

TRAINING DYNAMICS OF PARAMETRIC AND IN-CONTEXT KNOWLEDGE UTILIZATION IN LANGUAGE MODELS

Minsung Kim¹ Dong-Kyum Kim² Jea Kwon² Nakyeong Yang¹ Kyomin Jung¹
Meeyoung Cha²

¹Seoul National University ²Max Planck Institute for Security and Privacy (MPI-SP)
kms0805@snu.ac.kr mia.cha@mpi-sp.org

ABSTRACT

Large language models often encounter conflicts between in-context knowledge retrieved at inference time and parametric knowledge acquired during pretraining. Models that accept external knowledge uncritically are vulnerable to misinformation, whereas models that adhere rigidly to parametric knowledge fail to benefit from retrieval. Despite the widespread adoption of retrieval-augmented generation, we still lack a systematic understanding of what shapes knowledge-arbitration strategies during training. This gap risks producing pretrained models with undesirable arbitration behaviors and, consequently, wasting substantial computational resources after the pretraining budget has already been spent. To address this problem, we present the first controlled study of how training conditions influence models’ use of in-context and parametric knowledge, and how they arbitrate between them. We train transformer-based language models on a synthetic biographies corpus while systematically controlling various conditions. Our experiments reveal that intra-document repetition of facts fosters the development of both parametric and in-context capabilities. Moreover, training on a corpus that contains inconsistent information or distributional skew encourages models to develop robust strategies for leveraging parametric and in-context knowledge. Rather than viewing these non-ideal properties as artifacts to remove, our results indicate that they are important for learning robust arbitration. These insights offer concrete, empirical guidance for pretraining models that harmoniously integrate parametric and in-context knowledge.¹

1 INTRODUCTION

Large language models (Touvron et al., 2023; Brown et al., 2020; Biderman et al., 2023) store and use *parametric knowledge* (Geva et al., 2020; 2023; Meng et al., 2022) acquired during pretraining and increasingly leverage *in-context knowledge* through retrieval-augmented generation (Lewis et al., 2021; Ram et al., 2023; Shi et al., 2023), which supplies external documents at inference time. This allows models to incorporate up-to-date and domain-specific information beyond their training data. A central challenge appears when external documents conflict with parametric knowledge (Neeman et al., 2022), which forces the model to arbitrate between the two sources. The stakes are high when the retrieved content contains misinformation, noisy passages, or adversarially crafted text. Models that trust external sources uncritically become vulnerable to these risks, while models that rigidly rely on their parametric knowledge fail to benefit from valuable external information. Recent works (Xu et al., 2024) have studied how models behave under such knowledge conflicts, but most analyses have focused on analyzing or controlling the behavior of already-pretrained models (Ortu et al., 2024; Yu et al., 2023; Li et al.), without examining how training conditions shape arbitration. However, it is essential to understand during pretraining what factors determine how a model uses and arbitrates between its two knowledge sources, so as to avoid discovering undesirable arbitration behaviors only after pretraining has consumed substantial resources.

¹Our code is available at <https://github.com/kms0805/Training-Dynamics-of-PK-ICK>

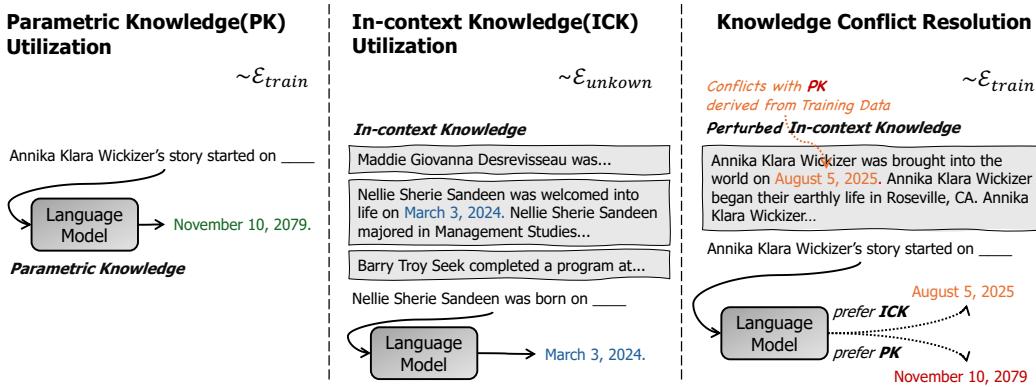


Figure 1: Three knowledge utilization scenarios. **Left:** parametric knowledge utilization where the model recalls knowledge encoded in its parameters and answers queries about entities seen during training. **Middle:** in-context knowledge utilization where the model extracts and uses knowledge provided only in the prompt and is evaluated on novel entities not seen during training. **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.

Determining the appropriate knowledge source for a model is often challenging, given the variable provenance and reliability of in-context information. Our work therefore defines a robust arbitration strategy based solely on the internal signals of the model, without considering external factors. We define this strategy by two principles: (1) for high-confidence, well-memorized knowledge, the model should follow its parametric knowledge, even when faced with conflicting in-context knowledge; and (2) for novel or unfamiliar information, the model should follow the provided in-context knowledge. This behavior, which mirrors patterns in human cognition (Koriat, 2011) and has been observed in modern large language models (Wu et al., 2024), motivates our investigation of the training factors that produce it. Consequently, our study is guided by two central research questions: **(RQ1)** What training conditions enable a model to develop distinct capabilities for the use of parametric and in-context knowledge? **(RQ2)** What specific characteristics of the training corpus induce the model to adopt a robust arbitration strategy between these two sources?

To answer these questions, we conduct controlled experiments, training transformer-based language models from scratch on synthetic biographies (Allen-Zhu & Li, 2024a,b; Zucchet et al., 2025). This framework enables precise manipulation of training conditions while isolating knowledge utilization from other confounding factors. Prior work (Allen-Zhu & Li, 2024a) shows the insights from this setup can transfer to real-world models (Touvron et al., 2023), which allows us to explore a wide range of training configurations efficiently. Building on earlier studies (Zucchet et al., 2025) that focused only on parametric knowledge, we systematically investigate the interaction between parametric and in-context knowledge. During training with varied conditions, we evaluate the model’s performance across three key knowledge utilization scenarios. First, we measure **parametric knowledge utilization** by the model’s ability to recall learned entity attributes from its parameters. Second, we assess **in-context knowledge utilization** by its capacity to extract and use knowledge from the context for novel entities that are not in the training data. Finally, we examine **knowledge conflict resolution** by observing which source the model follows when known entities are paired with perturbed contexts, where the in-context knowledge conflicts with the model’s parametric knowledge (Figure 1).

Our experiments led to the following findings: **Intra-document repetition** of facts is critical for the simultaneous emergence of both parametric knowledge and in-context knowledge utilization capabilities, and this in-context knowledge utilization ability emerges much earlier (Section 3). In addition, **a small degree of factual inconsistency** within a document encourages the model to favor its more confident parametric knowledge when conflicts arise, although during early training it initially prefers in-context knowledge (Section 4). Moreover, **a skewed frequency distribution of knowledge**, where long-tailed knowledge exists, preserves the model’s ability to use in-context

knowledge for unfamiliar entities. When these three conditions co-occur, they produce the desired arbitration pattern: the model defaults to parametric knowledge for well-learned entities but readily relies on in-context knowledge for rare or novel ones (Section 5). We validate these results on open-source LLMs, confirming that our findings extend beyond the synthetic setting (Section 6).

These results have critical implications for pretraining large language models for retrieval augmented generation. We find that data characteristics often seen as defects, such as modest inconsistencies and a skewed knowledge distribution, are actually beneficial features for developing models that can intelligently arbitrate between learned knowledge and new, in-context information. A direct implication of this finding is that preprocessing steps like aggressive cleaning, deduplication, and data balancing may inadvertently impair a model’s robust knowledge arbitration strategy.

2 EXPERIMENTAL SETUP

2.1 SYNTHETIC BIOGRAPHIES DATASET

We construct a synthetic biographies dataset following prior work (Allen-Zhu & Li, 2024a; Zuchet et al., 2025) (See details in Appendix A). Specifically, we generate synthetic biographical profiles, where each profile contains four attributes: `birth_date`, `birth_city`, `university`, and `major`. For each profile, we sample 7 distinct templates from a finite pool for each attribute. We use 6 templates to create training paragraphs with randomized attribute ordering, reserving 5 paragraphs for training and 1 for evaluation context. The remaining template for each attribute serves as test probes, which are cloze-style sentences designed to elicit the attribute values. Figure 8 illustrates the dataset structure. This deliberate separation ensures that the sentences used for training, context, and testing are never identical, compelling the model to utilize its parametric or in-context knowledge rather than relying on simple sequence memorization or repetition.

2.2 EVALUATION SETUP

During training, we periodically evaluate the model at each checkpoint under three knowledge utilization scenarios: parametric knowledge utilization, in-context knowledge utilization, and knowledge conflict resolution, to measure the model’s ability to utilize knowledge, as illustrated in Figure 1. We evaluate using the exact-match accuracy of the attributes generated by the model for the given input in each scenario. For each scenario, we randomly sample a set of k entities for evaluation, and in our experiments, we set $k = 200$.

Parametric Knowledge Utilization This scenario measures the model’s ability to utilize knowledge stored in its parameters. We evaluated this on entities seen during training, $e \sim \mathcal{E}_{\text{train}}$. The accuracy of parametric knowledge utilization is defined as $\text{Acc}_{\text{PKU}} = \mathbb{E}_{e \sim \mathcal{E}_{\text{train}}} \left[\frac{1}{|A_e|} \sum_{a \in A_e} \mathbf{1}\{M(p_a) = v_a\} \right]$, where A_e is the set of attributes of entity e , p_a is the test probe for attribute a , v_a is the ground-truth value, and $M(\cdot)$ is the model output.

In-Context Knowledge Utilization This scenario evaluates whether the model can utilize the knowledge provided only at inference time. We evaluated this on novel entities not seen during training, i.e., $e \sim \mathcal{E}_{\text{unknown}}$. For each unseen entity e , we construct a context C by concatenating C_e with paragraphs from two other random unseen entities, followed by shuffling. The accuracy of in-context knowledge utilization is defined as $\text{Acc}_{\text{ICKU}} = \mathbb{E}_{e \sim \mathcal{E}_{\text{unknown}}} \left[\frac{1}{|A_e|} \sum_{a \in A_e} \mathbf{1}\{M(C, p_a) = v_a\} \right]$.

Knowledge Conflict Resolution This scenario evaluates whether the model follows parametric knowledge (i.e., outputs the original training values) or in-context knowledge (i.e., outputs the values given in the perturbed context). For each training entity $e \sim \mathcal{E}_{\text{train}}$, we construct a perturbed context C'_e by randomly altering two attributes (`birth_date`, `major`). Preference for parametric knowledge is defined as $\text{Pref}_{\text{PK}} = \mathbb{E}_{e \sim \mathcal{E}_{\text{train}}} \left[\frac{1}{|A'_e|} \sum_{a \in A'_e} \mathbf{1}\{M(C'_e, p_a) = v_a\} \right]$, and preference for in-context knowledge as $\text{Pref}_{\text{ICK}} = \mathbb{E}_{e \sim \mathcal{E}_{\text{train}}} \left[\frac{1}{|A'_e|} \sum_{a \in A'_e} \mathbf{1}\{M(C'_e, p_a) = v'_a\} \right]$, where A'_e is the set of perturbed attributes, v_a denotes the original parametric value from training, and v'_a the conflicting value specified in C'_e .

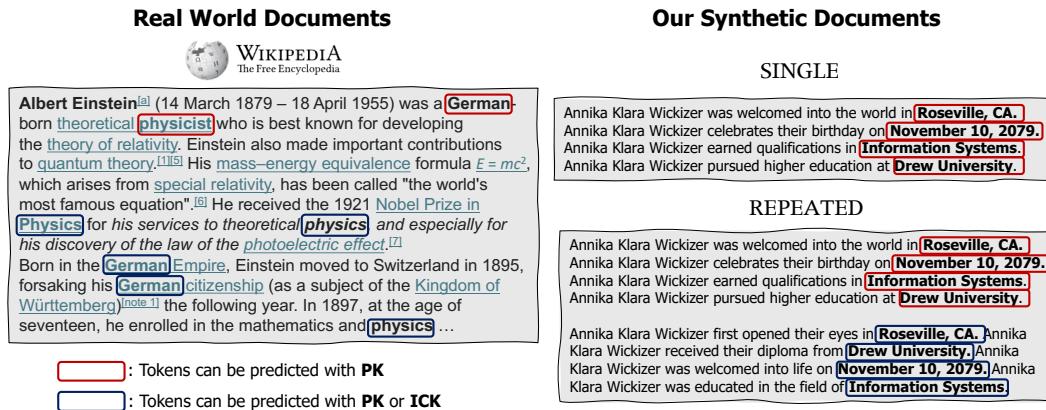


Figure 2: An example of intra-document repetition of key attributes (e.g., German, Physics) for a single entity, alongside our synthetic training-corpus variants. SINGLE uses one paragraph per entity and thus encourages reliance on parametric knowledge; REPEATED places two paraphrased paragraphs about the same entity in one document, allowing later mentions to leverage in-context knowledge or parametric knowledge.

2.3 TRAINING SETUP

We train an 8-layer decoder-only Transformer language model from scratch (Vaswani et al., 2017), adopting the detailed hyperparameters (Table 4) from prior work (Zucchet et al., 2025). For the training entity set ($\mathcal{E}_{\text{train}}$), we use profiles of $50k$ entities, and for the unknown entity set ($\mathcal{E}_{\text{unknown}}$), we use other $50k$ entity profiles that are unseen during training. Using the training paragraphs of $e \in \mathcal{E}_{\text{train}}$ from Section 2.1, we assemble documents according to the variants described below and use the resulting collection as the training corpus.

Training corpus variants. Our aim is to examine how a model uses and arbitrates between parametric knowledge and in-context knowledge. We start by hypothesizing the conditions under which a model acquires the ability to leverage both sources of knowledge. Our hypothesis is motivated by a common feature of real-world text: key attributes of an entity are often repeated within a single document (Figure 2). During next token prediction pretraining (Radford et al., 2019), the first mention of an attribute appears without prior in-document context, pushing the model to rely on parametric knowledge, whereas later mentions allow the model to leverage earlier context for prediction. We therefore hypothesize that *intra-document repetition* serves as a critical mechanism for the simultaneous development of both parametric and in-context knowledge utilization capabilities. To empirically test this hypothesis, we construct and analyze three corpus variants that systematically control for the presence of this repetition:

- **SINGLE:** Each training document contains exactly one training paragraph about a single entity. Attributes appear once per document, so the model cannot rely on in-context knowledge; predictions must be supported by parametric knowledge.
- **REPEATED:** Each training document concatenates two distinct paraphrased paragraphs about the same entity. Every attribute is mentioned twice within the same document using different sentence templates and randomized attribute order. The first mention requires parametric knowledge; the second can leverage in-document in-context knowledge from the earlier paragraph.
- **REPEATED + MIX:** To simulate a more complex and realistic scenario, we sample two paragraphs for each of 3 distinct entities and shuffle their order to form a single training document with 6 paragraphs. As in the REPEATED variant, the model can use parametric knowledge for the first mention of an attribute and either parametric or in-context knowledge for the second; additionally, it must retrieve relevant evidence amid distractor paragraphs about other entities.

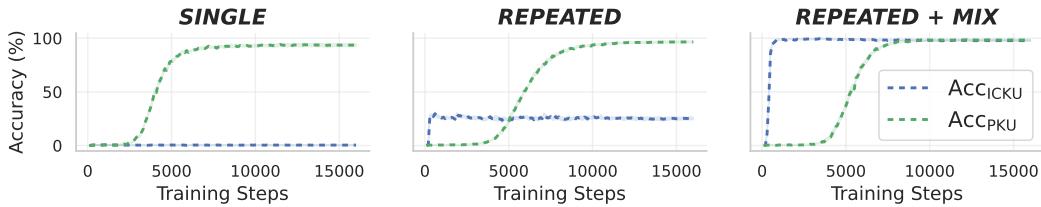


Figure 3: Accuracy of parametric knowledge utilization (Acc_{PKU}) and in-context knowledge utilization (Acc_{ICKU}) across training steps. **Left:** The model trained on the SINGLE corpus shows delayed parametric knowledge utilization and no activation of in-context knowledge utilization. **Middle:** In contrast, the REPEATED+MIX corpus induces early in-context knowledge utilization followed by parametric knowledge utilization. **Right:** The REPEATED corpus remains near random-guess performance on in-context knowledge utilization.

3 EMERGENCE OF PARAMETRIC AND IN-CONTEXT KNOWLEDGE UTILIZATION CAPABILITIES

Figure 3 shows the results of parametric knowledge utilization (Acc_{PKU}) and in-context knowledge utilization (Acc_{ICKU}) across our training on various corpora we constructed. The model trained on the SINGLE corpus developed parametric knowledge utilization midway through training, while in-context knowledge utilization did not become activated. In contrast, the model trained on the REPEATED+MIX corpus learned to utilize both forms of knowledge, with in-context knowledge utilization emerging earlier than parametric knowledge utilization and the latter gradually becoming activated afterward.

As revealed in previous work (Zucchet et al., 2025), parametric knowledge requires the formation of an attention circuit that connects subject entity tokens with attributes in order to store information about a particular subject in a key–value format (Meng et al., 2022; Geva et al., 2023; 2020). Once such a circuit is established, attribute errors flow along this pathway and are learned through back-propagation, which leads to a delayed rise in Acc_{PKU} during learning. By comparison, in-context knowledge requires only the operation of induction heads (Olsson et al., 2022) in the attention module, which copy and paste the attribute tokens necessary. This is a simpler mechanism that reflects learning of general utilization patterns rather than knowledge of specific entities, and therefore it is presumed to emerge earlier. Meanwhile, the model trained on the REPEATED corpus remains at the level of random guessing (~33%) on in-context knowledge utilization tasks. Here, the model is required to locate entity-specific information in the context; however, it instead learns only a shallow heuristic, randomly retrieving attributes of the correct type from the context without resolving entity identity.

From these observations, we confirm that the ability to utilize parametric knowledge does not automatically entail the ability to utilize in-context knowledge. Rather, when the training corpus provides a sufficiently complex environment through intra-document repetition that enables the use of in-context knowledge, the ability to utilize in-context knowledge tends to emerge first, followed by the gradual activation of parametric knowledge utilization.

4 EFFECTS OF FACTUAL INCONSISTENCY NOISE WITHIN A DOCUMENT

4.1 MODELS TRAINED WITHOUT NOISE OVER-RELY ON IN-CONTEXT KNOWLEDGE

The model trained on the REPEATED+MIX corpus was shown to leverage both parametric knowledge and in-context knowledge utilization. To investigate which type of knowledge the model prefers when these two conflict, we evaluated it using the knowledge conflict resolution scenario introduced in Section 2.2, measuring Pref_{ICK} and Pref_{PK} . The results (Figure 4(a) left) demonstrate that once in-context knowledge utilization is activated, the model invariably follows in-context knowledge in conflict situations and predicts attributes solely based on the perturbed context. In particular, this preference persists even after Acc_{PKU} reaches nearly 100% and the use of parametric knowledge has stabilized, showing that the model continues to rely on in-context knowledge whenever it is available.

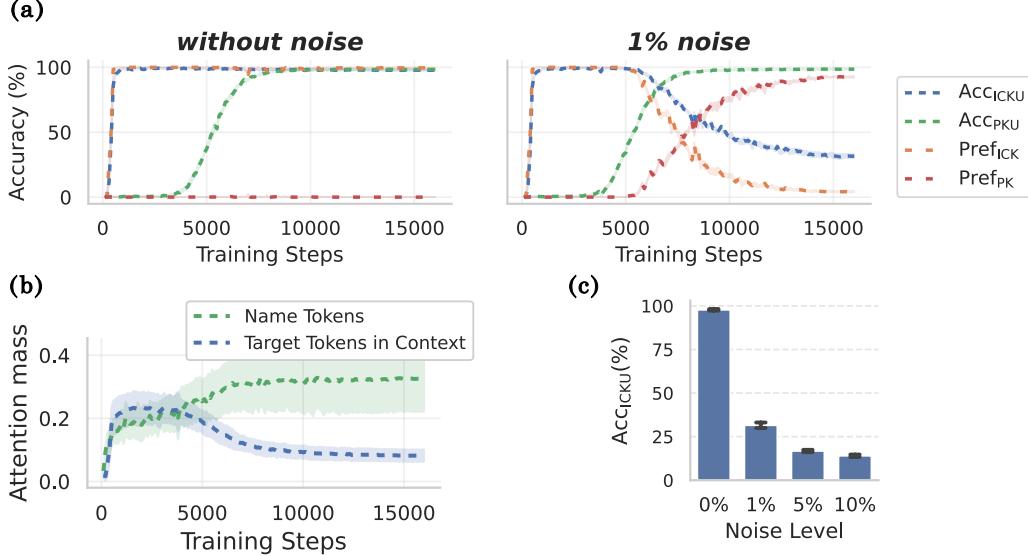


Figure 4: (a) Training dynamics of Acc_{ICKU} , Acc_{PKU} , Pref_{ICK} , and Pref_{PK} when trained on the REPEATED+MIX corpus without noise (**Left**) and with 1% noise (**Right**). When the training corpus contains no noise (i.e., no inconsistent knowledge within the same documents), the model consistently prefers in-context knowledge in knowledge conflicts, whereas even a small amount of noise induces a phase shift toward parametric knowledge preference as parametric knowledge utilization stabilizes. (b) 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. Green indicates the attention allocated to name tokens in the test probe, while blue indicates the attention allocated to target tokens in the context. (c) Acc_{ICKU} at the end of training across different noise levels.

We examined entropy and target token probability when predicting attributes using only parametric knowledge (that is, without any context, using only test probes) for entities in $\mathcal{E}_{\text{train}}$ and $\mathcal{E}_{\text{unknown}}$ to measure the confidence of the model in its parametric knowledge. As shown in Table 1, the model is considerably more confident (low entropy and high target probability) about the correct parametric knowledge for training entities compared to unseen entities. Nevertheless, when presented with contradictory in-context information, it still follows the in-context knowledge. This tendency to over-rely on external context deviates from the robust arbitration strategy that we aim to establish.

4.2 INCONSISTENCY NOISE MAKES MODELS USE PARAMETRIC KNOWLEDGE ROBUSTLY

However, in practice, large language models pretrained on real web corpus do not always follow in-context knowledge in conflict situations. Instead, they tend to prefer parametric knowledge for knowledge they have frequently encountered during training (Yu et al., 2023) or for information with high internal confidence (Wu et al., 2024), even when conflicting in-context knowledge is present. We hypothesize that the reason is the inevitable presence of noise in web corpora—such as typos, factual errors, or conflicting opinions—which introduces small inconsistencies within a document. Such noise likely prevented the model from always following only in-context knowledge when the two sources of knowledge conflicted.

To test this hypothesis, we trained models on the REPEATED+MIX corpus with a small degree of factual inconsistency noise within a document. In our setup, each entity is mentioned in two paragraphs within a document (a leading paragraph and a later one). On a per-entity basis, we perturb the `birth_date` and `major` values only in the leading paragraph with probability

Table 1: Target token probability and entropy measured at the last token of the test probe for entities in $\mathcal{E}_{\text{train}}$ and $\mathcal{E}_{\text{unknown}}$.

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

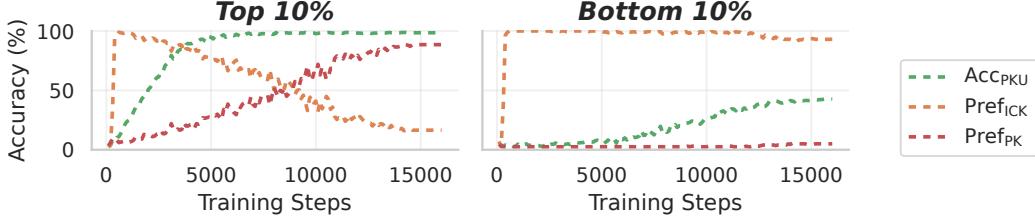


Figure 5: Acc_{PKU} , Pref_{ICK} , and Pref_{PK} for the top 10% (high-frequency) and bottom 10% (low-frequency) entities in the training corpus. For high-frequency entities, Pref_{ICK} is initially higher but gradually yields to Pref_{PK} ; for low-frequency entities, Pref_{ICK} remains consistently higher.

$p \in \{1\%, 5\%, 10\%\}$, replacing them with randomly sampled alternatives, and leave the later paragraph unchanged (Figure 9). The results are shown in Figure 4(a) right. Early in training, the model first activates in-context knowledge utilization, leading it to prefer in-context knowledge in knowledge conflicts. As parametric knowledge utilization stabilizes, however, the model gradually shifts toward always preferring parametric knowledge when conflicts arise. Remarkably, even very small amounts of noise (as little as 1%) were sufficient to induce this phase shift phenomenon.

However, in-context knowledge utilization performance gradually decreased, with the decline being more pronounced at higher noise levels (Figure 4(c)). We initially suspected that the degradation was caused by the model being overexposed to entities from $\mathcal{E}_{\text{train}}$, leading it to behave as if it knows unseen entities from $\mathcal{E}_{\text{unknown}}$. In other words, the model appears to overconfidently project its parametric knowledge onto unfamiliar entities, hallucinating attributes it never learned. Yet, as shown in Table 1, the model trained with 1% noise maintains very high confidence for the correct attributes of training entities, while showing the opposite pattern for unknown entities. This indicates that the model can, in fact, distinguish when it lacks parametric knowledge about new entities, as if it were simply forgetting how to utilize in-context knowledge.

In fact, when investigating the attention patterns at the last position of the test probe during in-context knowledge utilization for $\mathcal{E}_{\text{unknown}}$ entities, we observed that attention initially focuses heavily on target tokens in context but gradually shifts toward name tokens (Figure 4(b)). Prior works (Meng et al., 2022) have established that recalling parametric knowledge requires bringing information from relevant subject tokens through attention circuits (Zucchetti et al., 2025; Geva et al., 2023). Thus, even for unknown entities, the model appears to attempt recalling information from name tokens to utilize parametric knowledge, following the established mechanisms of parametric knowledge utilization. In other words, while the model recognizes that it lacks information about the entity from $\mathcal{E}_{\text{unknown}}$, it seems to have forgotten how to use in-context knowledge.

In conclusion, while a very small amount of inconsistency noise enables the model to robustly use parametric knowledge rather than uncritically following in-context knowledge when the two conflict, it also leads to an over-reliance on parametric knowledge, ultimately resulting in the degradation of in-context knowledge utilization.

5 EFFECTS OF SKEWED KNOWLEDGE DISTRIBUTION

5.1 SKEWED KNOWLEDGE DISTRIBUTION PRESERVES IN-CONTEXT KNOWLEDGE UTILIZATION ON UNFAMILIAR KNOWLEDGE

We hypothesize that to prevent the degradation of in-context knowledge utilization for unfamiliar knowledge, as observed in Section 4, the training data must continually expose the model to information that cannot be recalled purely from parametric knowledge. In other words, knowledge from long-tailed knowledge should appear repeatedly so that in-context knowledge utilization remains active and does not degenerate. To test this, we constructed a REPEATED+MIX corpus where entities are sampled according to a Zipfian distribution (Zipf, 2012)² (with small inconsistency noise as in Section 4). As shown in Table 2, training on this corpus yielded substantially less degradation in in-context knowledge utilization compared to training on a corpus with a uniform knowledge distribution.

²Zipfian distribution: $P(r) = r^{-\alpha} / \sum_{k=1}^N k^{-\alpha}$, where r is the rank (1 = most frequent).

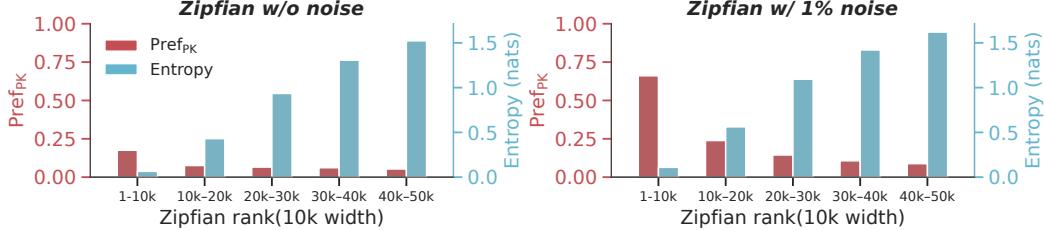


Figure 6: Bar plots of Pref_{PK} under knowledge conflict (red) and mean entropy in the parametric-knowledge-utilization setting (blue). Bins are ordered by Zipfian rank, where lower rank denotes higher frequency. **Left:** Results with zipfian training corpus without inconsistency noise. **Right:** Results with zipfian training corpus with a small amount(1%) of inconsistency noise.

We further evaluated preference measures across the top 10% and bottom 10% of entities by frequency (Figure 5) to examine under knowledge conflicts which knowledge the model follows for entities it saw frequently during training versus those it saw infrequently.

For high-frequency entities, the model initially preferred in-context knowledge but gradually shifted toward a robust reliance on parametric knowledge, as when trained under a uniform distribution. In contrast, for low-frequency entities, the model continued to prefer in-context knowledge. Importantly, this is not simply due to an inability to recall these entities from parametric knowledge. Because Acc_{CKU} exceeds Pref_{PK} , the model retains parametric knowledge for some of these entities, yet in conflict settings it still prefers in-context knowledge over parametric knowledge. This tendency to rely on in-context knowledge for low-frequency entities supports our hypothesis that the model continues to use in-context knowledge when unfamiliar knowledge arises during training, thereby preventing degradation of in-context knowledge utilization capability.

5.2 SKEWNESS ALONE FAILS TO BUILD ROBUST PARAMETRIC KNOWLEDGE PREFERENCE

We investigate whether, for a model trained on a corpus with a Zipfian distribution and a small amount of inconsistency noise, internal confidence in parametric knowledge calibrates to entity frequency and whether preference under conflict aligns with this confidence. We quantify confidence using the entropy of predictions during parametric knowledge utilization. As frequency increases, equivalently as Zipfian rank decreases, entropy declines and the preference for parametric knowledge in conflicts Pref_{PK} rises, as shown in Figure 6 right. In other words, confidence grows with training exposure, and the model correspondingly follows its parametric knowledge even when conflicting in-context knowledge is provided.

We then test whether the confidence-aligned robust knowledge arbitration strategy can emerge with skewness alone, without inconsistency noise. In this setting, confidence still tracks frequency, as in Figure 6 left. However, the model rarely follows parametric knowledge under conflict even when its parametric knowledge has high confidence (low entropy). Thus, a skewed knowledge distribution helps preserve in-context knowledge utilization but is insufficient on its own to yield a robust knowledge arbitration strategy.

6 VALIDATION ON REAL-WORLD MODELS

We showed that (i) intra-document repetition enables the joint emergence of parametric and in-context knowledge utilization, (ii) a small degree of factual inconsistency noise within a document biases conflict resolution toward confident parametric knowledge, and (iii) distributional skew with long-tailed knowledge preserves in-context utilization for unfamiliar entities. Because these properties arise naturally in web corpora, we test whether the same dynamics appear in a real-world open-source LLM. Using the publicly released checkpoints of PYTHIA-6.9B (Biderman et al., 2023), we

Table 2: Acc_{CKU} at the end of training on uniform vs. Zipfian ($\alpha = 1$) corpora. The Zipfian column shows the change relative to the corresponding uniform value in parentheses.

Noise	$\text{Acc}_{\text{CKU}} (\%)$	
	Uniform	Zipfian
1%	31.5	84.0 (+52.5)
5%	16.8	63.9 (+47.1)
10%	14.1	57.4 (+43.3)

evaluate parametric utilization, in-context utilization, and preference under knowledge conflict at each checkpoint (details in Figure 13 and Appendix E).

As shown in Figure 7, in-context utilization becomes effective earlier in training than parametric utilization. Under conflict, the model initially prefers in-context knowledge and later shifts toward preferring parametric knowledge as parametric utilization stabilizes. Meanwhile, the model continues to leverage in-context knowledge for novel entities, as reflected by high Acc_{ICKU} . These results indicate that repetition, small amounts of inconsistency noise, and skewed knowledge distributions in web-scale data naturally reproduce the dynamics observed in our synthetic setting, suggesting that our controlled findings extend to real-world scenarios.

7 RELATED WORKS

Large language models rely on both parametric and in-context knowledge (Lewis et al., 2021; Mallen et al., 2022; Ram et al., 2023; Shi et al., 2023). Recent studies show that model preferences in conflicts depend on confidence and training frequency (Wu et al., 2024; Yu et al., 2023), and can be steered through attention manipulation or contrastive decoding (Li et al.; Yu et al., 2023; Sun et al., 2025; Jin et al., 2024). However, these works mainly focus on post-pretraining behavior and provide limited insight into how the ability to handle the two sources develops during training. A complementary line of research investigates training dynamics of language models using synthetic datasets (Allen-Zhu & Li, 2024a;b; Zucchet et al., 2025), enabling controlled studies of how models acquire and store knowledge. While these studies illuminate the formation of parametric knowledge, they do not address the simultaneous development of in-context utilization (Olsson et al., 2022) or the dynamics of conflict resolution. We bridge these directions by conducting the first systematic analysis of how parametric and in-context knowledge utilization co-emerge and interact during pre-training.

In parallel, some studies (Chan et al., 2022) that investigate in-context learning with transformer classifiers on Omniglot datasets (Lake et al., 2019) report that a skewed data distribution is required for in-context and in-weight (parametric) learning to co-exist. In contrast, our results show co-existence even under a uniform distribution. We attribute this difference to the task setup: those classification tasks can be solved solely from exemplars provided in context, predicting only a class token conditioned on a query, whereas a language model performs next-token prediction for every data sequence and thus must rely on parametric knowledge for most initial content, which more strongly incentivizes reliance on parametric knowledge than in exemplars-conditioned classification. Building on this distinction, we study knowledge-conflict scenarios (Neeman et al., 2022) and settings closer to how real-world language models are trained (Brown et al., 2020).

8 CONCLUSION

We present the first systematic analysis of the training dynamics that govern parametric knowledge and in-context knowledge in language models. Our study shows that properties of the training corpus strongly shape the emergence of both knowledge-utilization capabilities and their effective arbitration. Intra-document repetition of facts is crucial for developing both parametric and in-context knowledge utilization. Moreover, small degrees of factual inconsistency together with skewed knowledge distributions are key to fostering a robust arbitration strategy between parametric and in-context knowledge. These findings challenge traditional data cleaning practices, highlighting that modest noise and skewed distributions can enhance a model’s ability to intelligently utilize both knowledge sources. Our results provide valuable guidelines for designing training corpora considering retrieval augmented generation settings, ensuring that models can effectively balance new information and prior knowledge with robust knowledge arbitration strategies.

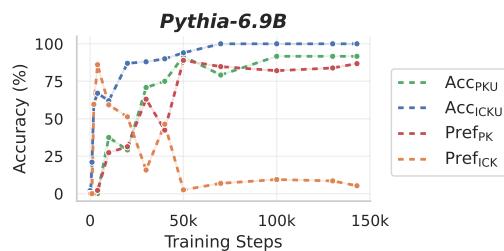


Figure 7: Acc_{ICKU} , Acc_{PKU} , Pref_{ICK} , and Pref_{PK} in Pythia checkpoints.

REPRODUCIBILITY STATEMENT

We describe the dataset construction process in detail in Section 2 and Appendix A. The hyperparameters and model configuration used in our experiments are provided in Appendix B. Furthermore, we will release code for experiments publicly. All experiments are implemented using the HuggingFace TRL library³ and conducted on a single NVIDIA A100 GPU. Each training run requires approximately 4–6 hours.

REFERENCES

- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction, 2024a. URL <https://arxiv.org/abs/2309.14316>.
- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation, 2024b. URL <https://arxiv.org/abs/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, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.
- 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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891, 2022.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023.
- Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*, 2023.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jie Xin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikanth (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70. URL <https://aclanthology.org/2024.findings-acl.70/>.
- Asher Koriat. The self-consistency model of subjective confidence. *Psychological Review*, 119: 80–113, 10 2011. doi: 10.1037/a0025648.
- Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. The omniglot challenge: a 3-year progress report, 2019. URL <https://arxiv.org/abs/1902.03477>.

³<https://huggingface.co/docs/trl/index>

- 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. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL <https://arxiv.org/abs/2005.11401>.
- Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In *Forty-second International Conference on Machine Learning*.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2022.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.
- Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. *arXiv preprint arXiv:2211.05655*, 2022.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- Francesco Ortú, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8420–8436, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.458. URL <https://aclanthology.org/2024.acl-long.458/>.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models, 2023. URL <https://arxiv.org/abs/2302.00083>.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. Replug: Retrieval-augmented black-box language models, 2023. URL <https://arxiv.org/abs/2301.12652>.
- Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. Redep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability, 2025. URL <https://arxiv.org/abs/2410.11414>.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Kevin Wu, Eric Wu, and James Y Zou. Clasheval: Quantifying the tug-of-war between an llm’s internal prior and external evidence. *Advances in Neural Information Processing Systems*, 37: 33402–33422, 2024.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey, 2024. URL <https://arxiv.org/abs/2403.08319>.

Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9924–9959, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.615. URL <https://aclanthology.org/2023.emnlp-main.615/>.

G.K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Martino Fine Books, 2012. ISBN 9781614273127. URL <https://books.google.co.kr/books?id=nR06MAEACAAJ>.

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

A SYNTHETIC BIOGRAPHIES DATASET CONSTRUCTION

Following prior work (Allen-Zhu & 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 from 1900–2099. 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 templates 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`.

We then create paragraphs containing each person’s biography with a randomized attribute order as follows: using 6 of the templates, we generate six paragraphs per entity; five are reserved for

⁴<https://github.com/smashew/NameDatabases/tree/master/NamesDatabases>

⁵<https://en.wikipedia.org/wiki/>

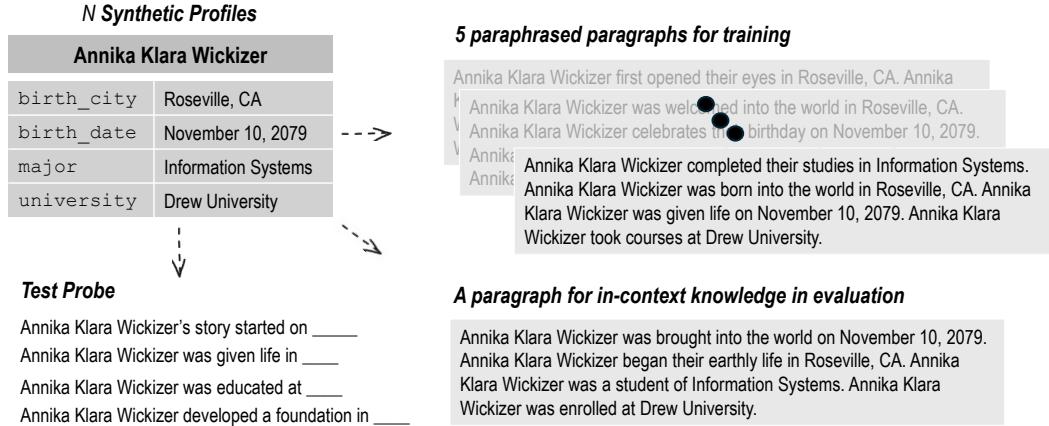


Figure 8: An example of the synthetic dataset. Each profile consists of four attributes (`birth_date`, `birth_city`, `university`, `major`), with paragraphs for training, a paragraph for in-context knowledge in evaluation, and test probes for eliciting the model to generate the attributes of each entity.

training and one is used as the evaluation in-context paragraph. The remaining (seventh) template is held out as a closed-style test probe designed to elicit the target attribute. An illustration of the resulting dataset is shown in Figure 8.

B DETAILS ON TRAINING LANGUAGE MODELS

Table 3: Model architecture.

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

Table 4: Training hyperparameters.

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

For our controlled experiments, we use a GPT-2-style decoder-only Transformer⁶. The model configuration is summarized in Table 3. Following Hoffmann et al. (2022), we adopt the settings used in Zucchet et al. (2025). The training hyperparameters are listed in Table 4.

⁶<https://huggingface.co/openai-community/gpt2>

C EXAMPLE OF FACTUAL INCONSISTENCY NOISE WITHIN A DOCUMENT

Figure 9 illustrates a document from the REPEATED+MIX corpus in which factual inconsistency noise has been injected. The value highlighted in pink was injected as noise with some probability and therefore does not match the latter original value, “November 10, 2079.”

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 9: Example of the document injected inconsistency noise

D ADDITIONAL EXPERIMENTAL RESULTS

We further examine the training dynamics by systematically varying several factors. Unless otherwise noted, all experiments are conducted on the REPEATED+MIX corpus.

D.1 EFFECT OF THE NUMBER OF TRAINING ENTITIES

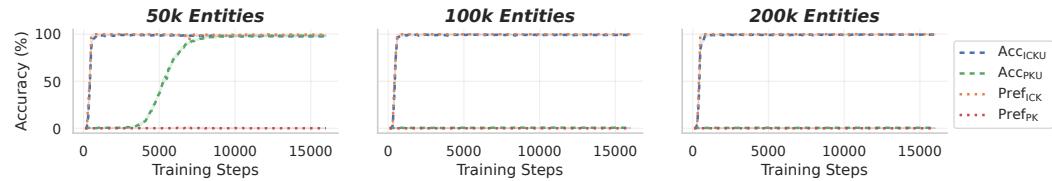


Figure 10: Training dynamics of AccICKU and AccPKU under different numbers of training entities.

Figure 10 compares REPEATED+MIX runs with 50k, 100k, and 200k training entities. With 50k entities, both in-context knowledge utilization (AccICKU) and parametric knowledge utilization (AccPKU) emerge, with AccICKU activating earlier and AccPKU 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.

D.2 EFFECT OF INTRA-DOCUMENT INCONSISTENCY NOISE

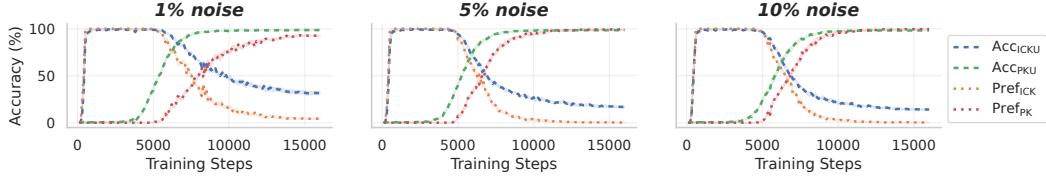


Figure 11: Training dynamics of Acc_{ICKU} and Acc_{PKU} under different levels of intra-document inconsistency noise.

Figure 11 examines training dynamics under intra-document factual 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 in-context knowledge (Pref_{ICK}) to preferring parametric knowledge (Pref_{PK}). Increasing noise accelerates this shift but also degrades Acc_{ICKU} at convergence, indicating over-reliance on parametric knowledge and a reduced ability to use in-context knowledge.

D.3 EFFECT OF DISTRIBUTIONAL SKEW

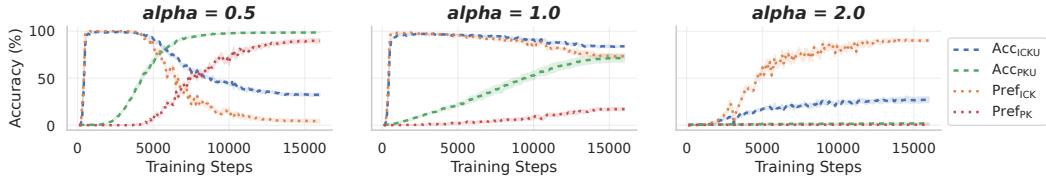


Figure 12: Training dynamics of Acc_{ICKU} and Acc_{PKU} as a function of the Zipf exponent α .

Figure 12 examines training dynamics under Zipfian sampling with $\alpha \in \{0.5, 1.0, 2.0\}$. A near-uniform regime ($\alpha=0.5$) yields progressive degeneration of Acc_{ICKU} over training, consistent with the model drifting toward parametric recall even for unfamiliar entities. An overly skewed regime ($\alpha=2.0$) produces undesirable dynamics—parametric utilization fails to activate—suggesting that extreme concentration of exposure undermines balanced capability growth. A moderate skew ($\alpha=1.0$) best preserves Acc_{ICKU} for rare or novel entities while still supporting stable Acc_{PKU} and a robust preference for parametric knowledge on frequently seen facts.

E EXPERIMENTAL DETAILS FOR REAL-WORLD LARGE LANGUAGE MODELS

We adapt the scenarios used in our controlled experiments so that they can be applied to models trained on real web corpora. Since web corpora contain abundant information about countries and their capitals, we designate 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–Capital Set* based on the country–capital data pairs used in Hernandez et al. (2023). Using this dataset, we build question–answer style test probes as illustrated in Figure 13, and define the **Parametric Knowledge Utilization** scenario. We then measure Acc_{PKU} by checking whether the model’s generations within 64 tokens contain the correct answer.

For the **In-Context Knowledge Utilization** scenario, we need to evaluate knowledge unseen during training. Therefore, we create 100 artificial country–capital pairs that do not exist in the real world, forming a *Synthetic Country–Capital Set*. As described in Section 2.2, we embed these pairs into a context and provide them to the model along with a test probe, measuring Acc_{ICKU} by verifying whether the correct answer appears within the first 64 generated tokens.

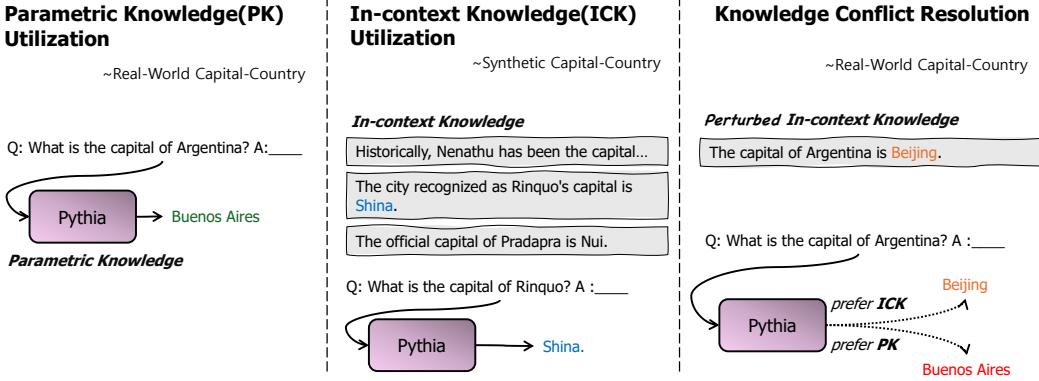


Figure 13: Three knowledge utilization scenarios in real-world large language models. **Left:** Parametric knowledge utilization, where the model recalls country–capital facts from real-world data that were encoded in its parameters during training. **Middle:** In-context knowledge utilization, where the model relies on synthetic country–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.

Finally, for **Knowledge Conflict Resolution**, we perturb the in-context knowledge by replacing the true answers in the *Real-World Country–Capital Set* with incorrect ones. We then provide these perturbed contexts together with the test probes and evaluate whether the model follows the perturbed in-context knowledge or the true answer. This allows us to measure Pref_{ICK} and Pref_{PK} .

F THE USE OF LARGE LANGUAGE MODELS

We used large language models solely to aid and polish the writing of this paper, including tasks such as grammar correction, wording refinement, and minor stylistic edits.