Stats or Facts: Decomposing Generalization in Language Models with Small-Scale Models

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

Abstract

Large language models learn both statistical patterns that make text fluent and factual associations between specific tokens that represent knowledge information. The complexity of natural language interweaving linguistic patterns and factual content challenges a systematic study of this capability. To address this, we introduce a Small-Scale Data Model (SSDM) designed to disentangle these components. The SSDM consists of a statistical stream of generic tokens, endowed with designated positional information, which composes with a separate factual stream of source-target token pairs representing knowledge. Partitioning the generating distribution of the statistical stream into sub-distributions, which we term *templates*, allows us to: (i) Independently vary the format of the templates (i.e., contextual structure) and the frequency with which facts appear within each template during training (i.e., contextual diversity); (ii) Measure both in-distribution and out-of-distribution generalization; and (iii) Distinguish between statistical, structural, and factual aspects of language model generalization. We demonstrate the flexibility of the SSDM by reporting example findings concerning: (a) the potentially catastrophic impact of low contextual diversity on either factual recall, statistical generalization, or both, contingent on the contextual structure; (b) observed stage-wise learning dynamics; and (c) hallucination.

1. Introduction

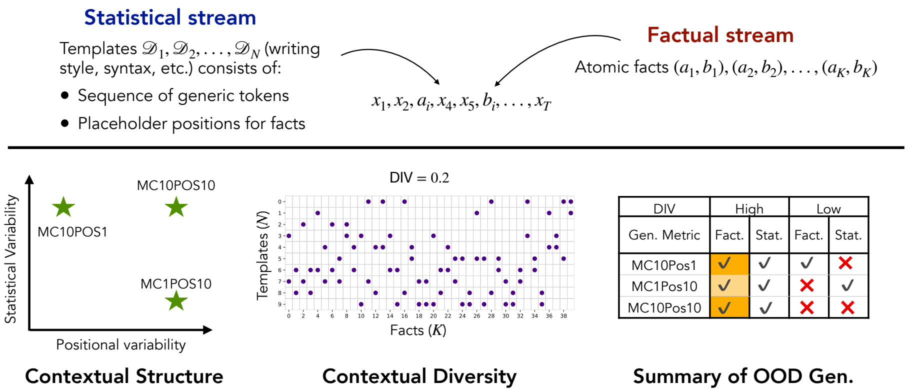
Natural language sequences carry two intertwined components. On one hand, they follow statistical and linguistic patterns – the syntax and word co-occurrence regularities that make sentences coherent and flow smoothly. On the other hand, they encode facts and relationships about the world – for example, that `Maryam Mirzakhani` won a `Fields Medal`. These two ingredients differ fundamentally: structural patterns govern the *coherency* of language, while factual content embeds *knowledge*. However, in natural text, these aspects, are tightly interlaced: each sentence simultaneously obeys linguistic patterns while conveying specific information. This distinction becomes evident when we consider statements like `The capital of France is Berlin`, which is structurally well-formed, yet contains incorrect factual information. Conversely, `Paris France capital is of the` contains factually correct information but lacks statistical coherence – demonstrating how these two components can vary independently while both are necessary for natural language.

Modern transformer-based (Vaswani et al., 2017) language models (LMs) are trained on massive corpora of such natural text sequences with one deceptively simple training objective: given a sequence of tokens, predict the next token (Radford et al., 2019). Through this next-token prediction training alone, LMs not only learn to generate contextually plausible sequences, but also learn significant amount of real-world knowledge: the model's 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., 2021; Allen-Zhu & Li, 2023; Akyürek et al., 2022; Yang et al., 2024; Meng et al., 2023; Mallen et al., 2023). This factual recall capability is remarkable precisely because the model receives no explicit fact supervision: it never encounters knowledge graphs or labeled facts or any other form of explicit differentiation between general and fact-related tokens. How can a simple next-token prediction objective guide models to distinguish and simultaneously learn these distinct types of information? Moreover, are there inherent trade-offs between acquiring linguistic structure versus factual knowledge?

Intuitively, the repetition of factual information across varying linguistic contexts during training—essentially appearing as *paraphrases* of the same underlying facts—likely plays a key role in facilitating factual recall from pure next-token

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

Accepted at Methods and Opportunities at Small Scale (MOSS), ICML 2025, Vancouver, Canada.


Statistical stream
Templates $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_N$ (writing style, syntax, etc.) consists of:
• Sequence of generic tokens
• Placeholder positions for facts
$x_1, x_2, a_i, x_4, x_5, b_i, \ldots, x_T$
Factual stream
Atomic facts $(a_1, b_1), (a_2, b_2), \ldots, (a_K, b_K)$
Statistical Variability
Positional variability
MC10POS1
MC10POS10
MC1POS10
Contextual Structure
DIV = 0.2
Templates (N)
Facts (K)
0
1
2
3
4
5
6
7
8
9
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
Contextual Diversity
Summary of OOD Gen.

DIV	High		Low	
Gen. Metric	Fact.	Stat.	Fact.	Stat.
MC10Pos1	✓	✓	✓	✗
MC1Pos10	✓	✓	✗	✓
MC10Pos10	✓	✓	✗	✗

Figure 1: Proposed testbed and summary of key findings. (Top) Our data model combines a statistical stream (templates with generic tokens and placeholder positions) and a factual stream (atomic source-target pairs), which together generate the training sequences. **(Bottom Left)** Visualization of three types of *contextual structure*—how templates vary along statistical and positional dimensions: MC10Pos1 (maximum statistical variability), MC1Pos10 (maximum positional variability), and MC10Pos10 (both statistical and positional variability). **(Bottom Middle)** Sample exposure matrix for contextual diversity DIV, with purple dots marking which templates each fact appears in. **(Bottom Right)** Summary of key findings: Low diversity impairs generalization—whether factual or statistical depends on structure—while high diversity enables both, with varying efficiencies (indicated by the intensity of the colored boxes.).

prediction supervision. But 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? If learning truly means isolating factual associations from linguistic structures, under what conditions is this possible? Moreover, is diversity only possibly impacting factual recall, or can it also impact the model's ability to learn statistical patterns?

The primary challenge in addressing these questions stems from the inherently complex nature of natural language data that intricately interleaves linguistic patterns and factual content. To disentangle the two, we introduce a small-scale synthetic model that isolates each component and lets us study their interaction under controlled conditions.

2. Small-scale data model (SSDM)

Our SSDM factorizes sequences into a *statistical* stream, representing linguistic patterns, and a *factual* stream, representing knowledge information. Our design is inspired by the synthetic-biography corpus of Allen-Zhu & Li (2023), but further abstracts the data to allow for the fine-grained control over both streams and their composition. See App. B for discussion.

Factual stream. We consider the factual stream as a set of K atomic facts $\mathcal{K} = \{(a_k \mapsto b_k)\}_{k=1}^{K}$ each given as an ordered *source-target* pair of distinct tokens from the vocabulary set $\mathcal{V} = \{1, ..., V\}$. The union $\mathcal{V}_{\mathcal{K}} = \mathcal{A} \cup \mathcal{B}$ of $\mathcal{A} = \{a_k\}_{k=1}^{K}$ and $\mathcal{B} = \{b_k\}_{k=1}^{K}$ defines the *fact vocabulary*. We denote $f : \mathcal{A} \to \mathcal{B}$ the deterministic one-to-one mapping, so each source token uniquely specifies its corresponding target $f(a_k) = b_k$. These facts represent abstract knowledge pieces with no inherent linguistic structure on their own: they become part of a language sequence only when paired with a template realization.

Statistical stream (Templates). We model templates as a mixture of N sub-distributions $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_N$, each representing a syntactic pattern, over the background sequences $\mathbf{z} = (z_1, \ldots, z_T)$ into which facts will be embedded. Sequence $\mathbf{z}$ consists of tokens from *generic vocabulary* $\mathcal{V}_{\mathcal{D}} := \mathcal{V} \setminus \mathcal{V}_{\mathcal{K}}$ with two positions reserved for the source/target fact tokens. For concreteness, we represent $\mathcal{D}_n$ by a first-order Markov chain (MC) with transition matrix $\mathbf{P}_n \in \Delta^{|\mathcal{V}_{\mathcal{D}}| \times |\mathcal{V}_{\mathcal{D}}|}$[1] and a placeholder **position pair** $\mathbf{q}_n = (q_{n,a}, q_{n,b})$, $q_{n,a} \leq T/2 < q_{n,b}$. We sample sequence $\mathbf{z} \sim \mathcal{D}_n$ by drawing an MC sequence of length $T - 2$ according to $\mathbf{P}_n$ to fill the $T - 2$ positions apart from positions $\mathbf{q}_n$ reserved for the fact tokens.

The final sequences $\mathbf{x} = (x_1, \ldots, x_T)$ are generated by sampling a fact pair (a_k, b_k), $k \sim [K]$ and a template n with context $\mathbf{z} \sim \mathcal{D}_n$; we then insert the fact in the reserved positions by setting $x_{q_a} = a_k,\ x_{q_b} = b_k$, and $x_t = z_t, \forall t \notin \{q_a, q_b\}$.

Our SSDM allows investigating generalization by precisely and independently varying: **(a) contextual diversity**—how many

[1] The choice of statistical n-gram generation in the form of MC is convenient for measuring statistical accuracy. One could adopt variants of traditional MCs (e.g., the contextual bigram (Ren et al., 2024)) without affecting our main findings.

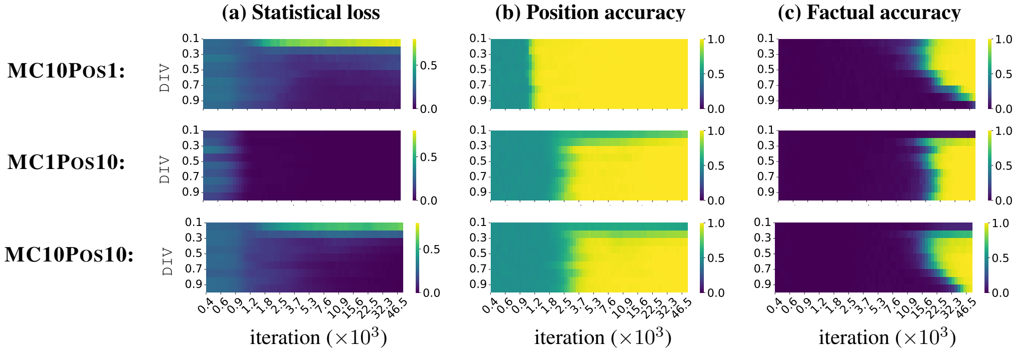


Figure 2: Generalization dynamics for varying contextual diversity levels and structures. Heatmaps show **(a)** $\texttt{Loss}_{\text{stat}}$ (5), **(b)** $\texttt{Acc}_{\text{pos}}$ (1) and **(c)** $\texttt{Acc}_{\text{fact}}$ (3) over training iterations (horizontal axis) and diversity level (vertical axis) for the three contextual structures defined in Sec. 2. Metrics are evaluated over OOD fact–template pairs. Brighter colors denote higher *loss* for $\texttt{Loss}_{\text{stat}}$ and higher *accuracy* for $\texttt{Acc}_{\text{pos}}$ and $\texttt{Acc}_{\text{fact}}$. Increasing diversity slows convergence, whereas very low diversity leads to catastrophic *statistical or/and factual* failure. See Sec. 3.1 and Apps. C and D for details.

different templates each fact appears in during training, **(b) contextual structure**—how template formats vary. See Fig. 1.

(a) Contextual diversity. Contextual diversity is controlled by a scalar $\texttt{DIV} \in (0, 1]$: each fact (a_k, b_k) is embedded in exactly $\texttt{DIV} \cdot N$ of the N templates, so larger $\texttt{DIV}$ gives richer contextual variety in the training set. These pairings are captured in a binary *in-distribution* (ID) exposure mask $\mathbf{E}_{\text{in}} \in \{0, 1\}^{N \times K}$ with $\mathbf{E}_{\text{in}}[n, k] = 1$ when fact k appears in template n. The 1's in the k-th column lists all ID templates in which fact k appears; while the zero entries mark its *out-of-distribution* (OOD) template set.

(b) Contextual structure. Contextual structure defines how templates differ in the placeholders $\mathbf{q}_n = (q_{n,a}, q_{n,b})$ and the transition matrix $\mathbf{P}_n$. To study structure independently of diversity, we fix $\texttt{DIV}$ and compare three structures, each with $N = 10$ templates: (i) **MC1POS**N keeps a common transition matrix $\mathbf{P}_n =: \mathbf{P}$ but assigns a distinct $\mathbf{q}_n$ to every template; (ii) **MC**N**POS1** fixes the position pair $\mathbf{q}_n =: \mathbf{q}$ across templates while giving each template a random $\mathbf{P}_n$; and (iii) **MC**N**POS**N varies both $\mathbf{P}_n$ and $\mathbf{q}_n$ across templates, which lets us test the model when both the positional and contextual patterns shift simultaneously. See Fig. 1 for illustration of the two axes (statistical and positional) of variability.

We train transformer models on sequences drawn from the ID template-fact pairs specified by $\mathbf{E}_{\text{in}}$ and evaluate their behavior on designed prompts: For a template n and fact (a_k, b_k), we feed the model a length $T/2$ prompt that contains the source token a_k at $q_{n,a}$ and the remaining (generic) tokens drawn from $\mathbf{P}_n$, and let it autoregressively complete the remaining $T/2$ tokens. We then score three aspects on the completion: (i) **Factual recall** ($\texttt{Acc}_{\text{fact}}$) checks whether the model places the correct target $b_k = f(a_k)$ at $q_{n,b}$. (ii) **Position accuracy** ($\texttt{Acc}_{\text{pos}}$) verifies adherence to the composition rule, i.e., fact tokens appear *only* at $q_{n,b}$ and that every other position in the completion is filled with a generic token. (iii) **Statistical loss** ($\texttt{Loss}_{\text{stat}}$) measures how closely the distribution of generic tokens matches the MC statistics of the underlying template. We measure these separately for ID and OOD template–fact pairings. Formal definitions of the metrics and experimental details are provided in Apps. C.2 and D.

3. Example Findings

3.1. Impacts of contextual diversity and structure (Fig. 2)

For ID sequences, all three metrics achieve perfect accuracy regardless of structure/diversity, which only affect convergence speed. We defer the ID discussion to App. D and focus here on the more nuanced OOD performance (see Fig. 2):

(1) Improvements in $\texttt{Acc}_{\text{pos}}$ during training slow down noticeably at low diversity ($\texttt{DIV} \leq 0.2$). However, once diversity is moderate or high, $\texttt{Acc}_{\text{pos}}$ converges to 1 at the same rate for any $\texttt{DIV}$ value. With MC10POS1–where all templates share the same placeholder positions–the impact of diversity on structural accuracy becomes negligible (but see item (2) below).

(2) Diversity is also critical for learning the underlying statistical component as measured by $\texttt{Loss}_{\text{stat}}$: with $\texttt{DIV} < 0.3$, late in the training, the model's $\texttt{Loss}_{\text{stat}}$ starts growing (light bars, top-right of Fig.2-a). In MC1POS10, however, where all templates share one transition matrix, $\texttt{Loss}_{\text{stat}}$ is largely insensitive to $\texttt{DIV}$ (unlike factual recall–see item (3)).

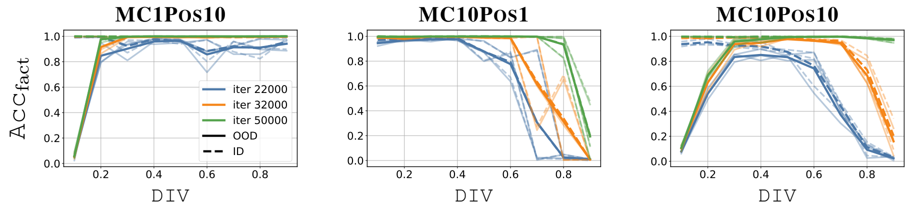


Figure 3: Diversity pays off with time. Factual accuracy versus diversity level after 22, 32 and 50k iterations. Depending on the template structure, very low diversity can remain unrecoverable—no amount of additional training restores OOD performance. High-diversity runs begin with lower accuracy but continue improving until they overtake low-diversity models, showing that diversity imposes an early cost yet delivers long-term gains. Each bold curve is average and shaded lines are individual runs for OOD (solid) and ID (dashed) data.

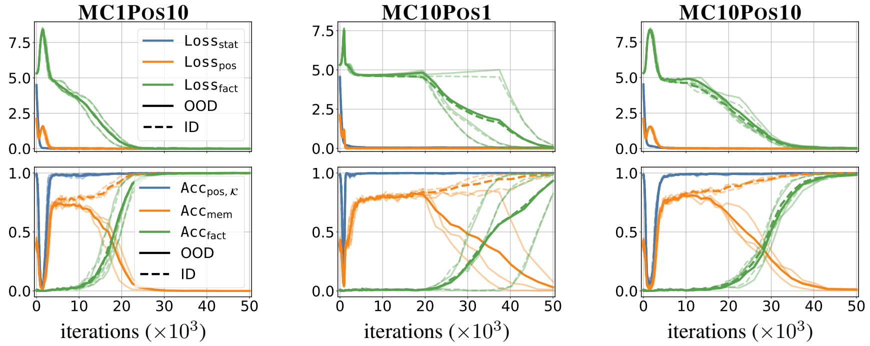


Figure 4: Stage-wise dynamics for $\texttt{DIV} = 0.8$. **(Top)** Statistical (blue), Position (orange), and factual-recall (green) losses. The model first matches the MC statistics, then learns the structure by identifying the position for the generic vs fact tokens, and only finally learns source–target mappings. **(Bottom)** Position accuracy for the target position (blue), memorization accuracy (orange), and factual accuracy (green). See Secs. 3.2, 3.3 and C.2.

(3) $\texttt{Acc}_{\text{fact}}$ shows the most intricate pattern: High diversity ($\texttt{DIV} = 0.9$), slows down improvements, while very low diversity ($\texttt{DIV} \leq 0.2$) leads to catastrophic factual errors. Contextual structure shapes the severity of these patterns: in MC1Pos10 the low-diversity collapse is severe and the high-diversity slowdown mild; in MC10Pos1 the fixed positional cue means low diversity merely delays progress without causing total failure. In MC10Pos10, both the severe collapse at low diversity and the slower convergence at high diversity are clearly apparent.

Diversity pays off with time. Fig. 3 plots $\texttt{Acc}_{\text{fact}}$ versus $\texttt{DIV}$ at three training checkpoints. With short training budget, intermediate diversity gives the best OOD recall. With long training, however, the most diverse setup catches up, while low diversity setups never recover—except in the MC10Pos1 that carries a strong positional cue for the facts. Thus, the optimal $\texttt{DIV}$ depends jointly on the training budget and the structure of the templates.

3.2. Stage-wise learning dynamics (Fig. 4-top)

In Fig. 4-(top row), we track the evolution of the statistical loss $\texttt{Loss}_{\text{stat}}$, structural loss $\texttt{Loss}_{\text{pos}}$, and factual loss $\texttt{Loss}_{\text{fact}}$ (see App. C.2 for definitions) during training. We focus on the high-diversity setting ($\texttt{DIV} = 0.8$) where along all three axes the model can generalize with long enough training. The curves reveal a consistent three-phase trajectory (ID: dashed; OOD: solid): the model first matches the MC statistics of generic tokens, then learns the position rule—faster in MC10Pos1, whose fixed placeholder gives a strong positional cue—and only after these two losses plateau does $\texttt{Loss}_{\text{fact}}$ start to *slowly* drop, showing that the model is finally learning the correct source-to-target mapping f.

3.3. From hallucination to generalization (Fig. 4-bottom)

In Fig. 4-(bottom row), we examine more closely the model's output at the reserved target position. We measure: 1) factual accuracy $\texttt{Acc}_{\text{fact}}$ (green), and 2) target-position accuracy $\texttt{Acc}_{\text{pos},\mathcal{K}}$ (blue), which measures the model's accuracy at placing *any* fact (vs generic) token at the placeholder position (see App. C.2). Consistent wit Sec. 3.2, the blue curve rises sharply to 1, showing that the model quickly learns *where* to place a fact token. By this stage, the model has also reached statistical generalization as shown in Fig. 4-(top row). Yet $\texttt{Acc}_{\text{fact}}$ remains low, meaning the model often inserts the wrong fact, producing syntactically plausible but factually incorrect completions—i.e., a transient "hallucination" phase. Only after extended training does $\texttt{Acc}_{\text{fact}}$ increase, marking a transition from hallucination to factual *generalization*. We also track *memorization accuracy* $\texttt{Acc}_{\text{mem}}$ (orange) that measures the fraction of examples in which the reserved target position is filled by only the fact tokens that had already appeared in this specific placeholder at the *training* stage (see App. C.2). Shortly after hallucination begins, $\texttt{Acc}_{\text{mem}}$ also increases but plateaus at this stage. Only with further training does the plateau break–$\texttt{Acc}_{\text{mem}}$ falls while $\texttt{Acc}_{\text{fact}}$ rises.

4. Conclusion

We introduce a lightweight synthetic data framework that gives us granular control over different components of data distribution, allowing us to study generalization of language models from different point-of-views. By varying contextual diversity in the training set, we quantify its impact on three generalization dimensions: statistical, structural (position), and factual accuracy. We defer details and complementary discussions to the App., where we discuss impact of model size on the observed dynamics (App. E), outline preliminary sketches for future theoretical analysis (App. F), probe the learned internal embeddings under varying diversity settings (App. G), and provide extended discussions on related works (App. B).

References

Akyürek, E., Bolukbasi, T., Liu, F., Xiong, B., Tenney, I., Andreas, J., and Guu, K. Towards tracing factual knowledge in language models back to the training data. *arXiv preprint arXiv:2205.11482*, 2022.

Allen-Zhu, Z. and Li, Y. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 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.

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

Edelman, E., Tsilivis, N., Edelman, B. L., Malach, E., and Goel, S. The evolution of statistical induction heads: In-context learning markov chains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qaRT6QTIqJ.

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'sfactual'predictions. *arXiv preprint arXiv:2207.14251*, 2022.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

Geva, M., Bastings, J., Filippova, K., and Globerson, A. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023.

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.

Li, S., Li, X., Shang, L., Dong, Z., Sun, C., Liu, B., Ji, Z., Jiang, X., and Liu, Q. How pre-trained language models capture factual knowledge? a causal-inspired analysis. *arXiv preprint arXiv:2203.16747*, 2022.

Liu, L. Z., Wang, Y., Kasai, J., Hajishirzi, H., and Smith, N. A. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*, 2021.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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.

Makkuva, A. V., Bondaschi, M., Girish, A., Nagle, A., Jaggi, M., Kim, H., and Gastpar, M. Attention with markov: A framework for principled analysis of transformers via markov chains, 2024. URL https://arxiv.org/abs/2402.04161.

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 Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

Nichani, E., Lee, J. D., and Bietti, A. Understanding factual recall in transformers via associative memories. *arXiv preprint arXiv:2412.06538*, 2024.

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. 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., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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. *arXiv preprint arXiv:2410.23438*, 2024.

Roberts, A., Raffel, C., and Shazeer, N. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020.

Sun, K., Xu, Y. E., Zha, H., Liu, Y., and Dong, X. L. Head-to-tail: how knowledgeable are large language models (llms)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yang, Z., Band, N., Li, S., Candes, E., and Hashimoto, T. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*, 2024.

Zucchet, N., Bornschein, J., Chan, S., Lampinen, A., Pascanu, R., and De, S. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*, 2025.

Table 1: Table of notations.

Symbol	**Meaning**
T	Sequence length.
$\mathbf{x} = (x_1, \ldots, x_T)$	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.
$f : \mathcal{A} \to \mathcal{B}$	One-to-one mapping, $f(a_k) = b_k$.
Statistical stream (templates)	
N	Number of templates.
$\mathcal{D}_n$	Distribution of template n.
$\mathcal{V}_D = \mathcal{V} \setminus \mathcal{V}_K$	Generic vocabulary.
$\mathbf{q} = (q_a, q_b)$	Placeholder positions ($q_a \leq T/2 < q_b$).
$\mathcal{P}_n$	Set of transition matrices of template n.
$\mathcal{Q}_n$	Set of position pairs of template n.
$\mathbf{P}_n$	A transition matrix drawn from $\mathcal{P}_n$.
Exposure and diversity	
$\mathrm{E}_{\text{in}} \in \{0,1\}^{N \times K}$	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).
$\texttt{Acc}_{\text{pos},\mathcal{K}}$	position accuracy at the target position.
$\texttt{Acc}_{\text{fact}}$	Factual accuracy (Eq. 3).
$\texttt{Loss}_{\text{stat}}$	Statistical loss (KL divergence to template MC – Eq. 5).
$\texttt{Loss}_{\text{pos}}$	Negative log-prob of placing fact tokens in the target placeholder and generic tokens elsewhere (Eq (2)).
$\texttt{Loss}_{\text{fact}}$	Negative log-prob of the correct target token at the target placeholder (Eq (4)).

A. Overview of supplementary material

Notations. We denote matrices/vectors/scalars as $\mathbf{A}/\mathbf{a}/a$ respectively. We view token-sequences as vectors and denote $\mathbf{a}_{\leq t}$ the subsequence of length t. 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 $[N] := \{1, \ldots, N\}$ and use $\mathbb{1}[\mathcal{C}]$ for the indicator function (1 if condition $\mathcal{C}$ is satisfied, 0 otherwise).

Overview.. Section B presents a detailed comparison with relevant works in the literature. In Sections C and D we gather the details omitted from the Secs 2 and 3 respectively. Section E provides additional experimental results and discussions. Section F introduces a minimal toy setting as a starting point for a theoretical analysis. Finally, Section G examines the model's internal embeddings and how training diversity affects them.

B. 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., 2023). A growing body of mechanistic-interpretability work explores *where* LMs store knowledge (Meng et al., 2023; Dai et al., 2021; Geva et al., 2020) and *how* they recall the correct fact (Geva et al., 2023; Lv et al., 2024). Other studies trace each learned fact back to the pre-training data to investigate which corpus patterns enable its acquisition (Elazar et al., 2022; Akyürek 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, 2023). 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, 2023; Nichani et al., 2024; 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 (2023), use a controlled synthetic biography dataset with fixed-sentence templates to examine factual recall: 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, Allen-Zhu & Li (2023) show that (i) when question templates are introduced *only* at the fine-tuning stage, factual recall is impossible unless each fact was seen in diverse contexts, i.e., in several templates, during pre-training, and (ii) even when questions are already present in pre-training data, more paraphrase diversity markedly boosts recall. In other words, successful recall requires *varied* exposure to each fact, not mere repetition in a single template.

While our data-generation scheme draws on this core insight, our synthetic testbed offers a more abstracted and finely controlled framework. By abstracting the biography setup further while retaining its template–fact structure, we gain more fine-grained control over other aspects like the statistical and structural composition, which were fixed in their work. We focus exclusively on pre-training experiments here, though the same abstract framework could be used to study fine-tuning similar to them as well. Yet, we show that even without finetuning and even in minimal settings like those described in App. F low-diversity can still be catastrophic for factual recall. 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.

Contemporaneous work by Zucchet et al. (2025) analyzes the factual recall dynamics in the same synthetic biography setup of Allen-Zhu & Li (2023), reporting a stage-wise dynamic as follows. 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 target. This observation is analogous to our stage-wise learning discussion in Sec. 3.2. 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 the same setup as Allen-Zhu & Li (2023), 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 $\mathbf{E}_{\text{in}}$ 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. (2024). 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. Within this abstract setting they prove theoretically that a single-layer transformer can memorize facts if the model size scales appropriately and they quantify how the capacity can be allocated between attention heads and MLP weights. 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 significantly 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. (2024). 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. (2024) 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.

Finally, we 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. (2024) 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.

C. Additional details of Section 2

C.1. Model probing for ID and OOD evaluation

We use controlled sequence probing to evaluate how the model's learning abilities evolve during training, with an emphasis on assessment across varying contextual diversity and contextual structures. Concretely, we prompt the model with an "incomplete" sequence $\texttt{prompt} := \mathbf{x}_{\leq T/2} = (x_1, ..., x_{T/2})$ of length $T/2$ that is sampled from a template $n \in [N]$ and includes a source token a from a factual pair $(a, b) \in \mathcal{K}$ at position $q_{n,a}$. We then allow the model to complete the sequence by generating the remaining $T/2$ tokens auto-regressively. The task requires that (i) the generic tokens generated follow the statistical pattern of $\mathcal{D}_n$, and (ii) the correct target token $b = f(a)$ appears exactly at position $q_{n,b}$.[2] For any given fact pair (a_k, b_k), the prompt sequence can be generated from a template n that was either seen (ID) or unseen (OOD) during training. We track performance separately on ID template-fact pairs (n, k) where $\mathbf{E}_{\text{in}}[n, k] = 1$ and OOD pairs where $\mathbf{E}_{\text{in}}[n, k] = 0$.

C.2. Evaluation metrics

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. To distinguish generated tokens at positions $t > T/2$ from $\texttt{prompt}$ tokens at $t \leq T/2$, denote $(\hat{x}_{T/2+1}, \ldots, \hat{x}_T)$ the tokens of the generated completion. 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. WLOG, 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 model's generated sequence should contain a token from the fact vocabulary $\mathcal{V}_\mathcal{K}$ *only* at the target position $q_{n,b}$ designated by the template, and all other positions should contain tokens from the generic vocabulary $\mathcal{V}_\mathcal{D}$. Formally, for each test sequence we define *position accuracy* and *loss* as:

$$\texttt{Acc}_{\text{pos}} = \mathbb{1}[\hat{x}_{q_{n,b}} \in \mathcal{V}_\mathcal{K}] + \Big(\sum\nolimits_{t=T/2, t\neq q_{n,b}}^{T} \mathbb{1}[\hat{x}_t \in \mathcal{V}_\mathcal{D}]\Big)/(T/2-1)\,, \tag{1}$$

$$\texttt{Loss}_{\text{pos}} = -\log(\sum_{v\in\mathcal{V}_\mathcal{K}} \hat{\mathbf{p}}_{q_{n,b}}(v)) - \frac{1}{T/2-1}\sum\nolimits_{t=T/2, t\neq q_b}^{T} \log(\sum_{v\in\mathcal{V}_\mathcal{D}} \hat{\mathbf{p}}_t(v)))\,. \tag{2}$$

We also use $\texttt{Acc}_{\text{pos},\mathcal{K}}$ to denote the first term in $\texttt{Acc}_{\text{pos}}$, i.e., $\texttt{Acc}_{\text{pos},\mathcal{K}} := \mathbb{1}[\hat{x}_{q_{n,b}} \in \mathcal{V}_\mathcal{K}]$ that measures position accuracy

[2] The desired position depends on the template n from which $\texttt{prompt}$ is sampled. Similarly, the target token also depends on the source token in $\texttt{prompt}$. For simplicity, we write $q_{n,b}$ instead of $q_{n(\texttt{prompt}),b(\texttt{prompt})}$.

only at the fact placeholder position.

2. **Factual accuracy/loss:** We define *factual accuracy* as the correct prediction of the target at the position specified by the template:

$$\texttt{Acc}_{\text{fact}} := \mathbb{1}\left[\hat{x}_{q_{n,b}} = f(a)\right] . \tag{3}$$

We can also accordingly define the *factual loss* as

$$\texttt{Loss}_{\text{fact}} := -\log\left(\hat{\mathbf{p}}_{q_{n,b}}(f(a))\right) . \tag{4}$$

3. **Statistical loss:** Denote $G_n \subseteq [T/2+1, T]$ 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:

$$\texttt{Loss}_{\text{stat}} := \left(\sum\nolimits_{t \in G_n} \mathrm{KL}\big(\tilde{\mathbf{p}}_t \,\|\, \mathbf{p}_t^*\big)\right) / |G_n| . \tag{5}$$

In Sec. 3.3, we measure an extra metric, *memorization accuracy*:

$$\texttt{Acc}_{\text{mem}} := \mathbb{1}[\hat{x}_{q_{n,b}} \in \mathcal{V}_{\mathcal{K},n}],$$

where $\mathcal{V}_{\mathcal{K},n} := \{b_k \,|\, \mathbf{E}_{\text{in}}[n,k] = 1\}$ is the set of targets that were paired with template n (of the prompt) in the training set. This metric measures the fraction of examples in which the reserved target position is filled by any targets repeated in the training set only. Note that for ID data, $\texttt{Acc}_{\text{mem}}$ goes to one, if $\texttt{Acc}_{\text{fact}}$ does so, since the correct target always belongs to $\mathcal{V}_{\mathcal{K},n}$. However, this is not the case for OOD data.

D. Additional details of Section 3

Experimental setup. In all experiments 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 $\mathcal{K}$ of $K = 100$ source-target pairs. For the MC, we let generic vocabulary set of size $|\mathcal{V}_{\mathcal{D}}| = 3$. We sweep diversity $\texttt{DIV}$ from $0.1 - 0.9$ and train the model for $50k$ iterations with AdamW (Loshchilov & Hutter, 2017) and a fixed learning rate of 10^{-4}. Unless otherwise noted, metrics are averaged over three random initializations over both model's initialization and data splits. Models were trained on a single Tesla V100-SXM2 GPU (16GB memory).

Impact of diversity on ID performance. Across the three contextual structures and the full range of diversity levels, the model ultimately reaches perfect ID performance on all three metrics. Yet, the rate at which this happens varies. As Fig. 5 show, both statistical loss ($\texttt{Loss}_{\text{stat}}$) and position accuracy ($\texttt{Acc}_{\text{pos}}$) converge to 0 and 1 respectively at nearly the same rate, largely unaffected by diversity level $\texttt{DIV}$. In contrast, factual recall ($\texttt{Acc}_{\text{fact}}$), 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. 5-(c) illustrates this delay clearly: as $\texttt{DIV}$ increases (lower rows in heat map), the yellow band marking perfect recall shifts rightward requiring additional training to reach full accuracy.

E. Additional discussion

E.1. 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: $(\mathbf{P}_n, \mathbf{q}_n)$ 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 $\mathbf{q}_{n'}$, 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

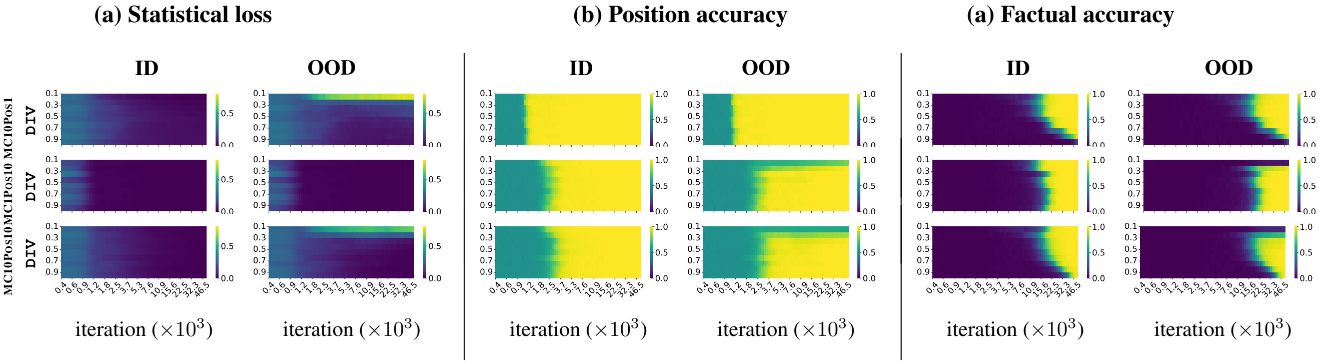


Figure 5: Generalization dynamics for varying contextual diversity levels and structures. Same as Fig. 2 with metric evaluated on both ID and OOD data.

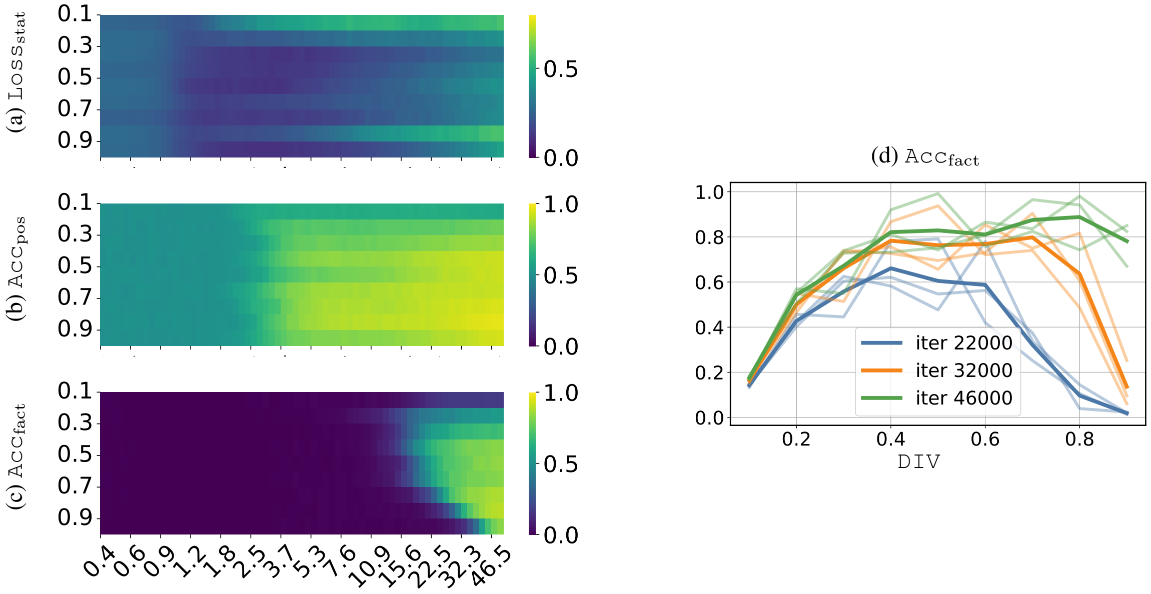


Figure 6: Structural-OOD performance in the MC10Pos10 setup. (a) $\texttt{Loss}_{\text{stat}}$, **(b)** $\texttt{Acc}_{\text{pos}}$ and **(c)** $\texttt{Acc}_{\text{fact}}$ for the experimental setup of Fig. 2 on sequences drawn from structural-OOD templates, defined in Sec. E.1. Panel **(d)** replots $\texttt{Acc}_{\text{fact}}$ versus $\texttt{DIV}$, at three training checkpoints as in Fig. 3.

these familiar subcomponents to generalize to the structural-OOD sequences. We call sequences generated from such templates *structural*-OOD sequences.

Fig. 6 plots $\texttt{Loss}_{\text{stat}}$, $\texttt{Acc}_{\text{pos}}$ and $\texttt{Acc}_{\text{fact}}$ on *structural*-OOD sequences in the same experimental setup of Fig. 2. 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–but the absolute value of the metrics remains markedly 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.

E.2. 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 Fig. 7, 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 improving *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.

F. 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 $|\mathcal{V}_{\mathcal{D}}| = 3$ tokens. For the source-target pair, we define each template $n \in [N]$ by a position pair $\mathbf{q}_n = (n, n+N)$. We compare the performance on two diversity levels: low diversity $\texttt{DIV} = 1/N$ and high diversity $\texttt{DIV} = (N-1)/N$.

Fig. 8 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.,

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

where $\mathcal{D}_n^k$ is the distribution over sequences drawn form 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 $\mathcal{L}_{\text{tot}}(\boldsymbol{\theta}) = \mathcal{L}_{\text{ID}}(\boldsymbol{\theta}) + \mathcal{L}_{\text{OOD}}(\boldsymbol{\theta})$, 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

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

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

During training, where we only get access to a subset of facts-template pairs (k, n) for which $\mathbf{E}_{\text{in}}[n,k] = 1$, 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.,

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

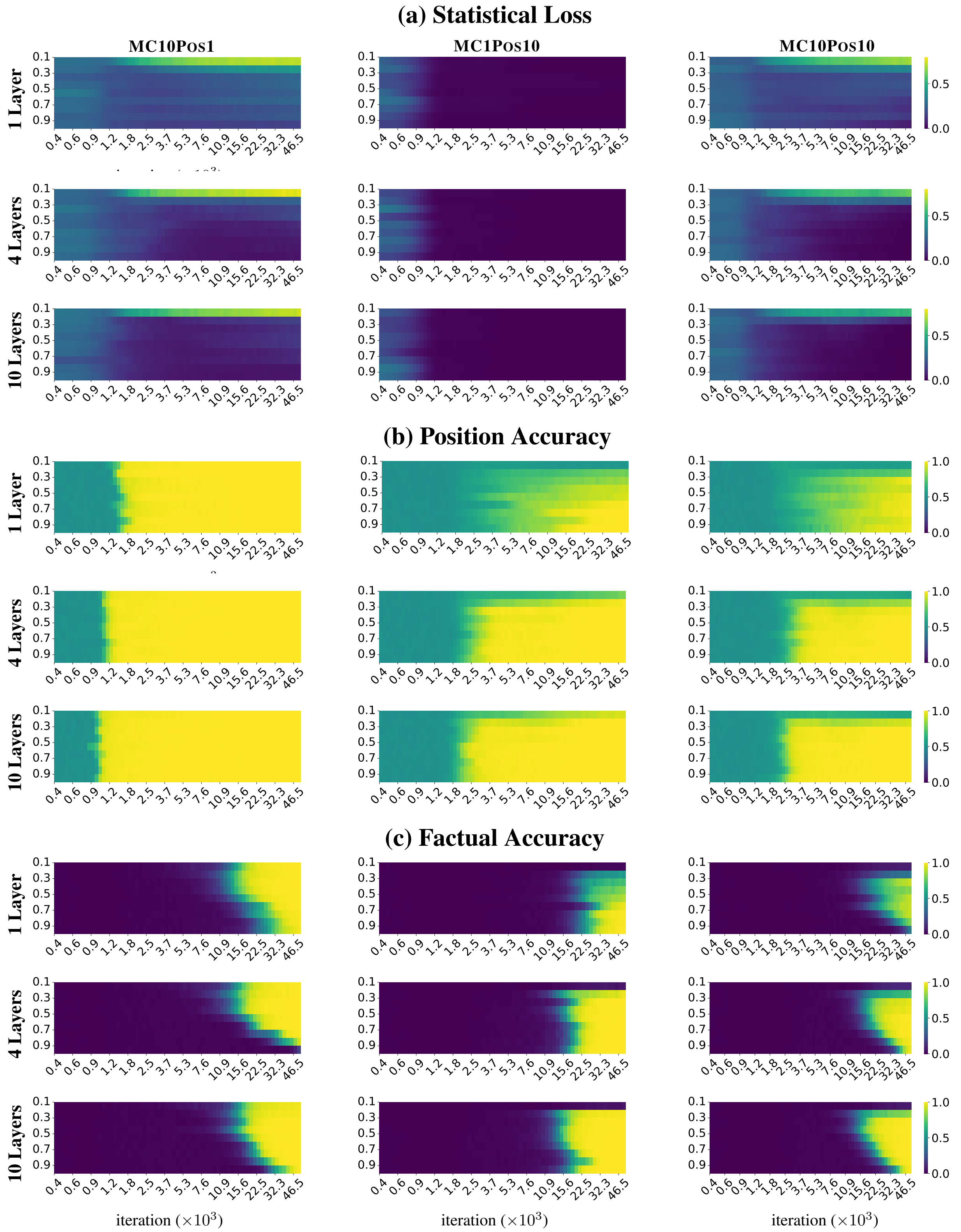


Figure 7: Impact of model size. Replication of the experiments in Sec. 3 (4-layer) with smaller (1-layer) and larger (10-layer) models on OOD sequences. With increased model capacity, we need fewer iterations to achieve the same level of performance. However, model capacity alone cannot alleviate the failure of factual recall at low diversity levels.

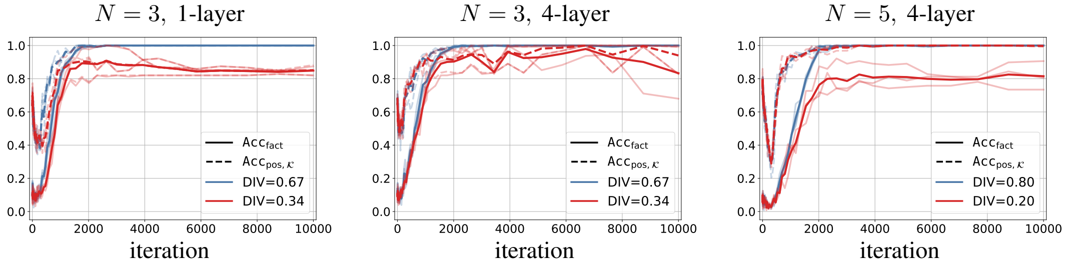


Figure 8: Minimal setup to replicate the impact of diversity. Factual recall $\mathtt{Acc}_{\mathtt{fact}}$ (solid) and target-position accuracy $\mathtt{Acc}_{\mathrm{pos},\mathcal{K}}$ (dashed). See App. F for discussion.

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 $\mathcal{L}_{\text{total}}$.

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 $\mathcal{L}_{\text{ID}}(\boldsymbol{\theta})$, they do not minimize the OOD risk $\mathcal{L}_{\text{ID}}(\boldsymbol{\theta})$. Thus, $\boldsymbol{\theta}_{\text{ID}}$ is a minimizer of the ID risk $\mathcal{L}_{\text{ID}}(\boldsymbol{\theta})$ but a different one than the total loss minimizer $\boldsymbol{\theta}^*$. On the other hand, as diversity increases, training finds generalizing solutions, i.e. $\boldsymbol{\theta}_{\text{ID}}$ 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 $\mathcal{L}_{\text{ID}}$, 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. 8-(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.

G. Representation Analysis

To analyze the structure of the model's 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 fact position $q_{n,a}$ 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 $\mathbf{h}_{\ell}^{(q_{n,a})}$ at fact position $q_{n,a}$ 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

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

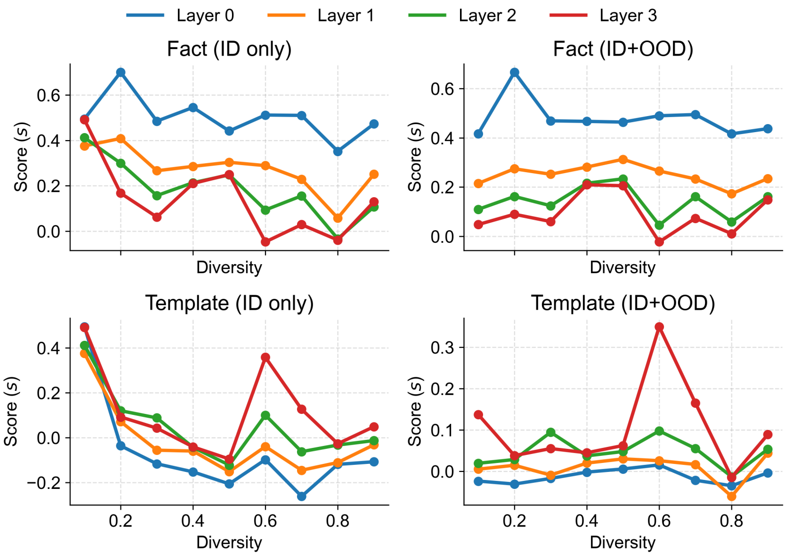


Figure 9: Clustering quality (see Sec. G) of hidden representations as a function of training diversity (DIV) for the MC1Pos10 template type. **Top**: Clustering scores when representations are labelled based on factual identity k. **Bottom**: Clustering scores when representations are labeled based 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.

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 9 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 surface form. 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 q_b. This is averaged over multiple in-distribution contexts where the fact appears, yielding a (fact × head) matrix of logit drops. Figure 10 shows these heatmaps for a model trained under low

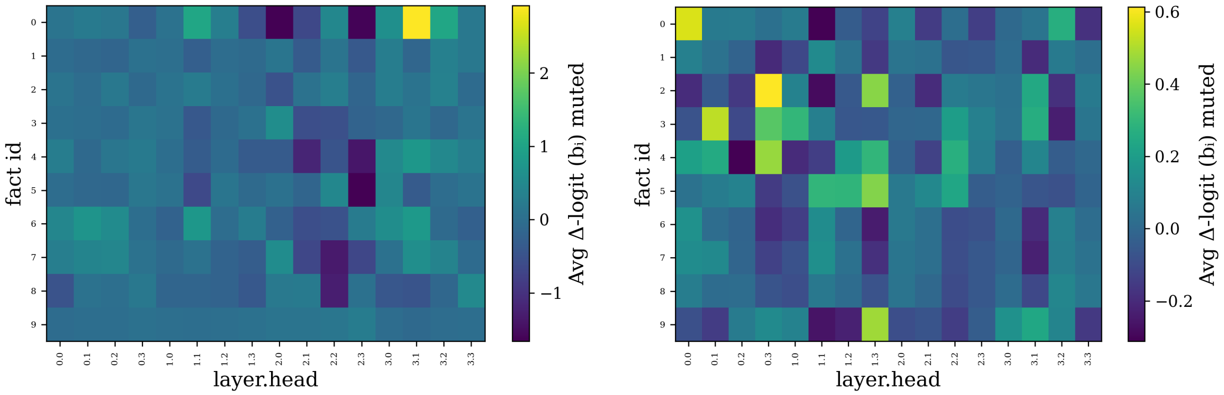


Figure 10: Per-fact head importance heatmaps in the factual recall task for the MC1Pos10 template type. Each row corresponds to a different fact, and each column to a specific attention head (indexed as layer.head). The color indicates the average change in the logit of the correct answer token b_i when the corresponding head is ablated, averaged over multiple in-distribution sequences for that fact. **Left**: model trained with low diversity ($\texttt{DIV} = 0.1$). **Right**: model trained with high diversity ($\texttt{DIV} = 0.9$). 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.

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.