How Training Data Shapes the Use of Parametric and In-Context Knowledge in Language Models

Minsung Kim[1], **Dong-Kyum Kim**[2], **Jea Kwon**[2], **Nakyeong Yang**[1]
Kyomin Jung[1,†], **Meeyoung Cha**[2,†]
[1]Seoul National University, [2]Max Planck Institute for Security and Privacy
{kms0805, kjung}@snu.ac.kr mia.cha@mpi-sp.org

Abstract

Large language models leverage both parametric knowledge acquired during pretraining and in-context knowledge provided at inference time. Crucially, when these sources conflict, models arbitrate based on their internal confidence, preferring parametric knowledge for high-confidence facts while deferring to context for less familiar ones. However, the training conditions that give rise to these fundamental behaviors remain unclear. Here we conduct controlled experiments using synthetic corpora to identify the specific data properties that shape knowledge utilization. Our results reveal a counterintuitive finding: the robust, balanced use of both knowledge sources is an emergent property that requires the co-occurrence of three factors typically considered detrimental, including (i) intra-document repetition, (ii) a moderate degree of intra-document inconsistency, and (iii) a skewed knowledge distribution. We further show that these dynamics arise in real-world language model pretraining and analyze how post-training procedures reshape arbitration strategies. Together, our findings provide empirical guidance for designing training data that supports the reliable integration of parametric and in-context knowledge in language models.[1]

1 Introduction

Large language models (LLMs) (Touvron et al., 2023; Brown et al., 2020; Biderman et al., 2023) encode vast amounts of world knowledge within their parameters during pretraining (Roberts et al., 2020; Petroni et al., 2019; Geva et al., 2020). However, reliance on this parametric knowledge is fundamentally limited: it becomes outdated as the world changes and lacks coverage of rare or domain-specific information. To address these limitations, retrieval-augmented generation (RAG) (Lewis et al., 2021; Ram et al., 2023a; Shi et al., 2023) provides external documents as in-context knowledge at inference time. Through this paradigm, models can effectively utilize information that lies outside their parameters, such as contemporary facts or domain-specific details.

LLMs acquire the ability to leverage both knowledge sources through standard next-token prediction on web corpora (Radford et al., 2018), without requiring explicit fine-tuning for retrieval-augmented generation (Ram et al., 2023b; Mallen et al., 2023; Shi et al., 2023). When these two sources conflict, models do not blindly follow in-context knowledge, which may itself be noisy or incorrectly retrieved. Instead, they exhibit confidence-dependent arbitration: they prefer their internal parametric knowledge for high-confidence facts (i.e., high-probability, low-entropy predictions) while deferring to in-context knowledge for less familiar information (Wu et al., 2024; Yu et al., 2023). Despite the widespread deployment of these systems, the specific training data properties that give rise to these fundamental behaviors remain poorly understood.

In this work, we present the first systematic identification of the training characteristics that enable models to robustly integrate parametric and in-context knowledge. We do so by training models on synthetic corpora (Allen-Zhu and Li, 2024a,b; Zucchet et al., 2025) with carefully varied properties to examine how these behaviors emerge and evolve. Specifically, we periodically evaluate three fundamental knowledge-related capabilities throughout the training process: parametric knowledge utilization, in-context knowledge utilization, and knowledge conflict resolution (Figure 1).

Our experiments reveal a counterintuitive finding: the robust use of both knowledge sources emerges only when three factors typically regarded as detrimental co-occur. First, **intra-document rep-**

†Corresponding authors.

[1]Code available at https://github.com/kms0805/how-training-data-shapes-pk-ick.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23242–23257

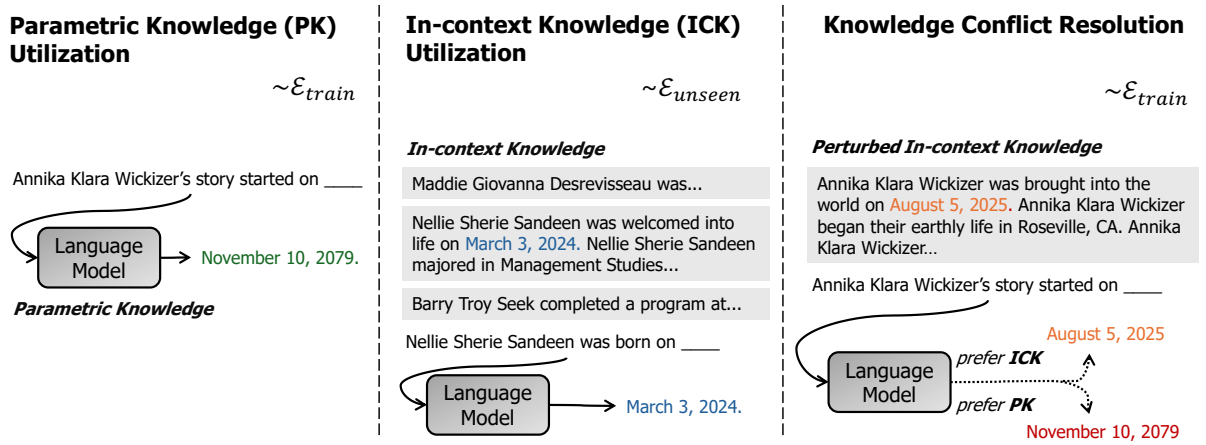


Figure 1: Evaluation framework for knowledge utilization and conflict resolution. **(Left)** Parametric knowledge utilization, where the model recalls knowledge encoded in its parameters about entities seen during training. **(Middle)** In-context knowledge utilization, where the model extracts and uses knowledge provided only in the prompt on unseen entities. **(Right)** Knowledge conflict resolution, where the model is queried about trained entities while the context provides conflicting information, and responses reveal the preference between parametric knowledge and in-context knowledge. These behaviors collectively track the co-emergence of dual knowledge capabilities.

etition creates the necessary training signal for the co-emergence of parametric and in-context knowledge capabilities (Section 3.1). Second, a moderate level of **intra-document inconsistency** prevents an over-reliance on in-context knowledge (Section 3.2). Third, a **skewed knowledge distribution** maintains a balanced reliance on both sources by ensuring that rare facts continue to require in-context knowledge, thereby preventing over-reliance on parametric knowledge (Section 3.3). These findings offer empirical guidance for designing data curation for large language models: aggressive preprocessing, such as extensive deduplication and data balancing, may inadvertently impair a model's ability to integrate diverse knowledge sources and resolve conflicts between them.

Our contributions are as follows:

- We present the first controlled study examining how specific training data characteristics shape the use of both parametric and in-context knowledge in language models.
- We identify that robust use of both knowledge sources emerges when three key factors co-occur: intra-document repetition, intra-document inconsistency, and skewed knowledge distributions.
- We demonstrate that these training dynamics generalize to real-world language models (Section 4.1) and establish how post-training procedures, such as instruction tuning, further reshape a model's internal arbitration strategies (Section 4.2).

2 Dataset and Setup

To study how the three knowledge-related capabilities of language models emerge during training, we design a controlled experimental framework that measures each capability and enables systematic manipulation of training-data properties.

2.1 Synthetic Biography Dataset

Following established protocols (Allen-Zhu and Li, 2024a; Zucchet et al., 2025), we construct a synthetic biography dataset in three steps (Figure 2). First, we generate synthetic entities, each defined by a profile of four attributes: `birth_date`, `birth_city`, `university`, and `major`. Second, for each attribute, we sample 7 distinct surface templates from a finite pool. Third, we instantiate these templates with the attribute values to create biography paragraphs. For each entity, 6 templates per attribute are used to create six paraphrased biography paragraphs: five are reserved as training paragraphs and one is held out as the evaluation context paragraph. The remaining template for each attribute is held out as a cloze-style test probe to obtain the corresponding attribute value. This separation ensures that training paragraphs, evaluation contexts, and test probes never share identical surface forms, encouraging the model to use parametric or in-context knowledge rather than simple string memorization. Detailed specifications are provided in Appendix A.

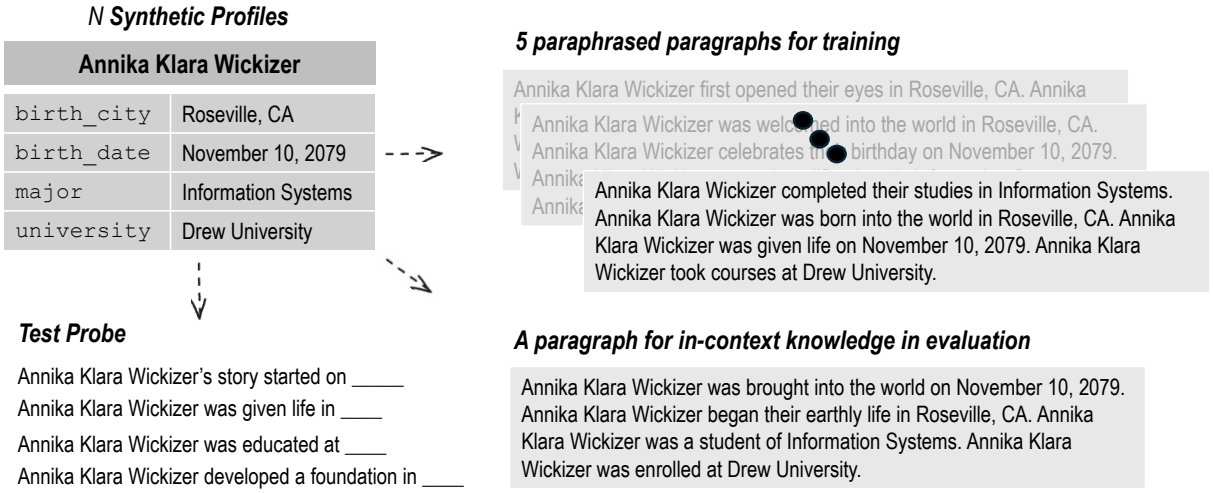


Figure 2: An example from the synthetic biography dataset. Each profile consists of four attributes (`birth_date`, `birth_city`, `university`, `major`), with paragraphs for training, a paragraph for in-context knowledge in evaluation, and test probes for eliciting the model to generate the attributes of each entity.

2.2 Training Setup

We train an 8-layer decoder-only Transformer (Vaswani et al., 2017) from scratch, adopting hyperparameters from prior work (Zucchet et al., 2025) (see Appendix B for details). We use $|\mathcal{E}_{\text{train}}| = 50\text{k}$ entities for training and hold out a separate set of $|\mathcal{E}_{\text{unseen}}| = 50\text{k}$ entities for evaluating in-context knowledge utilization on unseen entities. The model is trained on a corpus of training paragraphs from $\mathcal{E}_{\text{train}}$ using the next-token prediction objective (Radford et al., 2018).

2.3 Evaluation Protocol

We periodically evaluate models in three capabilities described below (Figure 1), using exact-match accuracy on 200 randomly sampled entities every 100 training steps.

Parametric Knowledge Utilization (PKU). This metric measures the model's ability to recall learned facts without contextual support. Given an entity $e \in \mathcal{E}_{\text{train}}$ and a test probe p_a for the attribute a, the model must generate the correct value v_a solely from its parameters.

$$\text{Acc}_{\text{PKU}} = \mathbb{E}_{e \sim \mathcal{E}_{\text{train}}} \left[\frac{1}{|A_e|} \sum_{a \in A_e} \mathbf{1}\{M(p_a) = v_a\} \right]$$

where A_e is the set of attributes for entity e, $M(\cdot)$ denotes the model output, and $\mathbf{1}\{\cdot\}$ is the indicator function.

In-Context Knowledge Utilization (ICKU). This metric evaluates the model's ability to extract and utilize knowledge provided in context for entities never seen during training. For $e \in \mathcal{E}_{\text{unseen}}$, we construct a context C containing e's held-out evaluation paragraph along with paragraphs from two other unseen entities as distractors:

$$\text{Acc}_{\text{ICKU}} = \mathbb{E}_{e \sim \mathcal{E}_{\text{unseen}}} \left[\frac{1}{|A_e|} \sum_{a \in A_e} \mathbf{1}\{M(C, p_a) = v_a\} \right]$$

Knowledge Conflict Resolution. This metric reveals the model's preference when parametric and in-context knowledge conflict. For $e \in \mathcal{E}_{\text{train}}$, we construct a perturbed context C'_e by replacing attribute values with randomly sampled alternatives, then measure how often the model follows each source:

$$\text{Pref}_{\text{PK}} = \mathbb{E}_{e \sim \mathcal{E}_{\text{train}}} \left[\frac{1}{|A_e|} \sum_{a \in A_e} \mathbf{1}\{M(C'_e, p_a) = v_a\} \right]$$

$$\text{Pref}_{\text{ICK}} = \mathbb{E}_{e \sim \mathcal{E}_{\text{train}}} \left[\frac{1}{|A_e|} \sum_{a \in A_e} \mathbf{1}\{M(C'_e, p_a) = v'_a\} \right]$$

where v_a denotes the original (parametric) value and v'_a denotes the perturbed (in-context) value. Higher Pref_{PK} indicates stronger reliance on parametric knowledge; higher Pref_{ICK} indicates stronger reliance on in-context knowledge. Note that $\text{Pref}_{\text{PK}} + \text{Pref}_{\text{ICK}}$ need not sum to 1, since the model may also produce outputs matching neither v_a nor v'_a.

3 Experiments

Building on the framework introduced in Section 2, we reverse-engineer how language models acquire these three capabilities (parametric knowledge utilization, in-context knowledge utilization, and

confidence-dependent arbitration under knowledge conflict) by systematically manipulating training-data properties and identifying which conditions support the emergence of each capability.

3.1 Effect of Intra-Document Repetition

Motivation and hypothesis. We first examine which factors enable models to use both parametric and in-context knowledge. We hypothesize that intra-document repetition, a common property of natural text in which some information is restated within the same document (Figure 3), plays a critical role. During the prediction of the next-token, the first mention of a fact requires parametric recall (Geva et al., 2023; Meng et al., 2022), while later mentions allow the model to take advantage of an earlier context. We hypothesize that this learning signal naturally enables both parametric recall and the use of in-context knowledge.

Design. To test this hypothesis, we construct two corpus variants that differ in whether attributes repeat within documents:

- SINGLE: Each document contains one paragraph per entity, so attributes appear only once.
- REPEATED: Each document contains two paraphrased paragraphs per entity. The first mention necessarily relies on parametric knowledge, while the second mention provides an opportunity for the model to use either parametric knowledge or in-context knowledge. To avoid trivial copying based solely on previously mentioned attribute types regardless of the subject, we mix multiple entities within each document. Specifically, we sample two paraphrased paragraphs for each of three distinct entities and shuffle all six paragraphs to form a single training document.

In both variants, paragraphs are sampled from the five training paragraphs reserved for each entity in Section 2.1.

Finding 1: Repetition yields co-emergence, with in-context knowledge utilization emerging first. Figure 4 shows that models trained on SINGLE develop only parametric recall, whereas models trained on REPEATED acquire both capabilities. Moreover, the ability to use in-context knowledge emerges earlier than parametric recall. One possible explanation for this ordering is a structural asymmetry: in-context knowledge can be used via general copying mechanisms (Olsson et al., 2022), whereas parametric recall requires jointly learning entity-specific knowledge and a recall mechanism, a combination that develops more gradually, as observed in prior work (Zucchet et al., 2025).

Finding 2: Clean REPEATED corpus induces over-reliance on context. Models trained on SINGLE cannot utilize in-context knowledge and therefore trivially prefer parametric knowledge under knowledge conflict. In contrast, models trained on REPEATED, despite possessing both capabilities, consistently prefer in-context knowledge under knowledge conflict (Figure 4), even when their parametric knowledge is highly confident. This is evidenced by significantly lower entropy and higher target probability for training entities (see Appendix D). Such over-reliance on context deviates from the behavior of real-world language models, which tend to prefer parametric knowledge for high-confidence facts (Yu et al., 2023; Wu et al., 2024).

3.2 Effect of Intra-Document Inconsistency

Motivation and hypothesis. The previous section showed that models trained on clean REPEATED data over-rely on in-context knowledge, even when their parametric knowledge is highly confident. This observation raises the question of which properties of natural web corpora discourage such unconditional reliance on in-context knowledge.

We hypothesize that a moderate degree of intra-document inconsistency plays this role. Real-world corpora inevitably contain noise (e.g., typos, imperfect statements, or synonymous paraphrases), making in-context evidence an imperfect signal. When contextual information is occasionally incorrect, the model may learn that parametric knowledge is more reliable for high-confidence facts. To test this hypothesis, we introduce controlled intra-document inconsistency into the training corpus.

Design. Starting from REPEATED, we inject inconsistency by perturbing the values of entity attributes in the leading paragraph of each document with probability $p \in \{1\%, 5\%, 10\%\}$, replacing them with randomly sampled alternative values, while leaving the later paragraph unchanged (Figure 10).

Finding 1: Intra-document inconsistency induces a preference transition. Figure 5 shows a consistent two-stage pattern. Early in training, the

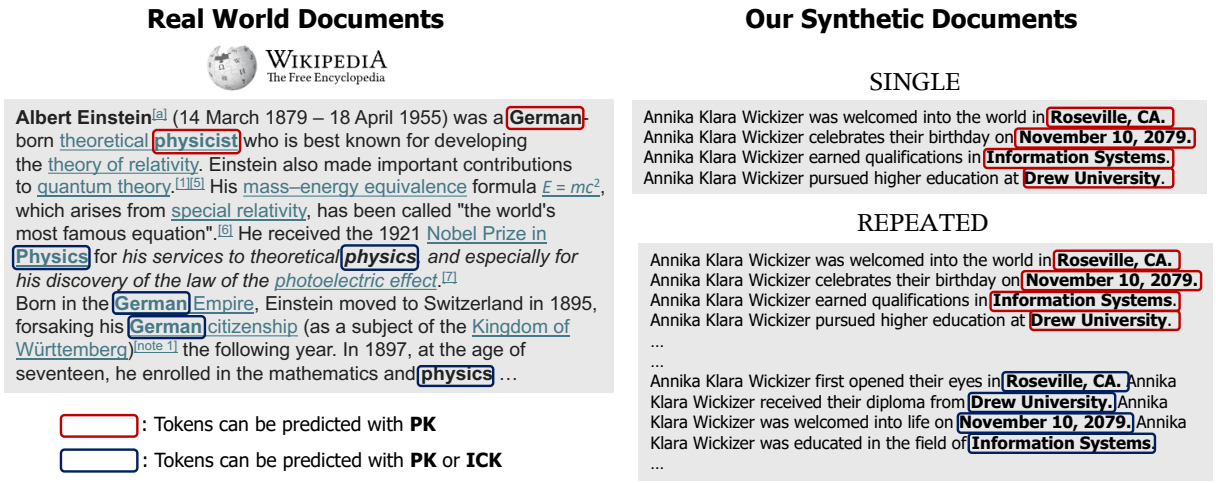


Figure 3: Intra-document repetition in the training corpus. (**Left**) Real-world Wikipedia text. (**Right**) Our synthetic corpus variants: SINGLE contains one paragraph per entity, while REPEATED contains two paraphrased paragraphs per entity, enabling in-context knowledge utilization on later mentions.

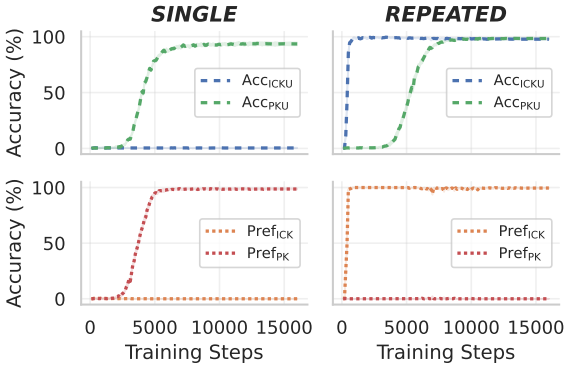


Figure 4: Evaluation results on SINGLE versus REPEATED. SINGLE develops only parametric knowledge utilization and always prefers parametric knowledge under conflict. REPEATED yields both capabilities, with in-context knowledge utilization emerging first, but consistently prefers in-context knowledge under conflict.

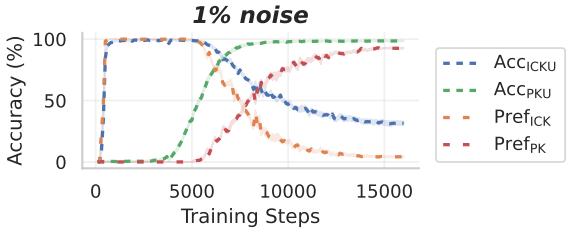


Figure 5: Effects of intra-document inconsistency. A 1% inconsistency rate shifts model preference from in-context to parametric knowledge as training progresses.

ability to use in-context knowledge emerges first, and the model prefers in-context knowledge under conflict. As parametric recall stabilizes, however, the model's preference gradually shifts toward parametric knowledge. Remarkably, even 1% inconsistency is sufficient to induce this transition. This behavior suggests that inconsistency imposes an effective ceiling on the reliability of context-based copying; once parametric accuracy exceeds this ceiling, the model increasingly favors parametric knowledge under conflict.

Finding 2: Inconsistency degrades in-context knowledge utilization. As the model increasingly relies on parametric knowledge, its ability to use in-context knowledge degrades at convergence. Notably, this degradation occurs even though evaluation is conducted on entities never seen during training: the model increasingly fails to use the provided context when answering questions about entirely unseen entities. One plausible explanation is the gradual forgetting of in-context circuits due to reduced usage (Olsson et al., 2022). As parametric knowledge becomes more advantageous for frequently observed entities during training, the circuits supporting context use receive diminishing learning signal and gradually deteriorate. Attention analysis in Appendix E supports this interpretation: when evaluating on unseen entities, attention initially concentrates on context tokens but progressively shifts toward subject name tokens, suggesting that the circuits responsible for in-context knowledge retrieval are used less during training and gradually forgotten.

3.3 Effect of Skewed Knowledge Distribution

Motivation and hypothesis. To prevent the degradation of in-context knowledge utilization

Noise	Uniform	Zipfian
1%	31.5%	84.0% (+52.5%)
5%	16.8%	63.9% (+47.1%)
10%	14.1%	57.4% (+43.3%)

Table 1: In-context knowledge utilization accuracy at the end of training. **Uniform** corresponds to the standard REPEATED corpus with uniform entity frequency, while **Zipfian** uses a skewed ($\alpha = 1$) entity frequency distribution. Zipfian sampling substantially mitigates degradation relative to uniform sampling under matched inconsistency levels.

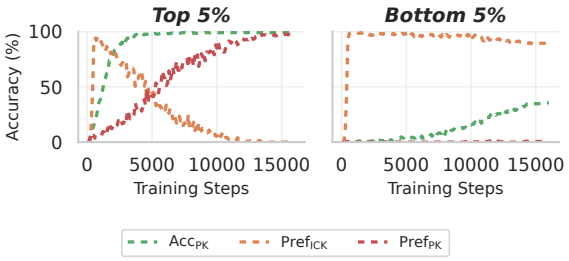


Figure 6: Arbitration behavior stratified by entity frequency. High-frequency entities transition to parametric preference, whereas low-frequency entities maintain in-context preference throughout training.

observed in the previous section, we hypothesize that the model must be continuously exposed to predictions that cannot be resolved by parametric knowledge alone during training. In natural web corpora, a vast amount of information exists where some knowledge appears very frequently while most knowledge appears only occasionally (long-tailed knowledge) (Mallen et al., 2023). We hypothesize that this skewed distribution of knowledge is key to maintaining balanced use of both in-context and parametric knowledge.

Design. We construct REPEATED corpora where entity occurrences follow a Zipfian distribution with parameter $\alpha = 1$,[2] and inject $p = 1\%$ inconsistency noise as in the previous section.

Finding 1: Long-tailed knowledge preserves in-context knowledge utilization capabilities. We hypothesized that long-tailed knowledge, for which sufficient parametric knowledge has not accumulated, would require continuous use of in-context knowledge, thereby preventing the degradation of in-context circuits observed in Section 3.2. Indeed, Table 1 shows substantially less degradation of in-

[2]The Zipfian distribution is defined as $P(r) = r^{-\alpha} / \sum_{k=1}^{N} k^{-\alpha}$, where r denotes the frequency rank.

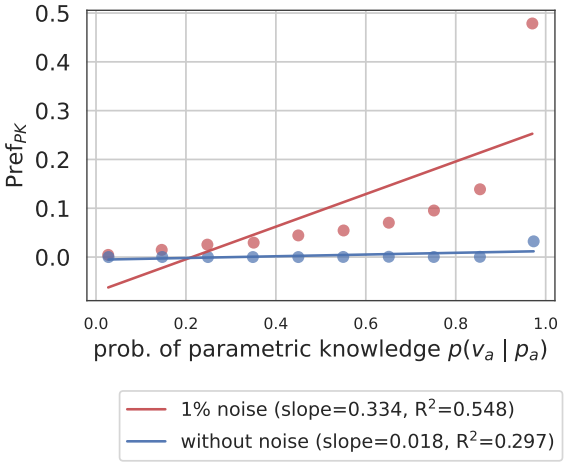


Figure 7: Parametric confidence (grouped into 10 bins) against parametric preference under conflict. (**Red**) With 1% inconsistency noise, higher confidence yields stronger parametric preference. (**Blue**) Without inconsistency noise, models show over-reliance on in-context knowledge across all confidence levels.

context knowledge utilization under Zipfian distribution across all noise levels. However, when inconsistency noise is too high ($p > 1\%$), in-context knowledge utilization appears to degrade too severely; even with Zipfian sampling, the recovery is only partial and final accuracy remains relatively low.

Finding 2: Frequency-dependent arbitration emerges. Unlike the uniform setting, where the model's confidence in parametric knowledge saturates uniformly across all trained entities, Zipfian sampling yields varying confidence in parametric knowledge that correlates with entity frequency (Appendix D, Figure 11), as observed in real-world language models (Mallen et al., 2023). This variation enables us to examine how the model's arbitration behavior under knowledge conflicts varies between entities that appear frequently versus infrequently in the training corpus. As shown in Figure 6, high-frequency entities (top 5% of all entities) transition toward parametric preference as training progresses, as observed in Section 3.2, while low-frequency entities (bottom 5% of all entities) maintain in-context preference throughout. Notably, for rare entities we observe that $\mathrm{Acc}_{\mathrm{PKU}}$ exceeds $\mathrm{Pref}_{\mathrm{PK}}$: the model can sometimes answer correctly via parametric recall, yet it still prefers contextual evidence under explicit conflict.

Finding 3: Skewed knowledge distribution alone is insufficient. So far, we have examined train-

ing on data with skewed knowledge distribution in the presence of inconsistency noise. However, does a long-tailed distribution alone, without inconsistency, produce confidence-calibrated arbitration? Figure 7 shows that the answer is no. We measure parametric confidence (grouped into 10 bins) against parametric preference under conflict. We first group the probabilities of predictions on parametric knowledge probes into ten equidistant bins and plot the average $\mathrm{Pref}_{\mathrm{PK}}$ for instances in each bin. Without inconsistency noise, models show a low parametric knowledge preference overall. Only the combination of skewed knowledge distribution and modest inconsistency yields the desired alignment between confidence and preference, where higher confidence leads to stronger preference for parametric knowledge.

3.4 Summary

Our experiments show that robust use of both knowledge sources emerges when the following three properties of the training corpus co-occur. First, intra-document repetition creates a training signal for both context use and parametric recall, enabling their co-emergence, with the former emerging earlier and the latter following later.

Second, intra-document inconsistency teaches the model not to blindly copy contextual information. Once parametric knowledge has been sufficiently learned and the model is confident in it, the model relies on its parametric prediction rather than defer to context. However, this shift introduces a new problem: the ability to use in-context knowledge degrades as parametric knowledge becomes sufficient for most predictions.

Third, a skewed knowledge distribution resolves this tension through the complementary roles of rare and frequent entities. Rare entities continue to require in-context knowledge, thereby preserving the model's ability to use context throughout training. For frequent entities, whose parametric knowledge is well-learned, the model instead becomes robust to noisy or misleading in-context evidence, producing confidence-dependent arbitration.

When all three properties co-occur, as they naturally do in real web corpora, models develop balanced reliance on both knowledge sources and confidence-dependent arbitration in our controlled setting. Additional hyperparameter experiments, including the number of entities in training data, noise levels, and degree of skewness, are provided in Appendix G.

4 Discussion and Implications

4.1 Do These Dynamics Emerge in Real-World Pretraining?

Our experiments provide a mechanistic account of knowledge utilization, revealing training dynamics that extend beyond previously studied end-state behaviors (Wu et al., 2024; Yu et al., 2023). We assess whether similar dynamics arise in real-world pretraining by examining open-source LLMs with publicly available intermediate checkpoints.

We evaluate Pythia (Biderman et al., 2023) on parametric knowledge utilization, in-context knowledge utilization, and knowledge conflict resolution at intermediate checkpoints throughout pretraining (experiment details in Appendix F).[3] As shown in Figure 8 (left), Pythia exhibits training dynamics consistent with our controlled experiments: the ability to use in-context knowledge emerges earlier than parametric recall, and the model initially prefers in-context knowledge under conflict but gradually shifts toward parametric knowledge, while maintaining high $\mathrm{Acc}_{\mathrm{ICKU}}$ for novel entities throughout training. These results show that the dynamics identified in our controlled experiments also arise during real-world pretraining, providing evidence that our findings capture behavior relevant beyond the synthetic setting.

To examine how this transition scales with model size, we analyze the preference gap ($\mathrm{Pref}_{\mathrm{ICK}} - \mathrm{Pref}_{\mathrm{PK}}$) for models ranging from 70M to 6.9B parameters (Figure 8, right). All models exhibit a consistent pattern: an initial dominance of in-context knowledge preference gradually shifts toward parametric knowledge preference over training. Notably, larger models show stronger parametric knowledge preference by the end of training, with the preference gap approaching -1 for the largest models, consistent with prior observations that larger models tend to rely more heavily on their parametric knowledge (Yu et al., 2023). This trend can be attributed to larger models developing parametric knowledge more effectively, leading to higher confidence in their internal knowledge and consequently a stronger preference for it when conflicts arise.

[3] We use Pythia because it provides fine-grained intermediate checkpoints throughout pretraining. Results on another model with publicly available pretraining checkpoints, Olmo (Groeneveld et al., 2024), are provided in Appendix F.

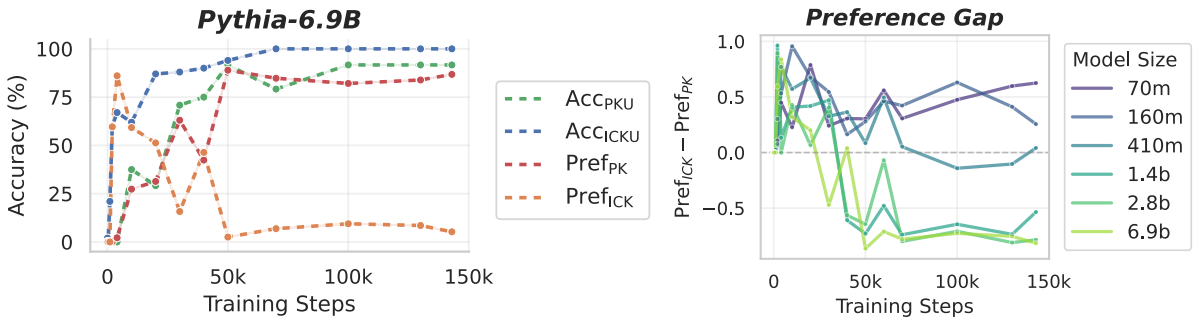


Figure 8: Evaluation results of knowledge utilization and conflict resolution in Pythia. (**Left**) $\mathrm{Acc}_{\mathrm{ICKU}}$, $\mathrm{Acc}_{\mathrm{PKU}}$, $\mathrm{Pref}_{\mathrm{ICK}}$, and $\mathrm{Pref}_{\mathrm{PK}}$ across training steps for Pythia-6.9B. (**Right**) Preference gap ($\mathrm{Pref}_{\mathrm{ICK}} - \mathrm{Pref}_{\mathrm{PK}}$) across different model sizes, showing a consistent pattern of initial increase followed by decline as training progresses.

	Pythia-6.9B		Olmo-7B	
	$\mathrm{Pref}_{\mathrm{PK}}$	$\mathrm{Pref}_{\mathrm{ICK}}$	$\mathrm{Pref}_{\mathrm{PK}}$	$\mathrm{Pref}_{\mathrm{ICK}}$
Before IT	0.8677	0.0525	0.5507	0.3894
After IT	0.1829	0.7771	0.2137	0.7155

Table 2: Conflict resolution before and after instruction tuning (IT) for Pythia-6.9B and Olmo-7B. Both models show a shift from parametric knowledge preference to in-context knowledge preference after IT.

4.2 Can Arbitration Strategies Be Reshaped via Post-Training?

Our findings reveal how the characteristics of the training corpus shape knowledge arbitration strategies during pretraining. A natural question arises: can these strategies be modified after pretraining through post-training procedures such as instruction tuning?

We examine whether instruction tuning affects arbitration behavior in real-world models. Specifically, we evaluate two models at their final pretraining checkpoints (Pythia-6.9B and Olmo-7B) using the same evaluation protocol as in Section 4.1, and compare them against their counterparts post-trained on the Tulu dataset (Wang et al., 2023), an instruction-following dataset.

As shown in Table 2, both base models exhibit a higher preference for parametric knowledge. After post-training, however, both models show a reversal: parametric knowledge preference drops while in-context knowledge preference increases. This suggests that instruction tuning, which typically involves data designed to encourage faithful adherence to context, can significantly alter the arbitration strategies established during pretraining.

Having observed that post-training can modify model behavior, we further investigate whether adjusting inconsistency noise in post-training data can by itself control the model's arbitration strategies as intended, given that this noise is a key factor for in-context knowledge reliance in our findings. We conduct post-training on our synthetic Zipfian corpus using answer-only loss with 1,000 entities for 500 steps, substantially smaller in both entity count and training steps than pretraining.

We examine two scenarios: (1) whether a model pretrained with 1% noise and post-trained on clean data increases its in-context reliance, and (2) whether a model pretrained without noise and post-trained with varying noise levels ($p \in \{1\%, 5\%, 10\%\}$) decreases its in-context reliance.

The results are shown in Figure 9. We plot the confidence-preference alignment by binning entities based on parametric knowledge probability and measuring $\mathrm{Pref}_{\mathrm{PK}}$ for each bin. The results show that adjusting noise levels alone can reshape this alignment. More broadly, noise level is one instance of a more general principle: controlling how much the training data allows the model to rely on context. This calibration of context reliability directly shapes the model's arbitration strategy between parametric and in-context knowledge. In practice, it offers a concrete lever for deployment: raising the context noise in post-training data encourages parametric reliance when retrieval sources are unreliable (e.g., web-based search), while lowering it encourages in-context reliance in domains with highly curated retrieval sources (e.g., legal or medical databases).

5 Related Work

Knowledge Utilization and Conflict Resolution in Language Models. Large language models store factual knowledge in their parameters during pretraining (Roberts et al., 2020; Petroni et al.,

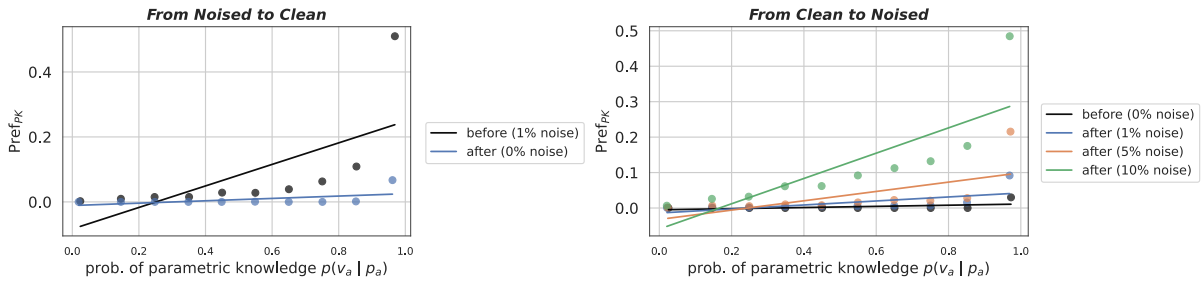


Figure 9: Alignment between parametric knowledge confidence and preference under conflict before and after post-training. (**Left**) Scenario 1: A model pretrained with 1% noise shows declined parametric knowledge preference after post-training with clean data that has no inconsistency noise. (**Right**) Scenario 2: A model pretrained without noise initially shows almost no parametric knowledge preference, but after post-training on data with noise, parametric knowledge preference gradually increases according to the model's confidence, with higher noise levels producing stronger confidence-calibrated parametric knowledge preference.

2019; Geva et al., 2020) and can also leverage in-context knowledge at inference time without explicit fine-tuning (Lewis et al., 2021; Ram et al., 2023a; Shi et al., 2023; Mallen et al., 2023; Ram et al., 2023b). When these sources conflict (Neeman et al., 2022), models exhibit confidence-dependent arbitration, preferring parametric knowledge for well-learned facts while deferring to in-context knowledge for less familiar information (Wu et al., 2024; Yu et al., 2023; Lee et al., 2026). Several methods have been proposed to steer this behavior through attention manipulation or contrastive decoding (Li et al.; Yu et al., 2023; Sun et al., 2025; Jin et al., 2024). However, prior work in this line focuses on analyzing or modifying post-hoc behavior (Kortukov et al., 2024; Xie et al., 2023; Longpre et al., 2021), leaving open the question of how these capabilities emerge during training.

Experiments with Controlled Training Data. Several recent works use controlled training setups to disentangle how data properties shape what models learn. For instance, Chan et al. (2022); Singh et al. (2023) investigate how data properties enable in-context and in-weight learning to co-exist in transformer-based classifiers. However, their work is limited to classification tasks, which may exhibit different dynamics from language models trained with next-token prediction. Other works present controlled studies with synthetic corpora to investigate parametric knowledge acquisition in language models (Allen-Zhu and Li, 2024a,b; Zucchet et al., 2025; Kim et al., 2026), but do not address in-context utilization (Olsson et al., 2022) or conflict resolution (Wu et al., 2024). We extend these directions by examining how training-data characteristics shape both the co-emergence of parametric and in-context knowledge utilization and the development of conflict arbitration strategies.

6 Conclusion

We presented a systematic analysis of how training data characteristics shape the use of parametric and in-context knowledge in language models through controlled experiments. We identified a counter-intuitive finding: robust use of both knowledge sources emerges when three properties commonly regarded as detrimental co-occur, including intra-document repetition, intra-document inconsistency, and skewed knowledge distribution. Experiments on real-world language models show that similar dynamics arise during pretraining, and our post-training experiments demonstrate that knowledge arbitration strategies can be reshaped by adjusting data characteristics. These findings offer practical guidance for designing training data that supports balanced use of parametric and in-context knowledge in language models.

7 Limitations

Our study primarily relies on a synthetic biography dataset and small-scale models. Research based on synthetic data offers the critical advantage of isolating and controlling complex factors while enabling numerous reproducible experiments, which would be infeasible in realistic scenarios. This design choice, however, naturally raises questions about the generalizability of our findings. At the same time, independently controlling properties such as repetition, noise, and frequency distribution in real-world training data is highly challenging, and repeatedly pretraining large-scale language

models to isolate each factor is prohibitively costly, which limits the feasibility of causal analysis in such settings. Accordingly, we adopt a complementary structure: establishing causal relationships through controlled experiments in the synthetic environment, and then showing that similar patterns arise as correlational evidence in real-world models whose pretraining checkpoints are publicly available. Through this approach, we provide a systematic understanding of which training data properties shape a model's knowledge utilization behavior, and we believe our work can serve as a foundation for follow-up studies in more realistic, larger-scale scenarios that are otherwise difficult to control. We hope our findings will be further extended and validated in such settings in future work.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2024-00348233). K. Jung is with ASRI, Seoul National University, Korea. The Institute of Engineering Research at Seoul National University provided computing resources.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024a. Physics of language models: Part 3.1, knowledge storage and extraction. *Preprint*, arXiv:2309.14316.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024b. Physics of language models: Part 3.2, knowledge manipulation. *Preprint*, arXiv:2309.14402.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. Olmo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.

Dong-Kyum Kim, Minsung Kim, Jea Kwon, Nakyeong Yang, and Meeyoung Cha. 2026. Bilinear representation mitigates reversal curse and enables consistent model editing. *Preprint*, arXiv:2509.21993.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. *Preprint*, arXiv:2404.16032.

Dongryeol Lee, Yerin Hwang, Taegwan Kang, Minwoo Lee, Younhyung Chae, and Kyomin Jung. 2026. Judging against the reference: Uncovering knowledge-driven failures in llm-judges on qa evaluation. *Preprint*, arXiv:2601.07506.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In *Forty-second International Conference on Machine Learning*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. 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)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. *arXiv preprint arXiv:2211.05655*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. 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)*, pages 2463–2473.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023a. In-context retrieval-augmented language models. *Preprint*, arXiv:2302.00083.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023b. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

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

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *Preprint*, arXiv:2301.12652.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. 2023. The transient nature of emergent in-context learning in transformers. *Advances in neural information processing systems*, 36:27801–27819.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *Preprint*, arXiv:2410.11414.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *Preprint*, arXiv:2306.04751.

Kevin Wu, Eric Wu, and James Y Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in Neural Information Processing Systems*, 37:33402–33422.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. 2025. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*.

A Biography Dataset Construction

Following prior work (Allen-Zhu and Li, 2024a; Zucchet et al., 2025), we first construct N synthetic person profiles. Each profile contains four attributes: `birth_date`, `birth_city`, `university`, and `major`. Names (first/middle/last) are sampled by randomly composing entries from a public name database.[4] For `birth_date`, we sample a date uniformly between 1900 and 2099. For `birth_city` and `university`, we sample from curated lists of 200 values each, and for `major` from a list of 100 values, all derived from Wikipedia.[5] For each attribute, we sample 7 distinct surface templates from a finite template pool. An example of template for `birth_date` is shown below.

An example of templates for `birth_date`

1. person was born on birth_date.
2. person came into the world on birth_date.
3. person entered this world on birth_date.
4. person was brought into the world on birth_date.
5. person took their first breath on birth_date.
6. person began their life journey on birth_date.
7. person celebrates their birthday on birth_date.
8. person first opened their eyes on birth_date.
9. person was welcomed into life on birth_date.
10. person arrived on birth_date.
11. person's story started on birth_date.
12. person was born to the world on birth_date.
13. person was delivered into the world on birth_date.
14. person was given life on birth_date.
15. person was welcomed into the world on birth_date.
16. person began their journey on Earth on birth_date.
17. person made their debut in the world on birth_date.
18. person became a part of the world on birth_date.
19. person was born into this life on birth_date.
20. person came to life on birth_date.

B Training Details

For our controlled experiments, we use a decoder-only Transformer following the GPT-2 architecture[6]. The model configuration is summarized in Table 3. The training hyperparameters are listed in Table 4. All experiments are implemented using the `Hugging Face TRL` library[7] and conducted on a single NVIDIA A100 GPU. Each training run requires approximately four to six hours.

[4] https://github.com/smashew/NameDatabases/tree/master/NamesDatabases

[5] https://en.wikipedia.org/wiki/

[6] https://huggingface.co/openai-community/gpt2

[7] https://huggingface.co/docs/trl/index

Component	Value
Embedding dimension	512
Layers	8
Attention heads	8
FFN inner dimension	2048
Context length	512

Table 3: Model architecture.

Hyperparameter	Value
Max training steps	16,000
Batch size	128
Learning rate	4×10^{-4}
Weight decay	0.10
LR scheduler	Cosine
Sequence length	512
Numerical precision	bfloat16

Table 4: Training hyperparameters.

C Intra-Document Inconsistency

Figure 10 illustrates a document from the REPEATED corpus in which intra-document inconsistency has been injected.

Annika Klara Wickizer was welcomed into the world in Roseville, CA. Annika Klara Wickizer celebrates their birthday on **August 5, 1999** . Annika Klara Wickizer earned qualifications in Information Systems. Annika Klara Wickizer pursued higher education at Drew University.

Dara Angila Honey was given life on April 6, 1978. Dara Angila Honey focused their academic efforts on Industrial. Dara Angila Honey entered this world in Indianapolis, IN. Dara Angila Honey achieved academic success at Fisk University.

Dara Angila Honey chose Industrial as their field of study. Dara Angila Honey completed a program at Fisk University. Dara Angila Honey was welcomed into life on April 6, 1978. Dara Angila Honey became a part of the world in Indianapolis, IN.

Annika Klara Wickizer first opened their eyes in Roseville, CA. Annika Klara Wickizer received their diploma from Drew University. Annika Klara Wickizer was welcomed into life on **November 10, 2079.** Annika Klara Wickizer was educated in the field of Information Systems.

Roselee Justine Woolem gained academic grounding in Business Analytics. Roselee Justine Woolem first opened their eyes in Phoenix, AZ. Roselee Justine Woolem studied at Hamilton College. Roselee Justine Woolem was brought into the world on August 12, 2083.

Roselee Justine Woolem entered this world on August 12, 2083. Roselee Justine Woolem majored in Business Analytics. Roselee Justine Woolem began their life in Phoenix, AZ. Roselee Justine Woolem developed expertise at Hamilton College.

Figure 10: Example of a document with injected inconsistency noise. The value highlighted in pink was injected as noise with some probability and therefore does not match the later unperturbed value, "November 10, 2079."

D Confidence for Parametric Knowledge

We measure two key metrics at the final token position of each test probe: (1) the probability assigned to the target token, and (2) the entropy of the probability distribution over the vocabulary. These

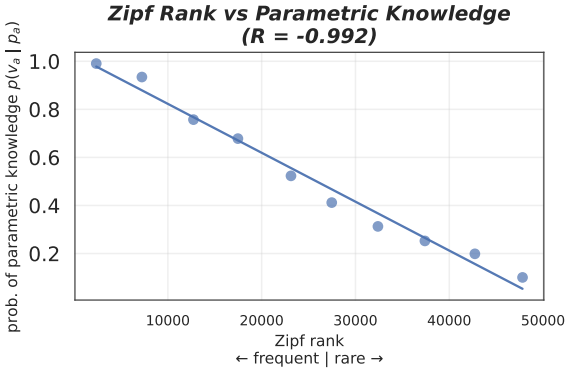


Figure 11: Relationship between entity frequency (Zipf rank) and parametric knowledge strength. The model exhibits a strong negative correlation between Zipf rank and the probability $p(v_a \mid p_a)$ of generating the correct attribute value given only the probe prompt. More frequent entities have stronger parametric knowledge while rare entities show weaker parametric knowledge.

metrics provide complementary perspectives on model confidence: the target probability reflects how strongly the model predicts the correct answer, while the entropy captures the overall uncertainty in the prediction.

	$\mathcal{E}_{\text{train}}$		$\mathcal{E}_{\text{unseen}}$	
	0% noise	1% noise	0% noise	1% noise
Target prob.	0.998	0.997	0.024	0.034
Entropy (nats)	0.011	0.016	0.955	1.236

Table 5: Target token probability and entropy measured at the last token of the test probe by entities

Table 5 presents these measurements for entities in both $\mathcal{E}_{\text{train}}$ (seen during training) and $\mathcal{E}_{\text{unseen}}$ (held-out entities) under two training conditions: without noise and with 1% inconsistency noise. For $\mathcal{E}_{\text{train}}$, the model exhibits extremely high confidence, assigning near-perfect probability to target tokens with very low entropy. This indicates that the model has successfully acquired and can reliably retrieve parametric knowledge for entities it encountered during training. In contrast, for $\mathcal{E}_{\text{unseen}}$, the model shows substantially lower confidence, with lower target probabilities and much higher entropy values.

Parametric knowledge varies with entity frequency. We investigate how the strength of parametric knowledge varies across entities in $\mathcal{E}_{\text{train}}$ as a function of their frequency in the pretraining corpus. Figure 11 shows the relationship between Zipf rank (where lower ranks indicate more frequent entities) and the probability $p(v_a \mid p_a)$ assigned to the correct attribute value when given only the probe prompt p_a, averaged across all attributes. The results reveal a strong negative correlation, demonstrating that parametric knowledge strength is tightly coupled with entity frequency.

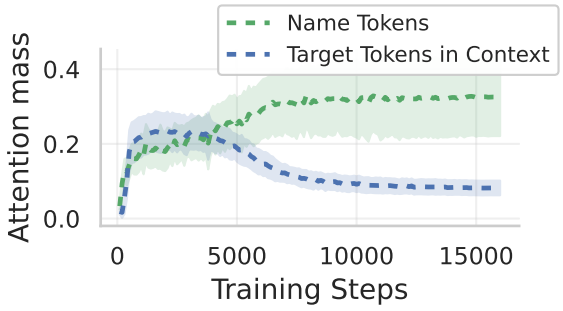


Figure 12: Changes in the layer-wise sum of attention mass at the last token of the test probe when the model trained with 1% noise performs in-context knowledge utilization for $\mathcal{E}_{\text{unseen}}$ entities. **Green** indicates the attention allocated to name tokens in the test probe, while **blue** indicates the attention allocated to target tokens in the context.

E Attention Pattern Analysis

We analyze attention patterns to investigate the mechanisms underlying the degradation of in-context knowledge utilization observed in Section 3.2. By examining the model trained on the REPEATED corpus with 1% inconsistency, we indirectly examine the circuits used for parametric versus in-context knowledge utilization.

We analyzed the attention patterns at the last token position of the test probe during in-context knowledge utilization for $\mathcal{E}_{\text{unseen}}$ entities. Figure 12 shows the layer-wise sum of attention mass over the course of training. We distinguish between two types of attention targets: (1) name tokens in the test probe (shown in green), which are more associated with parametric knowledge retrieval (Meng et al., 2022; Zucchet et al., 2025; Geva et al., 2023), and (2) target attribute tokens in the context (shown in blue), which are needed for in-context knowledge utilization (Olsson et al., 2022).

We find that early in training, attention is heavily concentrated on target attribute tokens in the context, consistent with successful in-context knowledge use mediated by in-context induction circuits (Olsson et al., 2022). However, as training progresses and parametric knowledge use stabilizes, attention gradually shifts toward name tokens

in the test probe. Notably, this shift occurs even when evaluating on $\mathcal{E}_{\text{unseen}}$ entities, for which the model has no parametric knowledge (see Table 5).

We hypothesize that the presence of inconsistency noise during training introduces imperfection in in-context knowledge utilization, making contextual information a less reliable signal. As a result, once parametric knowledge becomes sufficiently stable, the model increasingly defaults to parametric knowledge retrieval across all situations. Consequently, in-context knowledge utilization circuits receive progressively less training signal. In combination with regularization effects such as weight decay (Loshchilov and Hutter, 2017), this reduced usage leads to a gradual degradation of the model's ability to utilize in-context knowledge over the course of training.

This analysis helps explain how a skewed knowledge distribution (Section 3.3) can preserve in-context knowledge utilization. The continuous presence of unfamiliar or low-frequency entities in the training distribution forces the model to repeatedly rely on in-context knowledge, thereby preventing the complete abandonment of in-context knowledge circuits.

F Evaluation on Real-World LLMs

We adapt the evaluation scenarios used in our controlled experiments to settings applicable to large language models trained on real-world web corpora. Since such corpora contain abundant information about countries and their capitals, we define the set of training entities $\mathcal{E}_{\text{train}}$ as Real-World Countries and evaluate whether the model can correctly predict their corresponding capital cities.

To this end, we construct a Real-World Country and Capital Set based on the country and capital pairs introduced in Hernandez et al. (2023). We then build question and answer style probes as illustrated in Figure 13 and define the **Parametric Knowledge Utilization** (PKU) scenario. We measure Acc_{PKU} by checking whether the correct capital appears within the first 64 generated tokens.

For the **In-Context Knowledge Utilization** (ICKU) scenario, we evaluate the model's ability to use knowledge provided only in the prompt context. We construct 100 artificial country and capital pairs that do not correspond to any real-world entities, forming a Synthetic Country and Capital Set. These pairs are provided only within the prompt context, and Acc_{ICKU} is computed by verifying whether the correct synthetic capital is generated within 64 tokens.

Finally, for **Knowledge Conflict Resolution**, we perturb the in-context knowledge by replacing the true capitals in the Real-World Country and Capital Set with incorrect alternatives. Given these perturbed contexts and the corresponding test probes, we evaluate whether the model follows the in-context knowledge or instead relies on its parametric knowledge. This allows us to compute Pref_{ICK} and Pref_{PK}, reflecting the model's preference under explicit knowledge conflict.

Extending the results with the Pythia models in Section 4.1, we conduct the same set of experiments on Olmo-7B (Groeneveld et al., 2024). As shown in Figure 14, Olmo-7B exhibits qualitatively similar patterns to those observed in Pythia: in-context knowledge utilization emerges earlier, followed by the stabilization of parametric knowledge utilization, and preference shifts in resolving conflicts between parametric and in-context knowledge. These results suggest that the knowledge utilization dynamics identified in our analysis are not specific to a particular model family, but also appear across different large-scale language models trained on real-world data.

G Additional Experimental Results

We further characterize the influence of training data; unless otherwise noted, all experiments are conducted on the REPEATED corpus.

G.1 Effect of the Number of Training Entities

Figure 15a compares REPEATED runs with 50k, 100k, and 200k training entities. With 50k entities, both in-context knowledge utilization (Acc_{ICKU}) and parametric knowledge utilization (Acc_{PKU}) emerge, with Acc_{ICKU} emerging earlier and Acc_{PKU} following as training stabilizes. In contrast, for 100k and 200k entities, Acc_{PKU} fails to rise: the model learns to use in-context knowledge but does not develop robust parametric utilization. This suggests that when the number of entities exceeds the model's capacity, parametric knowledge cannot be stably acquired, leaving in-context knowledge utilization as the dominant strategy.

Figure 15b examines training dynamics under intra-document inconsistency levels of 1%, 5%, and 10%. Even 1% noise is sufficient to induce a phase shift in conflict-time preference: as Acc_{PKU} stabilizes, the model transitions from preferring

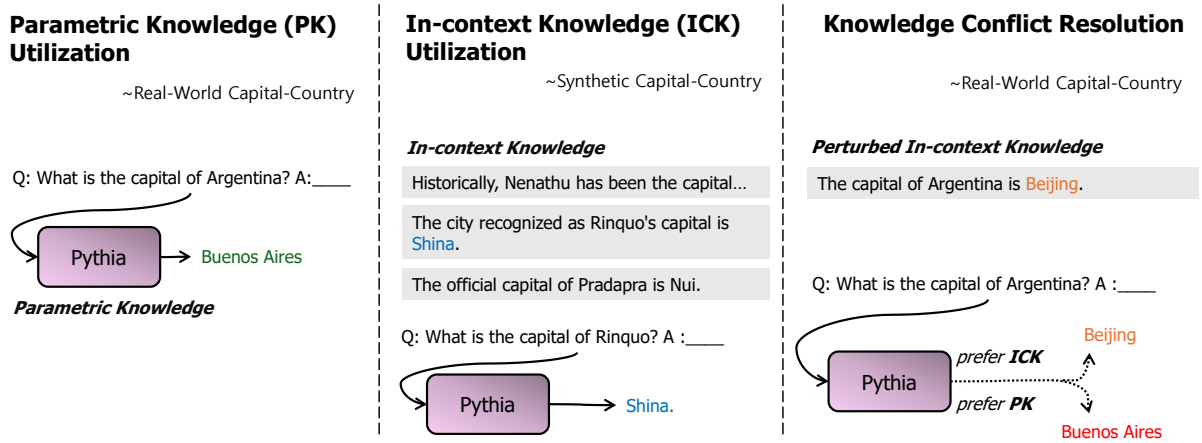


Figure 13: Three knowledge utilization scenarios in real-world large language models. (**Left**) Parametric knowledge utilization, where the model recalls country and capital facts from real-world data that were encoded in its parameters during training. (**Middle**) In-context knowledge utilization, where the model relies on synthetic country and capital pairs provided only in the context. (**Right**) Knowledge conflict resolution, where the model is queried about real-world countries while the prompt supplies perturbed (incorrect) capitals, allowing us to examine whether the model prefers parametric knowledge or the perturbed in-context knowledge.

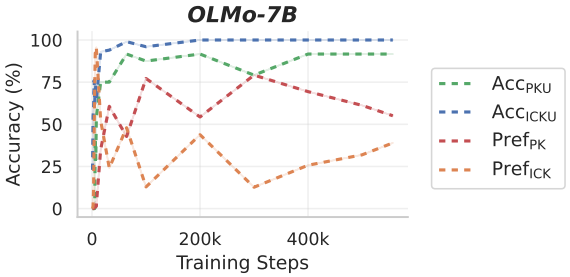


Figure 14: Evaluation results of knowledge utilization and conflict resolution in Olmo-7B.

in-context knowledge ($\mathrm{Pref}_{\mathrm{ICK}}$) to preferring parametric knowledge ($\mathrm{Pref}_{\mathrm{PK}}$). Increasing noise accelerates this shift but also degrades $\mathrm{Acc}_{\mathrm{ICKU}}$ at convergence, indicating over-reliance on parametric knowledge and a reduced ability to use in-context knowledge.

Figure 15c examines training dynamics under Zipfian sampling with $\alpha \in \{0.5, 1.0, 2.0\}$. A near-uniform regime (α=0.5) yields progressive degradation of $\mathrm{Acc}_{\mathrm{ICKU}}$ over training, consistent with the model drifting toward parametric recall even for unfamiliar entities. An overly skewed regime (α=2.0) produces undesirable dynamics in which parametric utilization fails to activate, because most entities appear too rarely to accumulate sufficient parametric learning signal. A moderate skew (α=1.0) best preserves $\mathrm{Acc}_{\mathrm{ICKU}}$ for rare or novel entities while still supporting stable $\mathrm{Acc}_{\mathrm{PKU}}$ and a robust preference for parametric knowledge on frequently seen facts.

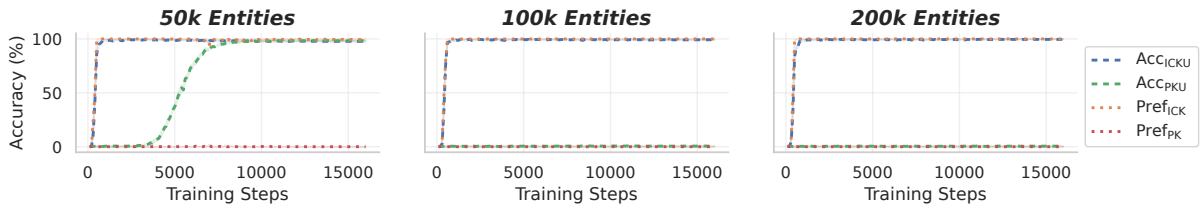
H Controlling In-Context Preference via Attention Head Amplification

We investigate whether the arbitration strategies identified in our controlled experiments can be further modulated at inference time through attention-head manipulation, providing a complementary view to our training-data analysis. Following Yu et al. (2023), we identify in-context heads, namely attention heads that contribute most to in-context knowledge utilization, and scale their activations by a factor α to amplify or diminish their influence during knowledge conflicts.

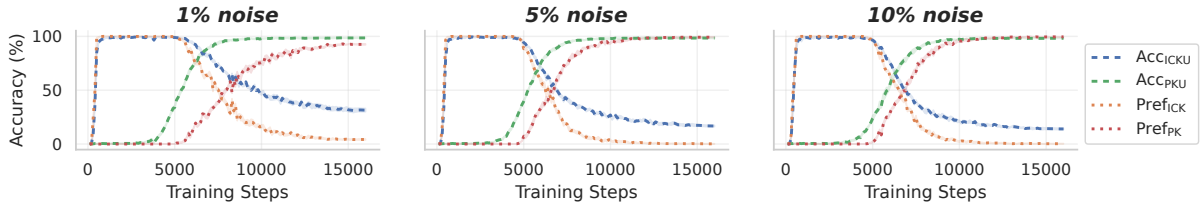
We apply this intervention to two models trained under different corpus conditions on the REPEATED corpus with Zipfian distribution: one trained without noise, and one trained with 1% inconsistency noise. We then measure conflict-time preferences for high-frequency entities across $\alpha \in \{0.0, 0.5, 1.0, 2.0\}$ (Table 6).

	Zipf, 0% noise		Zipf, 1% noise	
α	$\mathrm{Pref}_{\mathrm{PK}}$ (%)	$\mathrm{Pref}_{\mathrm{ICK}}$ (%)	$\mathrm{Pref}_{\mathrm{PK}}$ (%)	$\mathrm{Pref}_{\mathrm{ICK}}$ (%)
0.0	88.00	7.00	100.00	0.00
0.5	42.75	40.00	98.25	0.50
1.0	19.50	69.50	97.00	1.75
2.0	9.25	81.75	74.75	17.50

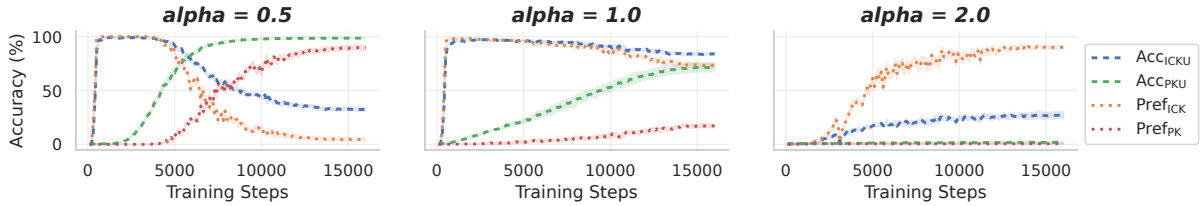
Table 6: Effect of in-context head amplification on conflict-time preference for high-frequency entities. α scales the activations of identified in-context heads.



(a) Effect of the number of training entities.



(b) Effect of intra-document inconsistency noise.



(c) Effect of knowledge distribution (Zipf exponent α).

Figure 15: Evaluation results during training of in-context knowledge utilization ($\mathrm{Acc}_{\mathrm{ICKU}}$), parametric knowledge utilization ($\mathrm{Acc}_{\mathrm{PKU}}$), and knowledge conflict preferences ($\mathrm{Pref}_{\mathrm{ICK}}$, $\mathrm{Pref}_{\mathrm{PK}}$) under varying corpus properties.

Both models show that in-context preference can be steered through α, consistent with observations on real-world models. However, the response magnitude differs substantially with training-data characteristics: the model trained with 1% noise exhibits a much stronger baseline preference for parametric knowledge and requires large amplification ($\alpha = 2$) to noticeably elicit in-context behavior, whereas the noise-free model transitions toward in-context preference even at $\alpha = 0.5$. This reflects the training-data-induced arbitration bias identified in our main experiments, in which the noisy pretraining corpus instills a persistent reliance on parametric knowledge that resists simple inference-time intervention. Since attention-head manipulation can also adversely affect the model's general capabilities (Li et al.), our training-data analysis and post-hoc inference-time interventions are best viewed as complementary rather than substitutes.

I The Use of Large Language Models

We used large language models to assist with the preparation of this paper. Specifically, they were employed for writing support, including grammar correction, wording refinement, and minor stylistic edits, as well as for developing code used in the experiments.