facts from background tokens but introduce a *structured* statistical stream: generic tokens are generated by a Markov process and fact placeholders occupy slots that can vary across different templates. This richer design allows us to introduce and systematically investigate how contextual structure and diversity affect performance, while extending the analysis beyond factual recall to include statistical and structural generalization, aspects not explored by Nichani et al. (2025). Through our effort to identify minimal toy settings where key tradeoffs, such as the impact of diversity on factual recall, are maintained, it might be possible to leverage some of the theoretical ideas from Nichani et al. (2025) to analyze our findings. However,

this would require various non-trivial extensions, particularly incorporating the impact of diversity and adapting for an autoregressive generation setting rather than their last-token prediction.

We also review a growing body of recent works that have used Markov chains, as we do here to model the statistical stream, to study various behavioral aspects of transformers in next-token prediction tasks. Makkuva et al. (2025) study the loss landscape properties of a single-layer transformer trained on sequences drawn from a fixed order-1 Markov chain, characterizing the influence of transition probabilities and architectural choices on the loss landscape. Edelman et al. (2024) demonstrates that transformers trained on sequences generated from random order-1 Markov chains develop the ability to perform in-context inference on unseen Markov chains by outputting bigram statistics inferred from the context. Park et al. (2025) extend this framework by examining the regime where training sequences are drawn from a fixed, finite set of Markov chains. Rajaraman et al. (2024) has analyzed the representational capacity of transformers for in-context learning of order-k Markov chains. None of these works combines MCs with factual information, as we do here. More broadly, synthetic tasks have proven valuable for dissecting LM behavior beyond our focus, including self-attention mechanisms (Tian et al., 2023; Li et al., 2024; Ren et al., 2024; Makkuva et al., 2025), in-context learning (Garg et al., 2022; Bietti et al., 2023; Edelman et al., 2024; Park et al., 2025), and distributional associations (Chen et al., 2025).

Additionally, our work is broadly related to prominent neuroscientific theories of language processing. For instance, some theories propose separate cortical circuits dedicated to syntactic and semantic processing, suggesting a functional segregation in the brain for handling language structure versus meaning (Matchin & Hickok, 2020). Other research points to a stage-wise progression in human language comprehension, where an initial, rapid phase of syntactic analysis is thought to precede and scaffold deeper semantic integration (Friederici, 2017).

B. Metric Overview

In this section, we review the metrics in Sec. 2.3 along with introducing the entropy-based metrics $\texttt{Loss}_{\text{fact}}$ and $\texttt{Loss}_{\text{pos}}$ used in Sec. 3.4 to study the learning dynamics. Recall that, whether ID or OOD, we measure the model's: (i) adherence to the composition rule between the two streams, (ii) accuracy of factual recall, and (iii) ability to follow the statistical patterns of the background template.

For each ID/OOD template-fact combination (n, k), we draw a sequence $\mathbf{x}_{1:T}$, use its first half ($\mathbf{x}_{1:T/2}$) as a prompt, and have the model generate the second half. We refer to the generated tokens at positions $t > T/2$ by $(\hat{x}_{T/2+1}, \ldots, \hat{x}_T)$. Let $\hat{\boldsymbol{\ell}}_t(\cdot) \in \mathbb{R}^V$ be the model's predicted *logits* at position $t \leq T$ conditioned on input $(x_1, \cdots, x_{T/2}, \hat{x}_{T/2+1}, \cdots, \hat{x}_{t-1})$. Also, let $\hat{\mathbf{p}}_t(\cdot) \in \Delta^{|\mathcal{V}|}$ be the softmax *probability* at this position. Without loss of generality, assume the index of the generic tokens is $[|\mathcal{V}_{\mathcal{D}}|]$.

1. **Position accuracy/loss:** To obey the composition rule between the factual and statistical streams, the generated sequence should contain a fact tokens from $\mathcal{V}_{\mathcal{K}}$ *only* at the target position j_n, and all other positions should contain tokens from the generic vocabulary $\mathcal{V}_{\mathcal{D}}$. Formally, we define *position accuracy* as:

$$\texttt{Acc}_{\text{pos}} = \tfrac{1}{2}\big(\mathbb{1}[\hat{x}_{j_n} \in \mathcal{V}_{\mathcal{K}}] + \tfrac{1}{T/2-1}\textstyle\sum_{t>T/2, t\neq j_n} \mathbb{1}[\hat{x}_t \in \mathcal{V}_{\mathcal{D}}]\big). \tag{4}$$

We also define the *position loss* by measuring how much of the probability distribution at each position is accumulated on the correct subset of the vocabulary. Specifically,

$$\texttt{Loss}_{\text{pos}} = -\tfrac{1}{2}\big(\log(\textstyle\sum_{v\in\mathcal{V}_{\mathcal{K}}} \hat{\mathbf{p}}_{j_n}(v)) + \tfrac{1}{T/2-1}\textstyle\sum_{t=T/2+1, t\neq j_n}^{T} \log(\sum_{v\in\mathcal{V}_{\mathcal{D}}} \hat{\mathbf{p}}_t(v))\big)\big). \tag{5}$$

2. **Factual accuracy/loss:** We define *factual accuracy* as the existence of the correct prediction of the target in the prompt completion generated by the model:

$$\texttt{Acc}_{\text{fact}} := \mathbb{1}\big[\{\hat{x}_t\}_{t>T/2} \cap \mathcal{V}_{\mathcal{K}} = \{b_k\}\big]. \tag{6}$$

We accordingly define the *factual loss* as

$$\texttt{Loss}_{\text{fact}} := -\log\left(\hat{\mathbf{p}}_{j_n}(b_k)\right) . \tag{7}$$

3. **Statistical loss:** Denote $\mathcal{G} \subseteq [T/2+1, T]$ the set of positions in the generated completion that are filled with generic tokens from $\mathcal{V}_{\mathcal{D}}$. For each such position t, we compare the model's distribution over generic tokens with the ground-truth

MC distribution ($\mathbf{P}_n$) of the template. Concretely, keep the first $|\mathcal{V}_\mathcal{D}|$ coordinates of the logit $\tilde{\boldsymbol{\ell}}_t = \boldsymbol{\ell}_t[\mathcal{V}_\mathcal{D}]$ corresponding to the generic tokens and compute the model's distribution over $\mathcal{V}_\mathcal{D}$ as $\tilde{\mathbf{p}}_t = \mathbb{S}(\tilde{\boldsymbol{\ell}}_t)$. Let $\mathbf{p}_t^*$ be the row of $\mathbf{P}_n$ that corresponds to the preceding generic token of $\hat{x}_t$. We measure *statistical loss* as:

$$\texttt{Loss}_{\text{stat}} := \frac{1}{|\mathcal{G}|} \sum\nolimits_{t \in \mathcal{G}} \mathrm{KL}\big(\tilde{\mathbf{p}}_t \,\|\, \mathbf{p}_t^*\big)\,. \tag{8}$$

C. Additional Experiment Results

C.1. Experiment Details of Sec. 3

In the experiments of Secs. 3, we use a decoder-only Transformer (Radford et al., 2018) with 4 layers and 4 heads trained auto-regressively with the standard next-token prediction loss. Unless stated otherwise, each training sequence has length $T = 50$. We use a template pool of size $N = 10$. We choose a generic vocabulary set of size $V_\mathcal{D} = 3$ for the MC and a fact set of $K = 100$ source-target pairs, which gives a total of $V = 203$ vocabularies. We allocate the first $2K$ tokens in the vocabulary to the fact stream and build up the fact set $\mathcal{K}$ by randomly pairing these tokens together. We build the statistical stream by generating N random transition matrices (sampled from the Dirichlet distribution with parameter $\mathbf{1}$) for each template, and N distinct position placeholders $\mathbf{I}_n = (i_n, j_n)$ along the sequence. We train the models for $30k$ iterations and a batch size of 64, with AdamW (Loshchilov & Hutter, 2019) and a fixed learning rate of 10^{-3}.

Unless otherwise specified, all reported metrics are averaged over three independent runs, each with different random seeds for both data generation (which controls the random generation of $\mathcal{K}$, $\{\mathbf{P}_n\}_{n\in[N]}$ and $\{\mathbf{I}_n\}_{n\in[N]}$) and model initialization. For experiments that sweep across the diversity level `DIV` (e.g., Fig. 2), we fix these seeds for a given run and vary only `DIV` from 0.1 to 0.9.

C.2. Experiment Details of Sec. 4

In the controlled experiments of Section 4, we use the same data and model initialization seeds for each experimental run as those used to train the corresponding well-trained model $\boldsymbol{\theta}_{\text{hi}}$. This ensures that the underlying fact and template definitions are identical and that training begins from the exact same initialization point. After initializing the model, we transfer the desired parameter subset from $\boldsymbol{\theta}_{\text{hi}}$, freeze it, and then retrain the model on a new dataset that differs only in its diversity level.

While the main text (Sec. 4.2 and Fig. 6) focused on the impact of interventions at the lowest diversity level ($\texttt{DIV} = 0.1$), Fig. 9 provides the complete results across all diversity levels and training iterations in a format similar to Fig. 2. Here, we only focus on the OOD performance. For completeness, the figure also includes results for interventions on other modules that had a limited impact and were therefore not discussed in the main text. Table 12 reports metrics at iteration 10k (mid-training), analogous to the final results shown in Fig. 6-(a). This mid-training snapshot more clearly demonstrates that intervening on $\mathbf{U}$ accelerates learning, even in low-diversity settings. We replicate the trends in Sec. 4, for a one-layer, one-head Transformer, as detailed in Figs. 10–11.

C.3. Impact of Model Size

Here, we repeat our main experiments – originally with 4-layer transformer – this time with 1-layer and 10-layer models. In this section, all experiments are run with a learning rate 10^{-4}. In Fig. 8, we gather heatmaps of (a) $\texttt{Loss}_{\text{stat}}$, (b) $\texttt{Acc}_{\text{pos}}$, and (c) $\texttt{Acc}_{\text{fact}}$ on OOD sequences for each model size.

Across all model sizes, the impact of diversity level and training duration is identical. For example, in terms of factual recall, low diversity causes a failure, moderate diversity is best for intermediate training length, and high diversity achieves optimal performance only after long training. The primary impact of depth is on the convergence speed: at any fixed iteration count, the 10-layer model achieves higher accuracy than the 4-layer, which in turn outperforms the 1-layer. Notably, larger models help slowly improve *position accuracy* even under low diversity (seen as brighter colors in the upper-right portions of panel b). However, the factual recall failure at extreme low diversity persists across depths, highlighting that model capacity alone cannot substitute for contextual variety in achieving robust OOD generalization.

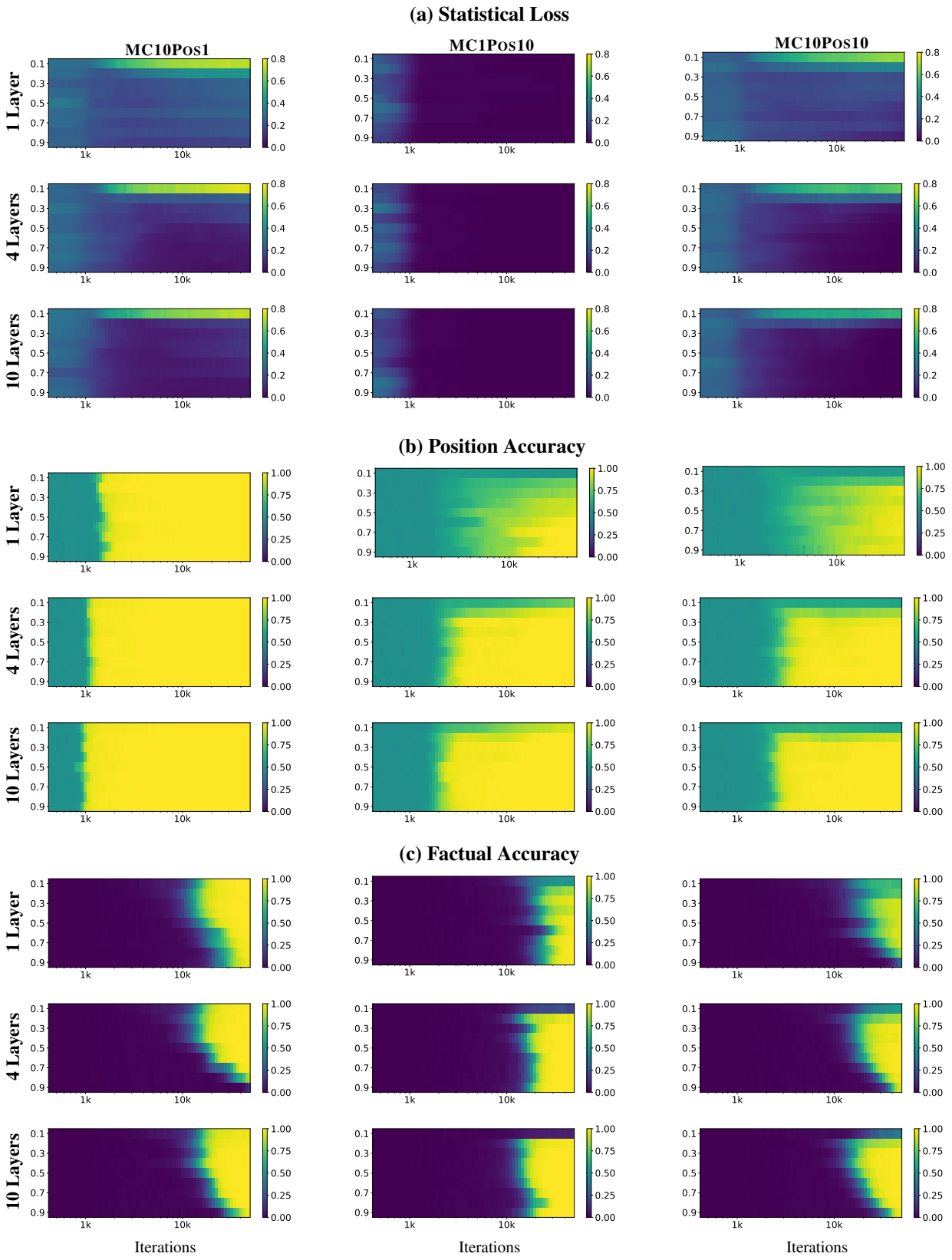
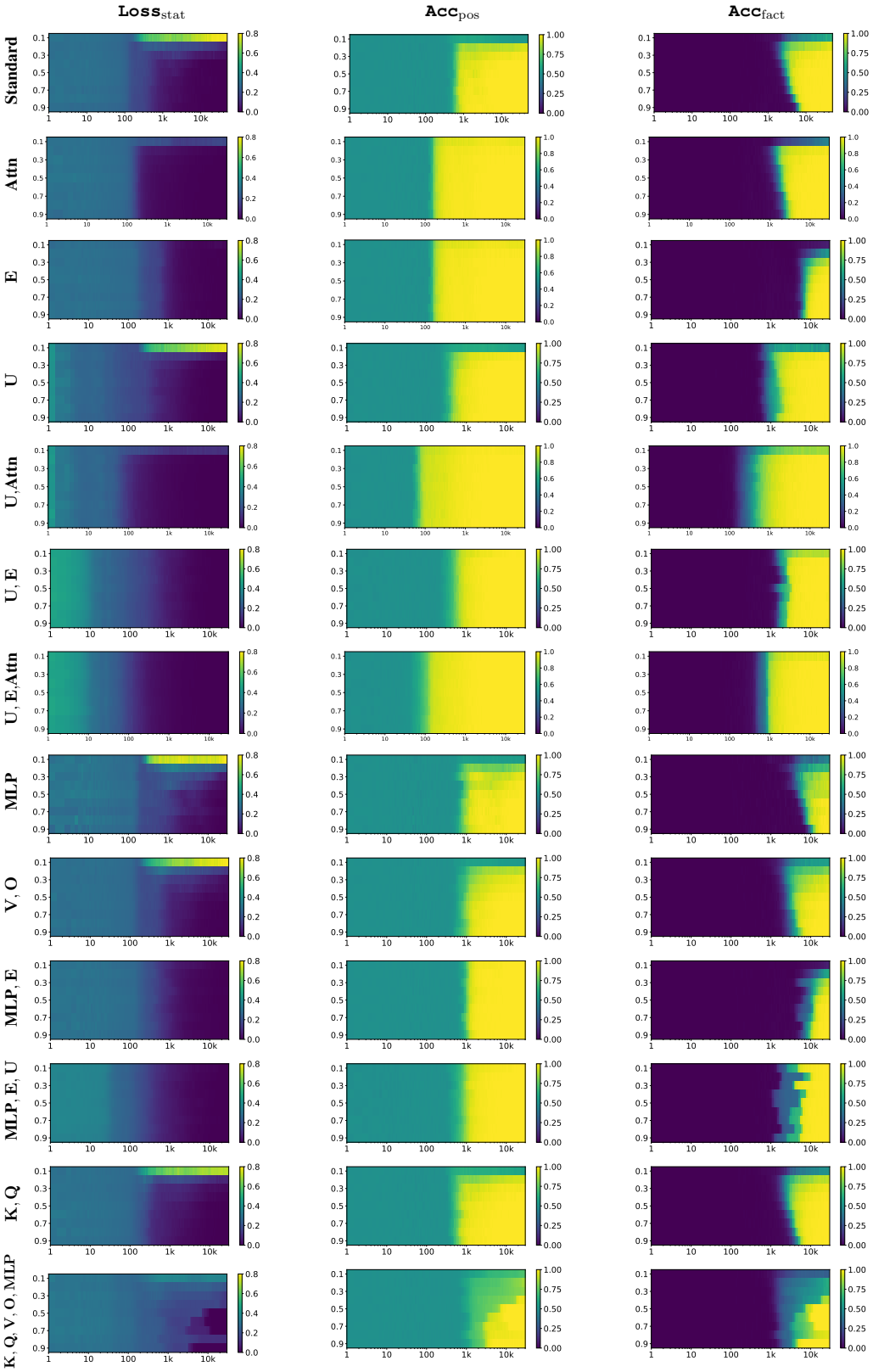


Figure 8. Impact of model size. Replication of the experiments in Sec. 3 (4-layer) with smaller (1-layer) and larger (10-layer) models on OOD sequences. With increased model capacity, we need fewer iterations to achieve the same level of performance. However, model capacity alone cannot alleviate the failure of factual recall at low diversity levels.



Failure at low diversity. Slowdown with high diversity

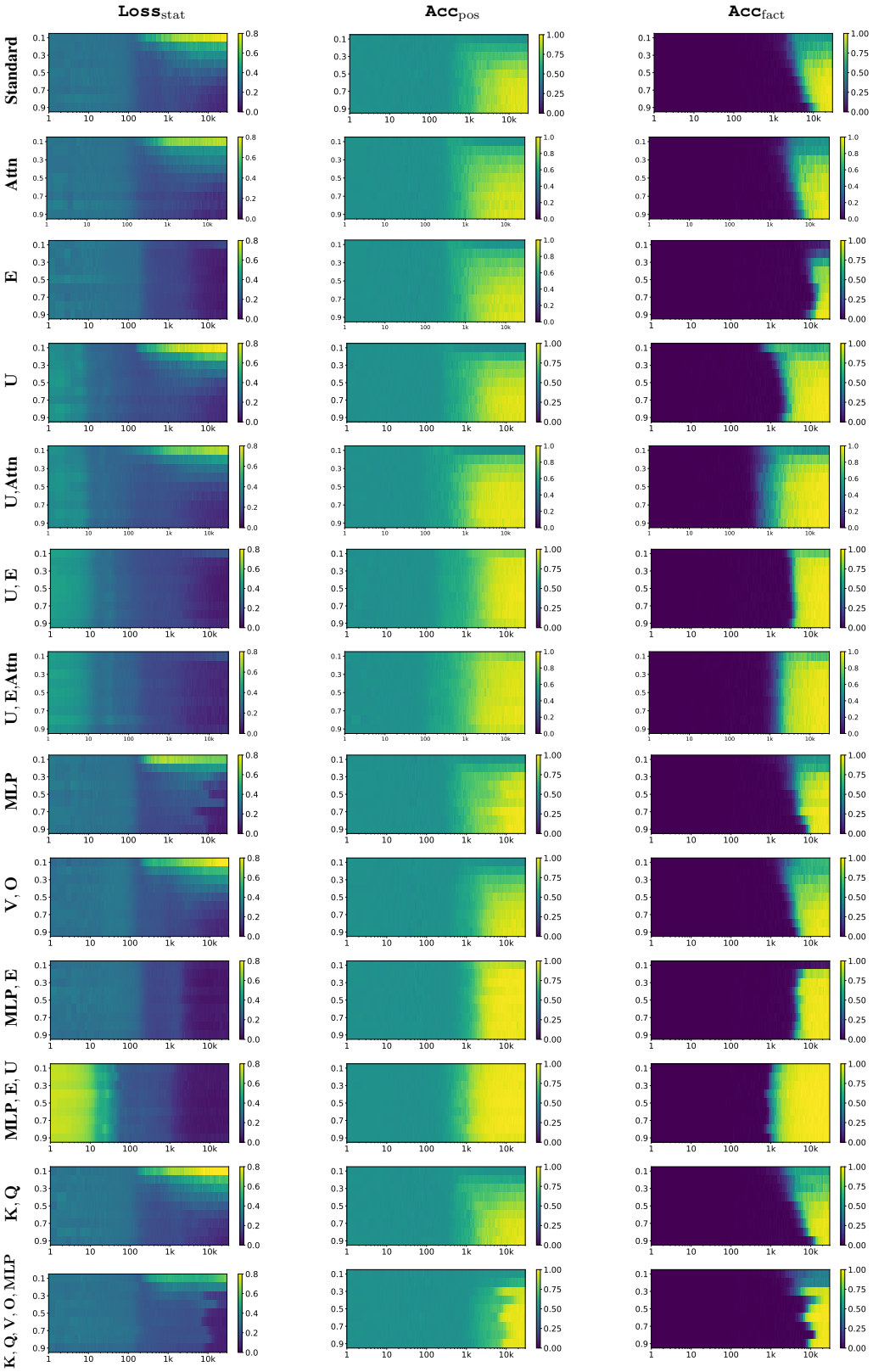
Stat & *pos* recover. No impact on *fact*.

No impact on *stat* and *pos*. Remedies slowdown of *fact*.

Improves all metrics, on both ends of diversity.

No improvements.

Figure 9. Experiments of Sec. 4.2 across different diversity levels and over training time. Each row label specifies which modules in the model, among the following, are frozen to that of $\boldsymbol{\theta}_{\text{hi}}$: softmax attention matrices **Attn**, token embedding / unembedding layer $\mathbf{E}/\mathbf{U}$, the weights of the fully-connected modules at each layer **MLP**, key, query, value, output matrices in self-attention module $\mathbf{K}, \mathbf{Q}, \mathbf{V}, \mathbf{O}$.



Failure at low diversity. Slowdown with high diversity

Stat & *pos* improve. No impact on *fact*.

No impact on *stat* and *pos*. Remedies slowdown of *fact*.

Improves all metrics, on both ends of diversity.

No improvements.

Figure 10. Same as the intervention analysis from Fig. 9 on a one-layer transformer. While the overall trends are similar to the four-layer model, the **Attn** intervention has a less significant impact here. This result aligns with the one-layer model's overall weaker performance (see also Fig. 8).

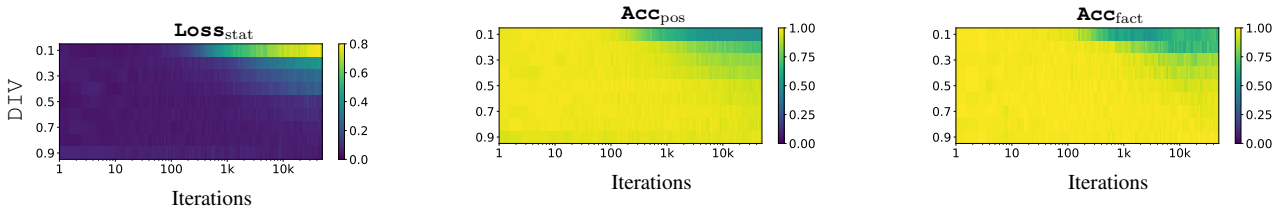


Figure 11. A one-layer transformer, initialized from a model trained on high-diversity data ($\texttt{DIV} = 0.9$), is then trained on a low-diversity dataset ($\texttt{DIV} = 0.1$). Although the model starts with perfect performance, its accuracy degrades across all metrics as training on the less diverse data progresses. This is similar to the degradation observed in Fig 5.

	standard	MLP	U	Attn	E	E + U	Attn + U	Attn + E + U
$\text{Loss}_{\text{stat}}$	0.74	0.64	0.75	0.17	0.01	0.00	0.11	0.00
Acc_{fact}	47.59	36.72	60.03	31.97	5.40	88.80	83.46	96.29
Acc_{pos}	54.49	57.06	58.75	97.07	99.40	99.47	99.14	99.93

Figure 12. Same metrics as Fig. 6-(a), but captured earlier in training (10k vs. 30k iterations). This mid-training snapshot more clearly reveals that the intervention on $\mathbf{U}$ accelerates learning facts, even on low-diversity data that yields sub-optimal performance.

C.4. Hyperparameter Tunings

Hyperparameter changes do not resolve low-diversity failure. In Figs. 13-(a,b), we examine whether standard optimization hyperparameters tuning can mitigate the failure modes observed in low-diversity regimes. We vary weight decay, learning rate, and learning rate schedulers, and find that none of these changes resolve the optimization bottleneck: when diversity is low, training consistently converges to a non-generalizing solution.

In Fig. 13-(a), we report OOD factual accuracy at the end of training (30K iterations) for weight decay values ranging from 10^{-5} to 10^{1}. As long as weight decay is in a reasonable range, models trained on high-diversity data generalize on the OOD templates. In contrast, under low-diversity settings such as $\texttt{DIV} = 0.1$ or 0.2, factual accuracy remains uniformly poor and largely insensitive to the choice of weight decay.

Next, we observe the same pattern when varying the learning rate or the scheduler in Fig. 13-(b). We test four settings: (i) constant learning rates (10^{-4}, 10^{-3}), (ii) a cosine schedule warming up from 10^{-4} to 10^{-3} and decaying to 10^{-5}, (iii) a cosine schedule warming up from 10^{-3} to 10^{-2} and decaying to 10^{-4}, and (iv) warmup–stable decay (WSD) (Hu et al., 2024) with a warmup from 10^{-4} to 10^{-3} followed by a stable phase and a final decay to 10^{-4}. Across all these choices, performance remains effectively unchanged at each diversity level. The only exception is a constant rate of 10^{-4}, where the training progress is slow and even high-diversity runs have not converged by the 30K-iteration mark.

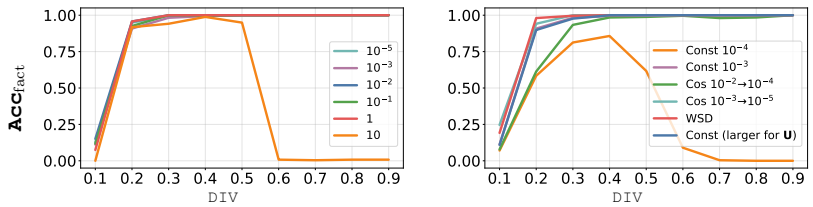
C.5. Variations on the Statistical Stream

C.5.1. HMM Statistical Stream

In this section, we replace the Markov-chain statistical stream with a hidden Markov model (HMM) to see the effect of having long-range dependencies in the statistical context. We focus on the MC10Pos10 setup and for each template, we generate a distinct HMM with $H = 15$ hidden states and $V_{\mathcal{D}} = 10$ observable states. The hidden transition matrix and emission matrix are drawn independently from Dirichlet distributions with parameter 0.1[1]. Aside from replacing the Markov chains with HMMs, all other aspects of the setup follow the experiment specification in Sec. 3.

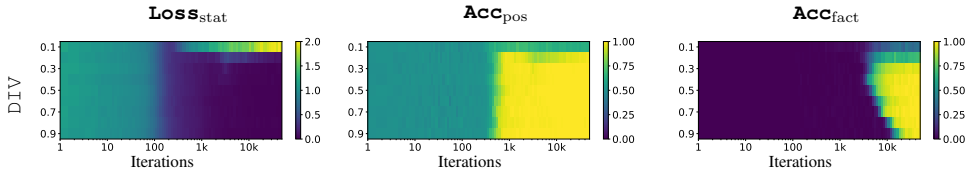
Fig. 14-(a) shows $\texttt{Loss}_{\text{stat}}$, $\texttt{Acc}_{\text{pos}}$, and $\texttt{Acc}_{\text{fact}}$ heatmaps across diversity and training iterations on OOD template-fact pairs. The heatmaps show that the qualitative effect of diversity remains unchanged: in low-diversity regimes, the model fails to generalize, while higher diversity improves OOD performance but results in a slower progress in performance, creating a temporal trade-off similar to the Markov chain case (Figs. 2 and 3).

[1]This choice yields relatively spiky transitions and emissions, producing more varied and context-dependent next-token distributions.



(a) (b)

Figure 13. Hyperparameter ablations. OOD factual accuracy at the final training iteration (30K) with varying **(a)** weight-decay values and **(b)** learning rate values and scheduler. Performance remains largely unchanged across diversity levels.



(a)

	standard	Attn	E	U	E + U	Attn + U	Attn + E + U
$\text{Loss}_{\text{stat}}$	0.32	0.20	0.01	0.27	0.01	0.25	0.01
Acc_{fact}	40.23	21.09	1.17	60.94	65.62	85.55	77.73
Acc_{pos}	71.53	99.79	100.00	71.43	100.00	100.00	100.00

(b)

Figure 14. Hidden Markov Models as statistical stream. (a) $\texttt{Loss}_{\text{stat}}$, $\texttt{Acc}_{\text{pos}}$, and $\texttt{Acc}_{\text{fact}}$. **(b)** Module intervention experiments analogous to Fig. 6: peak performance achieved during training. See App. C.5.1 for details.

The result of the targeted module interventions (following Sec. 4) in this setup also exhibits the same general pattern (Fig. 14-(b)): well-trained token embeddings and attention layers primarily improve positional and statistical generalization, while the unembedding layer is most directly associated with factual accuracy. We also find that reaching high factual accuracy with the module interventions is harder in the HMM setting: Interventions involving the unembedding layer improve factual accuracy considerably, but it does not lead to the full recovery seen in the Markov case (Fig. 6).

C.5.2. Removing the Statistical Component

Here we argue that for studying the impact of diversity, it is necessary to consider the joint learning and modeling of the *statistical* and *factual* streams. Using the same data model as in Sec. 3, we modify the training objective so that the model is required to learn only the factual association. Concretely, we replace the statistical stream with random i.i.d. tokens drawn from the generic vocabulary, and we drop the autoregressive (AR) loss over the statistical tokens and compute the loss only at the position of the target fact token b_k. This setup is similar in nature to the setup considered in Nichani et al. (2025), with the main difference that factual tokens appear at predefined positions (i.e., templates) rather than at random locations.

Fig. 15 reports factual accuracy on sequences drawn from OOD template-fact pairs. In contrast to the AR setting, the need for diversity disappears: at the lowest diversity levels, the model achieves perfect factual accuracy on OOD templates. In other words, the sharp degradation in performance observed in low-diversity regimes throughout Sec. 3 does not occur when the model is trained only on the factual target.

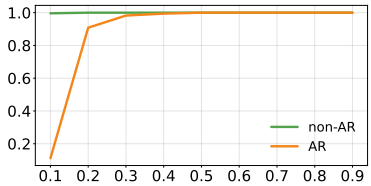
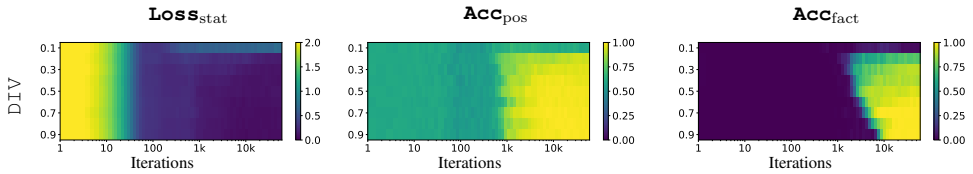


Figure 15. Removing the statistical component Factual accuracy at the final iteration for the non-AR objective (App. C.5.2) versus the standard autoregressive objective (MC10Pos10 in Sec. 3). See App. C.5.2 for details.



(a)

	standard	Attn	E	U	E + U	Attn + U	Attn + E + U
$\mathrm{Loss}_{\mathrm{stat}}$	0.30	0.30	0.12	0.32	0.27	0.37	0.22
$\mathrm{Acc}_{\mathrm{fact}}$	11.50	17.00	12.50	8.00	54.50	31.00	40.00
$\mathrm{Acc}_{\mathrm{pos}}$	67.60	78.40	93.87	60.12	87.12	74.31	92.05

(b)

Figure 16. Impact of token ambiguity. (a) $\mathtt{Loss}_{\mathrm{stat}}$, $\mathtt{Acc}_{\mathrm{pos}}$, and $\mathtt{Acc}_{\mathrm{fact}}$. **(b)** Module intervention experiments analogous to Fig. 6: peak performance achieved during training. See App. C.5.1 for details.

C.6. Effects of Token Ambiguity

In the data model introduced in Sec. 2, the factual tokens and the generic statistical tokens come from disjoint vocabularies. However, in more realistic settings, a token might take different roles. To explore how such token ambiguity might affect the results, we modify the statistical stream so that a small fraction of factual tokens can also appear within it.

Concretely, for each experiment we select 20% of the factual tokens (i.e., 40 tokens) uniformly at random and include them in the statistical vocabulary. The transition matrix of the statistical stream is then defined over the union of generic tokens and this overlapping fact subset. To limit the frequency with which these shared tokens appear as statistical predictions, we constrain the total transition probability mass to these tokens to be 10%. All other aspects of the experiment follow the setup of Sec. 3.

Fig. 16-(a) reports the OOD $\mathtt{Loss}_{\mathrm{stat}}$, $\mathtt{Acc}_{\mathrm{pos}}$, $\mathtt{Acc}_{\mathrm{fact}}$. The overall qualitative effect of diversity remains the same as in earlier experiments (Fig. 2): generalization fails in the low diversity regimes, and higher diversity improves OOD generalization at the cost of slower optimization. The key difference is that the model now requires *larger* diversity value to reach perfect performance across all metrics (see the less bright bands of factual accuracy for $\mathtt{DIV} = 0.3 - 0.5$). The presence of token overlap increases the ambiguity between the factual and statistical streams, which appears to strengthen the reliance on diverse contextual exposure.

The intervention results in Fig. 16-(b) (analogous to Fig. 6) also follow the broad structure observed in previous sections. Token embeddings and attention contribute most strongly to positional and statistical improvement, and the unembedding layer remains most closely associated with factual accuracy. However, the magnitude of the improvements caused by these module-wise interventions is smaller compared to the disjoint-vocabulary setting.

C.7. Non-uniform Diversity across Facts

The main experiments use a balanced exposure mask $\mathbf{M}_{\mathrm{in}}$, so that every fact appears in the same number of templates. Here, we relax this assumption. We partition the facts into two groups: a fraction μK of high-diversity facts and $(1-\mu)K$ low-diversity facts. High-diversity facts appear in 9 out of 10 templates, while low-diversity facts appear in only one template. We then evaluate both groups separately.

As shown in Fig. 17, increasing μ improves performance on the low-diversity facts. In particular, while in the uniform baseline in Fig. 2, OOD factual accuracy is near 10%. With moderate imbalance $\sim \mu = 0.5$, factual accuracy for the low-diversity facts jumps to 60%. Statistical loss and positional accuracy also improve, reaching near-zero loss and 90% accuracy, respectively. This suggests that diversity can help the model learn shared structure, such as how to separate template information from fact identity, and this shared structure can partially transfer to facts that were themselves observed in only a small number of contexts.

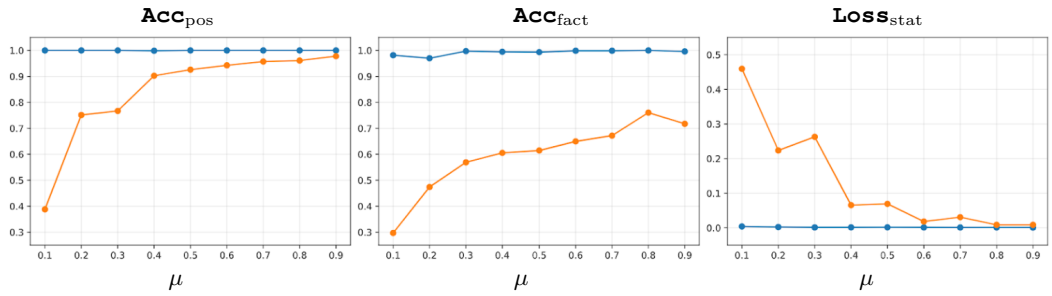


Figure 17. Non-unform representation of facts: Generalization can transfer from facts with diverse representation to other facts. We partition the facts into two groups: μK high-diversity facts and $(1-\mu)K$ low-diversity facts. The high-diversity facts appear in 9 out of 10 templates in the pretraining set, whereas the low-diversity facts appear in only 1 template. The curves report the final metric values on the high-diversity subset (blue) and the low-diversity subset (orange). As we increase μ, performance on the low-diversity facts improves sharply, suggesting partial transfer from the high-diversity facts.

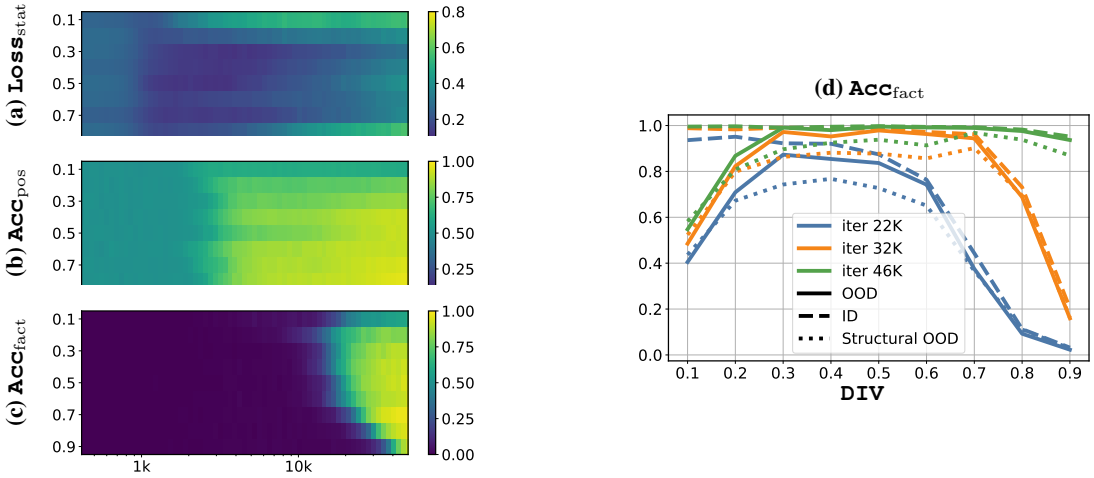


Figure 18. Structural-OOD performance in the MC10POS10 setup. (a) $\texttt{Loss}_{\text{stat}}$, **(b)** $\texttt{Acc}_{\text{pos}}$ and **(c)** $\texttt{Acc}_{\text{fact}}$ for the experimental setup of Fig. 2 on sequences drawn from structural-OOD templates, defined in Sec. C.8. Panel **(d)** replots $\texttt{Acc}_{\text{fact}}$ versus DIV, at three training checkpoints as in Fig. 3.

C.8. MC10POS10: *Structural*-OOD Data

The MC10POS10 setup enables us to evaluate a particularly challenging form of generalization beyond standard ID/OOD splits: *structural*-OOD templates that test pure compositional reasoning.

Recall that in MC10POS10, each of the $N = 10$ templates is uniquely specified by a transition matrix and position pair: $(\mathbf{P}_n, \mathbf{I}_n)$ for $n \in [N]$. As in our other contextual structure settings, we define ID and OOD templates for a given fact (a, b) based on whether they appeared with this fact pair during training.

The distinctive feature of MC10POS10 is that it allows us to test pure compositional generalization by forming new templates through pairing each $\mathbf{P}_n$ with a position pair $\mathbf{I}_{n'}$, for $n \neq n'$. Specifically, in this case, each mixed template combines a Markov chain and position both familiar in isolation but never jointly encountered, and the model must combine these familiar subcomponents to generalize to the structural-OOD sequences. We call sequences generated from such templates *structural*-OOD sequences.

Fig. 18 plots $\texttt{Loss}_{\text{stat}}$, $\texttt{Acc}_{\text{pos}}$ and $\texttt{Acc}_{\text{fact}}$ on *structural*-OOD sequences in the 4-layer experimental setup of Fig. 8. The qualitative trends mirror our standard OOD findings: Under low diversity levels, factual recall fails catastrophically; mid-training, optimal performance is achieved with intermediate diversity levels; and high diversity helps recovering the performance in long training regimes. However, the absolute value of the metrics remains lower than on the original OOD split, hinting the model needs even longer training to master completely novel template combinations. Position accuracy

shows a similar pattern: Sharp failure at low diversity, but little sensitivity to diversity elsewhere. Interestingly, $\texttt{Loss}_{\text{stat}}$ is hit hardest by these novel template compositions at test time. We leave a deeper investigation for future work, In panel (d), we further visualize, as in Fig. 3, that extended training uncovers the long-term benefits of high diversity, even though it may slow progress during the intermediate phase.

D. Minimal Setting to Understand the Impact of Low Diversity

Focusing on factual recall, we consider a minimal toy setting with N templates and $K = N$ fact pairs with sequences of length $T = 2 \times N$. For further simplicity compared to our other settings, we let the generic tokens $\mathcal{V}_{\mathcal{D}}$ to be drawn from uniform distribution over the $V_{\mathcal{D}} = 3$ tokens. For the source-target pair, we define each template $n \in [N]$ by a position pair $\mathbf{I}_n = (n, n+N)$. We compare the performance on two diversity levels: low diversity $\texttt{DIV} = 1/N$ and high diversity $\texttt{DIV} = (N-1)/N$.

Fig. 19 reports OOD factual recall in this setup for two diversity levels high (blue) and low (red) in three minimal settings: (a) $N = 3$ with a 1-layer model, (b) $N = 3$ with a 4-layer model, (c) $N = 5$ with a 4-layer model. In all cases ID performance reaches 100% by the end of training, so we only show the OOD results. As seen in panel (a), a 1-layer model is expressive enough to achieve perfect factual recall performance on the task, as it achieves perfect OOD (and ID) factual recall when trained with high diversity. However, with low diversity, the training algorithm fails to find this generalizing solution. Instead it converges to a solution that generalizes for the ID templates, but does not necessarily perform well on OOD templates. Increasing the model capacity roughly helps with the performance in the low-diversity case as shown in panel (b). However, increasing the task complexity by simply increasing N, the same large model of panel (b), fails again at finding the generalizing solution.

We can formally think of this failure under low-diversity as follows. Following the notation in Sec.2, the ultimate learning goal is to find model parameters $\boldsymbol{\theta}^*$ that minimize the next-token prediction (NTP) loss over the complete distribution over the choice of the templates and facts, i.e.,

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}} \left\{ \mathcal{L}_{\text{tot}}\left(\boldsymbol{\theta}\right) := \sum_{k \in [K]} \sum_{n \in [N]} \mathbb{E}_{\mathbf{x} \sim \{\mathcal{D}_n^k\}} \boldsymbol{\ell}_{\text{NTP}}\left(\mathbf{x};\, \boldsymbol{\theta}\right) \right\},$$

where $\mathcal{D}_n^k$ is the distribution over sequences drawn from the n-th template with the fact placeholders filled with the k-th fact (a_k, b_k), and $\boldsymbol{\ell}_{\text{NTP}}$ is the NTP loss on sequence $\mathbf{x}$ parameterized by model parameters $\boldsymbol{\theta}$. Note here that the total loss averages over *all* N templates. We assume henceforth that the model is sufficiently expressive such that $\mathcal{L}_{\text{tot}}(\boldsymbol{\theta}^*)$ attains the loss lower bound (over all possible parameterization). This is the case in all our settings.

We can now decompose this loss into two components as $\mathcal{L}_{\text{tot}}\left(\boldsymbol{\theta}\right) = \mathcal{L}_{\text{ID}}\left(\boldsymbol{\theta}\right) + \mathcal{L}_{\text{OOD}}\left(\boldsymbol{\theta}\right)$, where $\mathcal{L}_{\text{ID}}(\cdot)$ aggregates the ID templates and the complement $\mathcal{L}_{\text{OOD}}(\cdot)$ term contains the OOD templates for each fact. Concretely, let

$$\mathcal{L}_{\text{ID}}\left(\boldsymbol{\theta}\right) := \sum_{k \in [K]} \sum_{n\,:\,\mathbf{M}_{\text{in}}[n,k]=1} \mathbb{E}_{\mathbf{x} \sim \{\mathcal{D}_n^k\}} \boldsymbol{\ell}_{\text{NTP}}\left(\mathbf{x};\, \boldsymbol{\theta}\right),$$
$$\mathcal{L}_{\text{OOD}}\left(\boldsymbol{\theta}\right) := \sum_{k \in [K]} \sum_{n\,:\,\mathbf{M}_{\text{in}}[n,k]=0} \mathbb{E}_{\mathbf{x} \sim \{\mathcal{D}_n^k\}} \boldsymbol{\ell}_{\text{NTP}}\left(\mathbf{x};\, \boldsymbol{\theta}\right).$$

During training, where we only get access to a subset of facts-template pairs (k, n) for which $\mathbf{M}_{\text{in}}[n, k] = 1$, we are essentially minimizing $\mathcal{L}_{\text{ID}}(\boldsymbol{\theta})$. Intuitively this is the case because recall that we train the model such that at each iteration we see a fresh sequence $\mathbf{x}$ sampled from the ID templates and thus in the long run of many iterations, the training loss closely approximates the ID population loss $\mathcal{L}_{\text{ID}}(\boldsymbol{\theta})$. This is also empirically verified, since with sufficiently long training we always reach 100% ID accuracies. Thus, during training we find model parameters $\boldsymbol{\theta}_{\text{ID}}$ that minimize the ID population risk, i.e.,

$$\boldsymbol{\theta}_{\text{ID}} \in \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{ID}}(\boldsymbol{\theta}).$$

We remark that the set of minimizers can possibly contain multiple solutions (and we shortly argue that it does!). Also note that in the assumed setting of $\mathcal{L}_{\text{tot}}(\boldsymbol{\theta}^*)$ attaining the total-loss lower bound, a minimizer $\boldsymbol{\theta}^*$ of $\mathcal{L}_{\text{tot}}$ is certainly a minimizer of the ID risk.

The interesting question is: *Does training find model parameters $\boldsymbol{\theta}_{ID}$ that not only minimize the ID risk, but additionally minimize the OOD risk?* If that is the case, then $\boldsymbol{\theta}_{\text{ID}} = \boldsymbol{\theta}^*$, i.e., $\boldsymbol{\theta}_{\text{ID}}$ is a minimizer of the total loss $\mathcal{L}_{\text{total}}$.

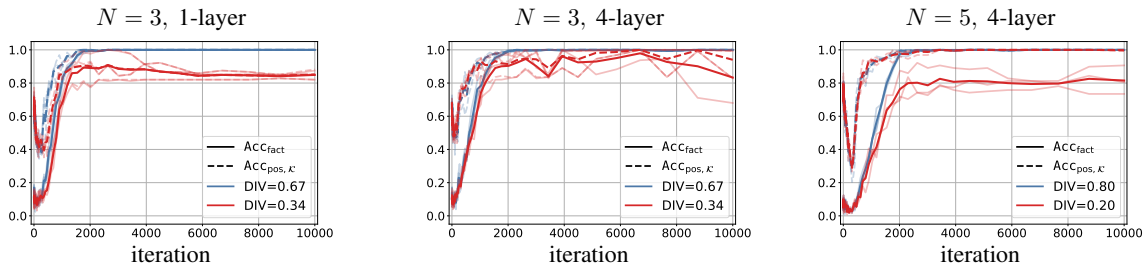


Figure 19. Minimal setup to replicate the impact of diversity. Factual recall $\mathtt{Acc}_{\mathrm{fact}}$ (solid) and target-position accuracy $\mathtt{Acc}_{\mathrm{pos},\mathcal{K}}$ (dashed). See Sec. D for discussion.

Our experiments (both in the original setup of Fig. 2 and even more evidently in the minimal setup of this section) reveal a compelling diversity-dependent dichotomy: On the one hand, under low-diversity, training converges to non-generalizing solutions that while they minimize $\mathcal{L}_{\mathrm{ID}}(\boldsymbol{\theta})$, they do not minimize the OOD risk $\mathcal{L}_{\mathrm{ID}}(\boldsymbol{\theta})$. Thus, $\boldsymbol{\theta}_{\mathrm{ID}}$ is a minimizer of the ID risk $\mathcal{L}_{\mathrm{ID}}(\boldsymbol{\theta})$ but a different one than the total loss minimizer $\boldsymbol{\theta}^*$. On the other hand, as diversity increases, training finds generalizing solutions, i.e. $\boldsymbol{\theta}_{\mathrm{ID}}$ is now a minimizer of both the ID and the total loss.

This dichotomy admits two possible explanations. With increasing diversity, either (1) the non-generalizing solutions are removed from the set of global optimizers of the ID loss $\mathcal{L}_{\mathrm{ID}}$, or (2) the landscape of the ID loss becomes more benign around the generalizing solutions (aka $\boldsymbol{\theta}^*$), which in turn makes it easier for the model to find them. Fig. 19-(b) also suggests that increased model capacity can partially help with making the ID landscape more benign.

Precisely characterizing how the context diversity and model capacity reshape the ID loss landscape is an exciting direction for future work. We believe the minimal setup and intuitions introduced in this section can serve as a starting point for such analysis.

D.1. Impact of Diversity on Implicit Bias

Here, we theoretically show on a simplified setting, motivated by our data model in main paper, how diversity can indeed lead to generalization failure.

We consider a model $f : [N] \times [K] \to \mathbb{R}^K$ that mirrors the *fact* and *template* structure of the data setup in Sec. 2. The model is parameterized by K token embeddings $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_K] \in \mathbb{R}^{d\times K}$ and N template/position embeddings $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_N] \in \mathbb{R}^{d\times N}$. For each pair (n, k), with ground-truth label k, the model computes a linear projection of the token embedding $\mathbf{z}_k$ and position embedding $\mathbf{p}_n$. Concretely, the model computes

$$f(n, k) = \mathbf{A}(\mathbf{z}_k + \mathbf{p}_n),$$

where $\mathbf{A} \in \mathbb{R}^{K\times d}$ represents a fixed projection. We train this model by minimizing cross-entropy loss on the ID (n, k) pairs defined by the binary exposure matrix $\mathbf{M}_{\mathrm{in}}$.

$$\mathcal{L}_{\mathrm{ID}}(f) := \sum_{k\in[K]} \sum_{n\,:\,\mathbf{M}_{\mathrm{in}}[n,k]=1} \boldsymbol{\ell}_{\mathrm{CE}}\left(f(n, k), \mathbf{e}_k\right).$$

For simplicity, let $d = K$ and $\mathbf{Z} = \alpha\mathbb{I}_K$ with trainable α. Reparameterizing $\mathbf{AP} = \mathbf{Q}$, the model becomes $f(n, k; \alpha, \mathbf{Q}) = \alpha\mathbf{A}\mathbf{e}_k + \mathbf{q}_n$. We further assume the columns of $\mathbf{A}$ satisfy $\arg\max_c \mathbf{A}_{ck} = k$, i.e., the maximum entry of each column lies on the diagonal, which we confirm empirically to be the case when training $\mathbf{A}$ jointly with the rest of the parameters.

Note that regardless of the choice of $\mathbf{M}_{\mathrm{in}}$, a generalizing minimizer of $\mathcal{L}_{\mathrm{ID}}$ exists at infinity by taking $\alpha \to \infty$ and $\mathbf{Q} = 0$. However, this might not be the only minimizer. Since the model is linear, we can invoke the implicit bias results to characterize where the trained model converges (Soudry et al., 2018; Lyu & Li, 2020).

Proposition D.1. *Let $\alpha_t, \mathbf{Q}_t$ be the iterates of gradient descent on $\mathcal{L}_{ID}$. As $t \to \infty$, the iterates converge in direction to the solution of the following SVM:*

$$\alpha_*, \mathbf{Q}_* = \arg\min_{\alpha,\mathbf{Q}} \ \alpha^2 + \|\mathbf{Q}\|_F^2 \quad s.t. \quad \alpha(\mathbf{A}_{kk} - \mathbf{A}_{ck}) + (\mathbf{Q}_{kn} - \mathbf{Q}_{cn}) \geq 1, \quad \forall (n, k) : \mathbf{M}_{\mathrm{in}}[n, k] = 1, \ \forall c \neq k. \quad (9)$$

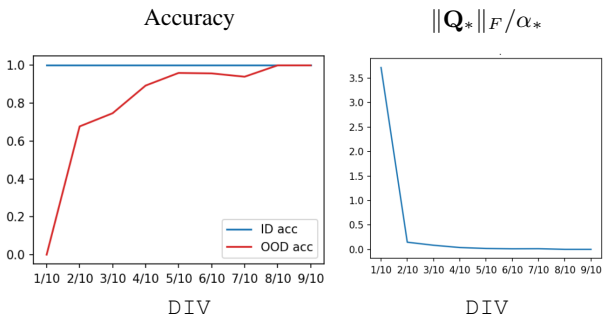


Figure 20. Impact of diversity on the SVM solutions. Optimal solutions of the SVM (9) for different diversity levels `DIV`, for $N = K = 10$, and 10 random projection $\mathbf{A}$. **(left)** Accuracy of $f(n, k; \alpha_*, \mathbf{Q}_*)$ on the ID/OOD pairs (n, k). **(right)** Relative norm of $\mathbf{Q}_*$ to α_* as diversity increasees. See Sec. D.1 for details.

Crucially, any feasible point in this program corresponds to a global minimizer of the ID loss. Also, since the projection matrix satisfy $\mathbf{A}_{kk} - \mathbf{A}_{ck} > 0$, a perfectly generalizing solution, one where the model ignores template cues ($\mathbf{Q} = 0$) and relies solely on the invariant fact signal ($\alpha > 0$), is always feasible. However, whether the model actually generalizes depends on the implicit bias of the optimization: it will converge to the specific feasible point that minimizes the L_2 norm, and this solution depends on the set of constraints that are specified by the exposure matrix $\mathbf{M}_{\text{in}}$.

In Fig. 20, we show how the optimal $\alpha_*, \mathbf{Q}_*$ changes with diversity level `DIV`, for randomly generated $\mathbf{A}$ matrices. We observe that as diversity increases, $\|\mathbf{Q}_*\|_F/\alpha_*$ decreases to zero and the prediction accuracy on OOD (n, k)'s reach from 0 to 1. This confirms that higher diversity makes the generalizing solution the unique min-norm optimizer by reshaping the set of margin constraints.

Below, we show for a special case of $\mathbf{A}$ how low diversity fails at generalizing on the OOD set.

Proposition D.2. *Let $N = K$ and assume $\mathbf{M}_{\text{in}}$ is balanced (each row/column has `DIV`$\cdot N$ true entries). Let $\mathbf{A} = \beta \mathbb{I}_K$ with $\beta < 1/K$. If `DIV`$= 1/K$, the min-norm solution yields zero OOD accuracy. If `DIV`$> 1/K$, the model recovers the generalizing solution $\mathbf{Q}_* = 0$, achieving perfect OOD accuracy.*

Proof. With $\mathbf{A} = \beta \mathbb{I}_K$, the SVM reduces to minimizing $\alpha^2 + \|\mathbf{Q}\|_F^2$ subject to $\alpha\beta + \mathbf{Q}_{kn} - \mathbf{Q}_{cn} \geq 1$ for all ID pairs (n, k).

With `DIV` $= 1/K$, i.e., $\mathbf{M}_{\text{in}} = \mathbb{I}$, each column of $\mathbf{Q}$ appear in the constraints for a single k and for each k we have $K - 1$ constraints of the form $\alpha\beta + \mathbf{Q}_{kk} - \mathbf{Q}_{ck} \geq 1$, $c \neq k$. By symmetry, $\mathbf{Q}$ takes a structure where its diagonal elements $\mathbf{Q}_{kk} = x$ and off-diagonals $\mathbf{Q}_{ck} = y$. The problem then reduces to $\arg\min_{\alpha,x,y} \alpha^2 + K\,x^2 + K(K-1)y^2$ s.t. $\alpha\beta + \beta(x - y) \geq 1$ with optimal values $\alpha_* = \frac{K\beta}{1+K\beta^2}$ and $\mathbf{Q}_* = \frac{1}{1+K\beta^2}\mathbb{I}$. Then, $\arg\max f(n, k) = n \neq k$ for OOD (n, k) since $\beta < 1/K$. Thus, the trained model will not generalize to unseen (n, k) pairs in the training set.

When `DIV` $> 1/K$, for each n there exists $c \neq k$ such that $\mathbf{M}_{\text{in}}[n, k] = \mathbf{M}_{\text{in}}[n, c] = 1$. Let $x = \mathbf{Q}_{kn} - \mathbf{Q}_{cn}$. Then, for any feasible point $\mathbf{Q}$, we have $\alpha\beta + \beta\,x \geq 1$ as well as $\alpha\beta - \beta\,x \geq 1$ which implies by summing the two inequalities that $\alpha \geq 1/\beta$. This means that the optimal objective value is $\alpha_*^2 + \|\mathbf{Q}_*\|_F^2 \geq 1/\beta^2$. However, $\alpha = 1/\beta, \mathbf{Q} = 0$ achieves the lower bound, and since the objective is strongly convex, this is the unique minimizer. □

E. Representation Analysis

Building on our experiments probing the model's last-layer representations (Sec. 4.1), we now examine the clustering properties of hidden representations from all layers and the impact of training diversity.

To analyze the structure of these internal representations, for a given sequence drawn from the n-th template and carrying fact (a_k, b_k), we probe hidden layer representations (for any layer ℓ) at source fact position i_n and test whether the transformer encodes a *template-invariant* representations of each fact (a_k, b_k).

For every (fact, template) pair (both ID and OOD), we first sample M sequences. For each sequence, we collect hidden vectors $\mathbf{h}_\ell^{(i_n)}$ at source fact position i_n from the ℓ-th transformer layer. For each layer ℓ, we then stack the vectors into $\mathbf{H}^{(\ell)} \in \mathbb{R}^{P \times d}$ and keep the top $d' = \min(30, d, P)$ principal components, where d denotes the dimensionality of the hidden

layer representations, and $P = M \times N \times K$ denotes the total number of hidden vectors extracted per layer. This gives us a PCA-reduced matrix $\widetilde{\mathbf{H}}^{(\ell)} \in \mathbb{R}^{P \times d'}$ where $\widetilde{\mathbf{h}}_i^{(\ell)},\ i \in [P]$ denotes the PCA-reduced representation at fact position for the i-th sequence. Recall that N denotes the number of templates, and K denotes the number of atomic facts. We set M to 250 in our experiments.

For every layer ℓ, we take the PCA-reduced matrix $\widetilde{\mathbf{H}}^{(\ell)} \in \mathbb{R}^{P \times d'}$ and evaluate clustering quality of the vector embeddings hen they are labeled in two different ways: 1) each vector tagged with the fact index k, and 2) each vector tagged with the template index n. We measure the clustering quality with `silhouette_score` (SKLEARN). For a given hidden layer ℓ, and every representation $\widetilde{\mathbf{h}}_i^{(\ell)},\ i \in [P]$, we compute 1) $e_i^{(\ell)}$, the average Euclidean distance to all other vectors that share its label and 2) $f_i^{(\ell)}$, the smallest average distance to a group with a *different* label. Formally, if C_i is the set of indices with the same label, then

$$e_i^{(\ell)} \;=\; \frac{1}{|C_i| - 1} \sum_{j \in C_i,\, j \neq i} \big\|\widetilde{\mathbf{h}}_i^{(\ell)} - \widetilde{\mathbf{h}}_j^{(\ell)}\big\|_2, \quad f_i^{(\ell)} \;=\; \min_{C \neq C_i} \frac{1}{|C|} \sum_{j \in C} \big\|\widetilde{\mathbf{h}}_i^{(\ell)} - \widetilde{\mathbf{h}}_j^{(\ell)}\big\|_2.$$

Using these two metrics, the score for each vector embedding $\widetilde{\mathbf{h}}_i^{(\ell)}$ is defined as

$$s_i^{(\ell)} \;=\; \frac{f_i^{(\ell)} - e_i^{(\ell)}}{\max\{f_i^{(\ell)}, e_i^{(\ell)}\}} \in [-1, 1], \quad i \in [P].$$

The silhouette value attains 1 when $\widetilde{\mathbf{h}}_i^{(\ell)}$ lies well inside a compact cluster whose members share the same label, drops to 0 when clusters of different labels overlap, and becomes negative if the vector is closer to another label's cluster than to its own. The layer-level score $s^{(\ell)}$ is the average of these values across all vectors in the layer, i.e., $s^{(\ell)} = \frac{1}{P} \sum_{i \in [P]} s_i^{(\ell)}$. To differentiate the two labeling schemes, we denote the score as $s_{\text{fact}}^{(\ell)}$ when clusters are labeled using factual indices k, and as $s_{\text{tmpl}}^{(\ell)}$ when template indices n are used as labels. If $s_{\text{fact}}^{(\ell)}$ is high, it indicates that the representations are *template-invariant*: the hidden representations learned for any given fact a only depends on the fact itself and not the context template it appears in during training. In turn, if $s_{\text{tmpl}}^{(\ell)}$ is high, it suggests that the fact hidden representations from the same template cluster together even when the facts differ.

Figure 21 reveals three consistent trends in the clustering structure of hidden representations as training diversity grows. First, *factual identity is always the dominant organizing principle*: across layers, the fact curves sit well above the template curves, indicating stronger clustering by fact than by template. Second, this *fact-centric structure generalizes to unseen pairings*—the ID-only and ID+OOD fact curves roughly remain similar, showing that vectors from unseen fact–template combinations fall into the same clusters as their seen counterparts. Third, *greater template diversity progressively weakens template-based structure while fact-based structure remains intact*, so the gap between the two widens.

Fact Heads. To better understand how factual knowledge is stored across attention heads, we compute a per-fact head attribution heatmap. For each fact, we iteratively ablate individual heads (by zeroing their contribution) and measure the drop in the model's confidence for the correct token b. This is averaged over multiple in-distribution contexts where the fact appears, yielding a (fact × head) matrix of logit drops. Figure 22 shows these heatmaps for a model trained under low diversity (left) and high diversity (right). In the low diversity case, head importance is broadly distributed, with no head clearly emerging as critical for any fact. By contrast, in the high diversity regime, certain heads show strong and localized importance for specific facts, suggesting that the model has developed specialized storage heads. This points to a more structured encoding strategy that emerges only when the model sees the same fact across various different templates.

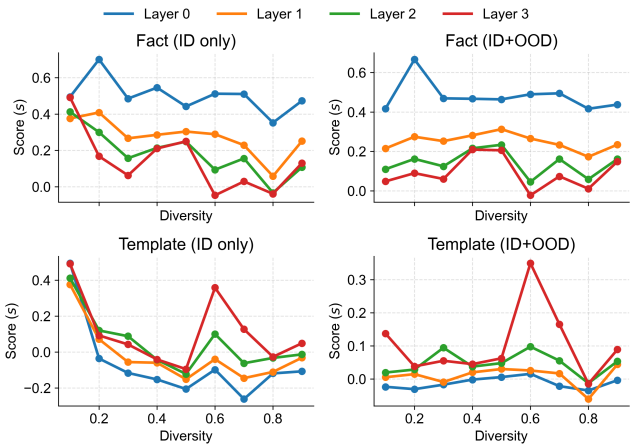


Figure 21. Clustering quality (see Sec. E) of hidden representations as a function of training diversity (DIV) for the MC1POS10 template type. **Top**: Clustering scores when representations are labelled based on factual identity k. **Bottom**: Clustering scores when representations are labeled based on template identity n. *Left* shows computation of score using representations of *ID* (template, fact) pairs. *Right* shows the clustering score using representations of both *ID* and *OOD* (template, fact) pairs. **Fact clustering dominates:** In both "Fact" panels (top row) every layer's curve sits well above the corresponding "Template" curves (bottom row). Hidden vectors therefore cluster primarily by the underlying fact rather than by the template. **Strong fact clustering persists on unseen templates:** The two fact curves—one computed on seen (ID-only) pairs, the other on the full ID + OOD set are roughly similar. Vectors for unseen fact–template combinations land in the same clusters as their seen counterparts, showing that the model abstracts the fact beyond the specific templates it saw during training. **Template invariance improves with diversity:** Moving from low to high diversity the template scores drift toward (or below) 0, while the fact scores remain roughly stable. This widening gap indicates that training on a broader mix of templates gradually removes template details from the representations while still keeping the facts separated.

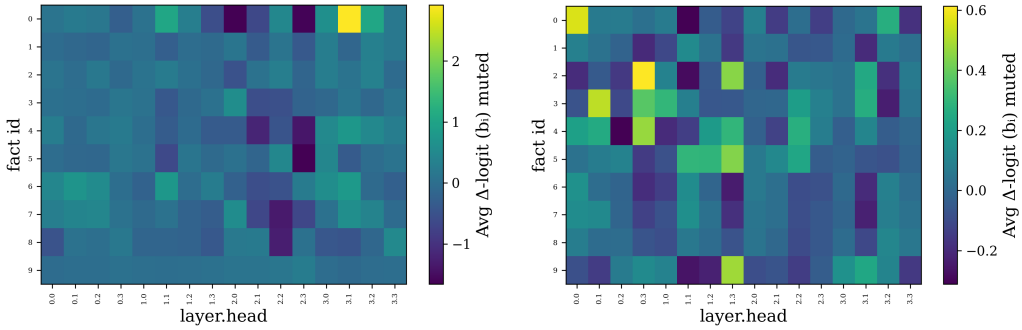


Figure 22. Per-fact head importance heatmaps in the factual recall task for the MC1POS10 template type. Each row corresponds to a different fact, and each column to a specific attention head (indexed as layer.head). The color indicates the average change in the logit of the correct answer token b_i when the corresponding head is ablated, averaged over multiple in-distribution sequences for that fact. **Left**: model trained with low diversity (DIV $= 0.1$). **Right**: model trained with high diversity (DIV $= 0.9$). In the low diversity regime, head importance is diffuse and uniform across heads, suggesting no clear specialization. In contrast, at high diversity, certain heads become more consistently important for specific facts, indicating emergent specialization and more structured factual encoding.