Understanding Fact Recall in Language Models: Why Two-Stage Training Encourages Memorization but Mixed Training Teaches Knowledge

Ying Zhang[1] Benjamin Heinzerling[1,2] Dongyuan Li[3] Kentaro Inui[1,4]
[1]RIKEN Center for Advanced Intelligence Project [2]Tohoku University
[3]The University of Tokyo [4]MBZUAI
{ying.zhang,benjamin.heinzerling}@riken.jp lidy@csis.u-tokyo.ac.jp
kentaro.inui@mbzuai.ac.ae

Abstract

Fact recall, the ability of language models (LMs) to retrieve specific factual knowledge, remains a challenging task despite their impressive general capabilities. Common training strategies often struggle to promote robust recall behavior with two-stage training, which first trains a model with fact-storing examples (e.g., factual statements) and then with fact-recalling examples (question–answer pairs), tending to encourage rote memorization rather than generalizable fact retrieval. In contrast, mixed training, which jointly uses both types of examples, has been empirically shown to improve the ability to recall facts, but the underlying mechanisms are still poorly understood. This research investigates how these training strategies relate to their ability to recall facts. Our analysis on synthetic fact recall datasets with the Llama-3.2B model reveals that mixed training encouraging format-invariant representations across both fact-storing and fact-recalling contexts. These findings suggest that maintaining representation consistency between storage and retrieval may play a key role in enabling LMs to generalize factual knowledge across task formulations.

1 Introduction

If someone knows that "Barack Obama was born in Hawaii," they should easily answer the question "Where was Barack Obama born?" We refer to this process of answering questions using previously acquired knowledge as "**Fact Recall**". Fact recall is an important capability involved in many aspects of language understanding, but is a challenging problem for LMs trained via next-token prediction based on word co-occurrence [1], as this objective encourages **rote memorization**. For example, a model trained on "Barack Obama was born in Hawaii" can predict "Hawaii" when prompted with "Barack Obama was born in," but fails on "Where was Barack Obama born?" [2]. This differs from human fact recall, which naturally involves a deeper understanding to abstract the entities "Barack Obama" and "Hawaii" and identify their relation "was born in."

Recent work has explored how LMs store and recall facts via specific neurons and attention patterns [3, 4, 5, 6, 7, 8, 9, 10], and developed training strategies to improve fact recall [2, 11]. Notably, Zhu et al. [2] show that **two-stage training**, which first trains on fact-storing (e.g., factual statements) and then on fact-recalling (question–answer pairs), primarily leads to rote memorization characterized by very low (9.7%) fact recall accuracy on unseen questions. In contrast, **mixed training**, which mixes fact-storing and fact-recalling examples, helps the model learn facts as generalizable knowledge that can be retrieved across different query forms. We refer to this ability as teaching knowledge, which is evidenced by a much higher accuracy (88.6%) in fact recall. Although promising, this raises key questions: **Why does two-stage training appear to encourage memorization while mixed training appears to teach knowledge?**

In this study, we address this question by examining the internal representations of LMs. We find that while two-stage training results in divergent representations for fact-storing and fact-recalling, mixed training promotes format-invariant representations. In detail, we applied both mixed training and two-stage training to fine-tune the

LM: Llama-3.2-3B, using biographical statements (BIO) for fact-storing and question–answer pairs (QA) for fact-recalling. By training separate linear projections to probe the model's top layers, we disentangle the effects of encoded features across different input formats from their associated extraction logic. Our results indicate that the suboptimal recall performance of two-stage training is not primarily driven by a complete loss of stored knowledge or the structural format differences between BIO and QA. Instead, it is caused by representation inconsistency stemming from differences in sequence length or the distal position of attributes in longer sequences. These findings reveal how mixed training facilitates knowledge retrieval by enforcing representation consistency across varying contexts. Our work sheds light on how optimization strategies fundamentally shape the internal mechanisms of LMs.

2 Preliminaries

2.1 Preliminary backgrounds

Factual Knowledge: We define a fact as a triple (s, r, a) that maps a subject entity s and a relation type r to an attribute a. For example, (*Barack Obama, wasBornOn, August-4-1961*). We present each fact in two natural language formats: a *biographical statement* (BIO) and a *question–answer* pair (QA). In the BIO format, the fact is expressed as a declarative sentence (e.g., *Barack Obama's life journey began on August 4, 1961*). In the QA format, the fact is expressed as a question about the subject, with the attribute as the answer (e.g., *Q: What is the birth date of Barack Obama? A: August 4, 1961*). **Fact Recall:** In fact recall tasks, when asked a question about a subject–relation pair (s, r), the model needs to recall the corresponding fact and output the attribute a [6, 7].

2.2 Preliminary experiments

Experimental Setup. To ensure a clean analysis and avoid interference from pre-existing factual knowledge in LMs, we constructed synthetic BIO and QA datasets following the procedure of Zhu et al. [2], both covering the same 10,000 unique individuals. For each individual, we generated one biographical entry consisting of six sentences, each corresponding to one attribute, and the entry maintained a consistent order: birth date, birth city, attended university, job, employer, and blood type. For the QA dataset, 5,000 individuals were used for model training, forming the QA *in-distribution* set, while the remaining 5,000 were held out for evaluation, forming the QA *out-of-distribution* set. See Appendix 6.1 for further details. We analyze the decoder-only LM based on the Transformer architecture: Llama-3.2 (28 layers, 3B parameters) with a 128K vocabulary. Each Transformer layer consists of a multi-head self-attention (MHA) module and an MLP block, using the standard sequential design. All experimental results are averaged over 3 random seeds.

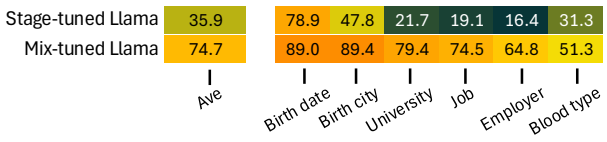


Figure 1: Exact match accuracy of fine-tuned models on QA out-of-distribution set.

Experimental Findings. We investigate two fine-tuning strategies differing in training order: (1) *Two-Stage Training*. We first fine-tune the pre-trained LM on the BIO dataset alone, allowing it to memorize the biographical statements. We then fine-tune this model on the QA in-distribution dataset to adapt it to the question–answer format. The result is a *Stage-tuned* model. (2) *Mixed Training*. We fine-tune the pre-trained LM on a mixed dataset containing all BIO statements combined with the QA in-distribution examples. BIO and QA examples are randomly shuffled into training batches. The resulting model is referred to as the *Mix-tuned* model. See Appendix 6.2 for detailed fine-tuning procedures. After training, we evaluate the Stage-tuned and Mix-tuned models on the QA out-of-distribution data. We report exact match accuracy as the evaluation metric, treating the input prefix as the prompt and evaluating the predicted attribute span. Figure 1 summarizes the performance. The Mix-tuned model outperforms the Stage-tuned model by over 38 points in accuracy, demonstrating superior generalization. Figure 5 in Appendix 6.2 further confirms that both models successfully store BIO statements and adapt to QA in-distribution, indicating that the performance gap not relates to partial forgetting but lies in extracting underlying knowledge. These results indicate that two-stage training encourages rote memorization, while mixed tuning teaches knowledge. In the following section, we investigate the in-

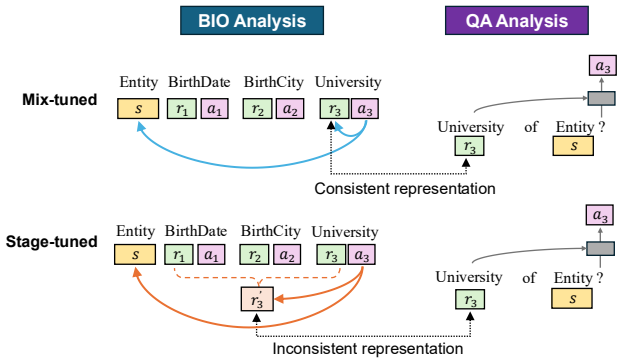


Figure 2: Representation consistency hypothesis.

ternal differences between fine-tuned models that account for this performance gap.

3 Explaining the divergence in fact recall: the role of representations

To answer our question—*why does two-stage training appear to encourage memorization while mixed training appears to teach knowledge?*— we first exclude catastrophic forgetting as the primary cause (§ 2.2) and instead analyze the model's internal representations and extraction mechanisms. As shown in Figure 1, while both models accurately recall initial attributes (e.g., "birth date" >78.9%), the Stage-tuned model exhibits a sharp performance decay as attributes become more distal. In contrast, the Mix-tuned model maintains robust recall across the entire sequence. This divergence suggests a systemic retrieval bottleneck in the Stage-tuned model: as the prompt format shifts from BIO to QA, the model struggles to retrieve information located deeper in the original sequence. We hypothesize that this failure stems from representation inconsistency (§ 3.1)—a state where the model fails to anchor subject-relation (s, r) pairs in a format-independent latent space. While two-stage training allows these encodings to diverge, mixed training explicitly encourages consistency, a hypothesis we empirically validate in § 3.2.

3.1 The representation consistency hypothesis

As illustrated in Figure 2, two-stage training lacks the constraints necessary to ensure cross-format alignment. In the initial stage, the model encodes attribute a_3 conditioned on a long BIO prefix (s, r'_3). However, the subsequent QA fine-tuning utilizes a significantly shorter and structurally distinct context (r_3, s). This temporal and structural separation allows the model to develop divergent encodings for the same relation from different input formats. Consequently, the model fails to generalize the stored knowledge to out-of-distribution QAs, with the failure being particularly pronounced for distal attributes whose BIO encoding was conditioned on more complex prefixes.

Conversely, mixed training acts as a multi-task learning objective, simultaneously optimizing for both (s, r'_3) and (r_3, s). This joint pressure forces the model to maximize parameter utility by converging on a format-invariant encoding mechanism. By marginalizing format-specific noise, the model develops consistent relation representations that reside in a shared latent space, thereby facilitating robust fact recall across disparate prompt structures.

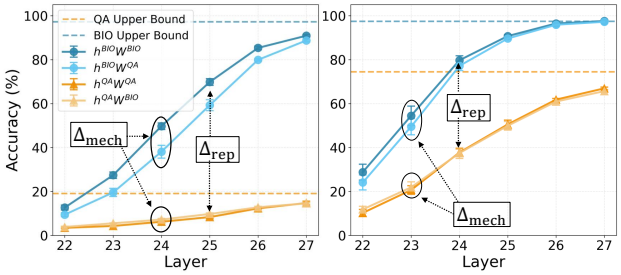
3.2 Probing representation consistency

Let the hidden state h at layer l represent the core semantic information of the subject and relation. We define the model's internal extraction function as $o = g(h)$, which maps the state h to the first token of the target attribute. To validate our hypothesis, we investigate two potential drivers of representation consistency: (1) Feature Alignment, where (s, r) is encoded in a format-invariant manner ($h^{\text{BIO}} \approx h^{\text{QA}}$), and (2) Mechanism Alignment, where the model employs a unified extraction logic ($g^{\text{BIO}} \approx g^{\text{QA}}$).

Directly comparing internal states or non-linear parameters is challenging due to task-irrelevant noise and complexity. However, since factual attributes are linearly decodable in top layers [12], we formalize the extraction process $g(\cdot)$ as a linear projection $W \in \mathbb{R}^{d \times |V|}$, where d is the hidden size and $|V|$ is the vocabulary size. By training separate probes, W^{BIO} and W^{QA}, we can disentangle feature stability from mechanism consistency. To quantify these effects, we define the following metrics:

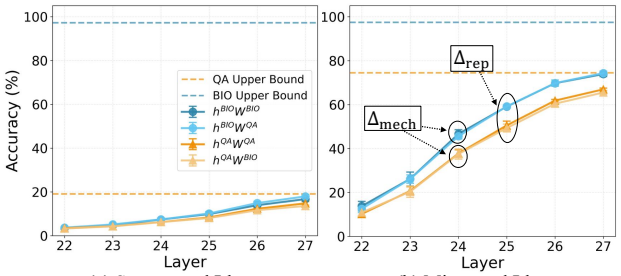
1. **Feature Consistency** (Δ_{rep}): Measures if features from different formats are decodable by a fixed probe (e.g., W^{QA}):
   $\Delta_{rep} = |\text{Acc}(h^{\text{QA}}W^{\text{QA}}) - \text{Acc}(h^{\text{BIO}}W^{\text{QA}})|$.
2. **Mechanism Consistency** (Δ_{mech}): Measures if distinct extraction logics yield consistent results on a single representation (e.g., h^{QA}):
   $\Delta_{mech} = |\text{Acc}(h^{\text{QA}}W^{\text{BIO}}) - \text{Acc}(h^{\text{QA}}W^{\text{QA}})|$.

Experimental Setup. To disentangle the effects of sequence length (r_3 vs. r'_3) and structural formatting (e.g.,



(a) Stage-tuned Llama (b) Mix-tuned Llama

Figure 3: Accuracy for first token of attribute "job", when BIO = BIO_{long}.



(a) Stage-tuned Llama (b) Mix-tuned Llama

Figure 4: Accuracy for first token of attribute "job", when BIO = BIO_{short}.

(s, r_3) vs. (r_3, s)), we evaluate three distinct formats: the full BIO sequence (BIO_{long}: (s, r_3')), a truncated BIO-style fact (BIO_{short}: (s, r_3)), and the QA format (r_3, s). To prevent h from containing already-extracted attribute information, we define the hidden state as the sum of the residual stream and the self-attention output, bypassing the MLP layers which are conventionally associated with knowledge retrieval [4].

Results. As shown in Figure 3, the Stage-tuned model exhibits a significantly larger Δ_{rep} than the Mix-tuned model when using BIO_{long}, indicating a fundamental feature inconsistency. While both models maintain relatively small Δ_{mech} values, suggesting largely aligned extraction mechanisms, the Stage-tuned model shows a higher Δ_{mech} ($\approx 20\%$) for h^{BIO} at Layer 24, reflecting a greater divergence in its retrieval logic. Interestingly, when BIO = BIO_{short}, Figure 4 reveals that both models show nearly zero Δ_{mech}, and a much smaller $\Delta_{rep}(< 10\%)$. These results suggests that the observed inconsistency is not primarily driven by structural formatting (e.g., (s, r_3) vs. (r_3, s)), but rather by the sequence length or the distal position of attributes in long sequences. In summary, our analysis clarifies a critical distinction: while Stage-tuning merely achieves passive storage where knowledge is "locked" within specific training formats, Mixed-tuning fosters an active, format-invariant representation space. By explicitly aligning long-sequence encodings with short-query structures, Mixed-tuning ensures that factual information remains accessible and generalizable, whereas Stage-tuning suffers from a latent-space mismatch that renders stored facts invisible to out-of-distribution queries.

4 Related work

Mechanistic interpretability aims to explain model behavior by uncovering its internal mechanisms [13]. Prior work has investigated how MLP neurons [13] and attention heads [14] contribute to particular linguistic or reasoning tasks. Recent work has analyzed how different optimization strategies affect a model's internal capabilities [15]. Our work contributes to this by providing an internal analysis of why mixed training, as an optimization strategy, effectively facilitates knowledge retrieval.

Understanding factual knowledge in language models falls into two main directions: understanding how models *store* facts [7, 16], and how they *recall* facts [2, 5, 6]. For recalling, [5] show that *extract heads* and MLP blocks work together to recall facts. [6] show that generalization in fact recall correlates with how well facts are stored. Our work extends this line of research by providing a mechanistic analysis of how mixed training promotes consistent knowledge representation for retrieval.

5 Conclusion

We investigate the mechanistic divergence between two-stage and mixed training strategies for fact recall. Our analysis reveals that two-stage training fails to generalize due to representation inconsistency, where the model develops divergent encodings across fact-storing and recalling formats. Conversely, mixed training facilitates generalization by enforcing a format-invariant latent space that anchors factual representations regardless of input structure. By disentangling feature encoding from extraction logic, we provide a new perspective on how optimization strategies shape the internal knowledge mechanisms of LMs. Our model-agnostic framework provides a foundation for future research into broader cross-task alignment.

Acknowledgment

The project was supported by RIKEN Incentive Research Project and by the Japan Science and Technology Agency under Grant No. JST BOOST JPMJBY24F9. We thank Yuta Hitomi and Ryoma Ishigaki for insightful discussions.

References

[1] DeepSeek-AI. Deepseek-v3 technical report. **arXiv preprint arXiv:2412.19437**, 2024.

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

[3] Yuheng Chen, Pengfei Cao, Yubo Chen, et al. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 38, pp. 17817–17825, 2024.

[4] Damai Dai, Li Dong, Yaru Hao, et al. Knowledge neurons in pretrained transformers. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics**, pp. 8493–8502. Association for Computational Linguistics, 2022.

[5] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 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. Association for Computational Linguistics, 2023.

[6] Gaurav Rohit Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. Understanding finetuning for factual knowledge extraction. In **International Conference on Machine Learning**, pp. 15540–15558. PMLR, 2024.

[7] Benjamin Heinzerling and Kentaro Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 1772–1791. Association for Computational Linguistics, 2021.

[8] Xiyu Liu, Zhengxiao Liu, Naibin Gu, et al. Relation also knows: Rethinking the recall and editing of factual associations in auto-regressive transformer language models. In **Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence**, 2025.

[9] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In **Advances in Neural Information Processing Systems**, pp. 17359–17372, 2022.

[10] Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. What does the knowledge neuron thesis have to do with knowledge? In **International Conference on Learning Representations**, 2024.

[11] Olga Golovneva, Zeyuan Allen-Zhu, Jason E Weston, and Sainbayar Sukhbaatar. Reverse training to nurse the reversal curse. In **Conference on Language Modeling**, 2024.

[12] Evan Hernandez, Arnab Sen Sharma, Tal Haklay, et al. Linearity of relation decoding in transformer language models. In **International Conference on Learning Representations**, 2024.

[13] Wes Gurnee, Theo Horsley, Zifan Carl Guo, et al. Universal neurons in GPT2 language models. **Transactions on Machine Learning Research**, 2024.

[14] Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, et al. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In **International Conference on Learning Representations**, 2024.

[15] Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, et al. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In **International Conference on Learning Representations**, 2024.

[16] Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric attributes in language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, pp. 175–195. Association for Computational Linguistics, 2024.

6 Appendix

6.1 Details on data preparation

6.1.1 BIO dataset

This research followed the general setup of [2] with some modifications to generate synthetic datasets. Specifically, we generated profiles for $N = 10,000$ individuals. For each person, we independently and randomly sampled their first and last name, gender, birth date, birth city, attended university, job, employer, and blood type from uniform distributions. Example biographies:

- *Meghan Charles's birthday is remembered on November 10, 2092. He took birth in Amyville, GA. He took advantage of the diverse programs offered at Cedar Crest College. He was involved in work as a chartered public finance accountant. He aligned his professional ambitions with Henson, Ellis and Sexton. He has a blood type of O-.*

6.1.2 QA dataset

For each individual, we generated six fixed questions corresponding to the six attributes. The question served as the prompt, and the model generated the attribute as the answer. Accuracy was again computed by exact match. Examples:

- *What is the birth date of Meghan Charles? November 10, 2092.*
- *What is the birth city of Meghan Charles? Amyville, GA.*
- *Which university did Meghan Charles study? Cedar Crest College.*
- *What is the job of Meghan Charles ? chartered public finance accountant.*
- *Which company did Meghan Charles work for? Henson, Ellis and Sexton.*
- *What is the blood type of Meghan Charles ? O-.*

6.2 Details on BIO/QA fine-tuning

We fine-tuned all models using the AdamW optimizer with an initial learning rate of 0.0001 and cosine learning rate decay on NVIDIA A100 40GB. All models were trained with next-token prediction. For BIO fine-tuning and mixed training, we used a linear warm-up of 1,600 steps and a batch size of 32. For QA fine-tuning, we used no warm-up and a batch size of 256. Table 1 summarizes the total number of training updates for each model setting.

Table 1: Training updates for different model and training strategies.

Model	Training Type	Updates	Training Hours (1 run)
Llama	BIO	6,820	10.37
	QA	400	4.36
	Mix	10,571	22.16

Figure 5 shows the performance of fine-tuned models. Both the QA-tuned and Mix-tuned models perform well on data seen during training. Specifically, they retain biographical facts (BIO Accuracy $\geq$ 97.2%) and adapt well to the question-answer format (QA in_dist accuracy $\geq$ 98.4%).

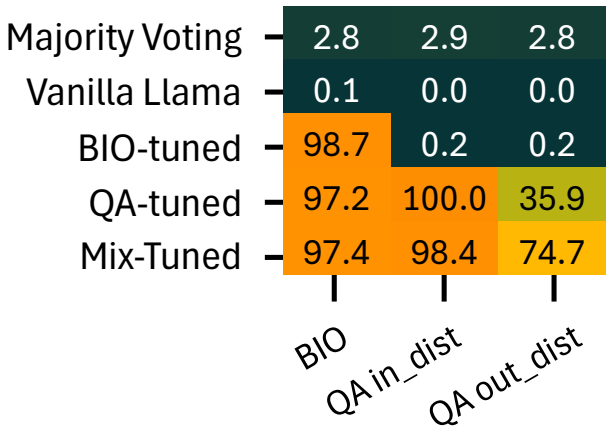


Figure 5: **Performance of fine-tuned Llama and Pythia on BIO, QA in-distribution, and QA out-of-distribution sets.** High accuracy on BIO and QA in-distribution sets indicates successful retention of biographical facts and adaptation to the question–answer format, respectively.