Investigating the Effect of Training Data Order on Small Language Model Fact Retention

Yair Ben Shimol yairb2@mail.tau.ac.il
Ido Tamir idotamir1@mail.tau.ac.il
Harel Ben Shoshan harelb2@mail.tau.ac.il

Abstract

This project explores how the ordering of training data affects fact retention in small language models (LLM). We investigate whether facts presented at the beginning or end of a training dataset are more likely to be retained and retrieved. We fine-tune small open-source models and evaluate their ability to recall fictional biographical facts inserted at different corpus positions. Our findings shed light on whether the ordering of the data set biases memorization and influences the accuracy of the retrieval.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across a wide range of natural language processing tasks. Despite their success, the mechanisms by which these models memorize and retain factual information are only partially understood. A growing body of work suggests that memorization in LLMs is not uniform and that factors such as dataset composition, token frequency, and fine-tuning procedures can significantly affect retention. However, one important and underexplored factor is the *global ordering* of the training data.

Transformer-based models are trained on sequences of tokens that are typically drawn from large corpora. Although the internal architecture processes input locally within batches, the order in which the training data is presented during optimization may introduce positional biases in memorization. Specifically, it remains unclear whether the information presented early in the training corpus is retained differently than the information encountered later.

In this work, we investigate the extent to which training data order influences fact retention in small-language models. To this end, we fine-tune three open source transformer models of different sizes (160M, 410M, and 1B parameters) on a controlled corpus combining a base dataset with synthetic question–answer pairs. Our synthetic data set consists of 30 made-up questions, each paired with two fictitious answers designed to be non-sensical and thus free of prior model bias.

We design two experiments: (1) comparing retention of synthetic facts when placed at the beginning versus the end of the corpus, and (2) introducing contradictory facts, where one answer is placed early and the other late, to test which ordering dominates. The evaluation is based on the probability ranking and top-k accuracy. Our study provides empirical evidence on whether early or late placement of data exerts a stronger influence during fine-tuning, with implications for dataset design in resource-constrained scenarios.

2 Related Work

2.1 Memorization in Language Models

A growing literature has investigated memorization in large language models (LLMs). Carlini et al. (2023) showed that models can memorize rare or unique training sequences and that memorization scales with model capacity and data duplication. Tirumala et al. (2022) analyzed training dynamics, showing that larger models memorize faster and retain longer. Speicher et al. (2024) examined memorization using synthetic random strings, revealing distinct phases and token-level dynamics. Wei et al. (2024) provide a broad survey of memorization, covering mechanisms, evaluation methods, and implications for privacy. While these studies explain what and how LLMs memorize, they do not isolate whether the *order* of exposure to data matters.

2.2 Catastrophic Forgetting and Order Effects

Order sensitivity is also well documented in continual learning. Catastrophic forgetting describes how new knowledge can overwrite earlier information. To mitigate this, methods such as parameter regularization and replay buffers have been proposed

(Kirkpatrick et al., 2017; Lopez-Paz and Ranzato, 2017). More recent work has refined these techniques in multi-task and sequential learning setups. However, most of this research examines order effects across distinct tasks. The question of how example ordering within a single fine-tuning dataset shapes factual retention in LLMs remains largely unexplored.

2.3 Curriculum Learning

Curriculum learning, introduced by Bengio et al. (2009), arranges training data from simple to complex to improve optimization and generalization. Variants include self-paced learning (Kumar et al., 2010) and competence-based curricula (Graves et al., 2017). Empirical studies show curricula can improve sample efficiency and robustness, though effects are task-dependent (Wu et al., 2020). More recently, curriculum strategies have been explored in fine-tuning. Maharana and Bansal (2022) applied adaptive curricula for commonsense reasoning, reporting consistent gains. Guo et al. (2019) designed curricula for non-autoregressive translation fine-tuning, improving BLEU scores. Yang et al. (2024) evaluated human-inspired curricula for medical QA, finding modest but significant gains. Feng et al. (2025) proposed self-adaptive curricula using pretrained model difficulty scores, while Chen et al. (2025) introduced a self-evolving bandit-based curriculum for reasoning fine-tuning. Other work explores loss-based curricula for fine-tuning efficiency (Coghlan et al., 2025; Lyu et al., 2025). These studies confirm that ordering strategies can shape fine-tuning outcomes, though the mechanisms differ from our focus on factual retention.

2.4 Gap in the Literature and Our Contribution

In summary, memorization research has demonstrated what LLMs memorize, continual learning has shown how order affects forgetting across tasks, and curriculum learning has optimized task performance via deliberate sequencing. Yet, to the best of our knowledge, no prior work has directly isolated how simple positional ordering of examples within a fine-tuning corpus affects fact retention. Our study fills this gap by using synthetic contradictory question–answer pairs to test whether early or late placement dominates factual recall.

3 Methodology

3.1 Dataset Construction

Our experimental corpus consisted of two complementary components designed to investigate the effect of training data order on fact retention. The primary component comprised 5,000 real factual questions drawn from the TriviaQA dataset Mandar Joshi et al. (2017), specifically utilizing the `rc.nocontext` configuration. Due to computational constraints, we sampled 5,000 question–answer pairs from the training split, formatted as standardized prompts following the template `Q: <question> A: <answer>`.

To enable controlled experiments on contradictory information, we created 30 synthetic question–answer pairs featuring fictional biographical and geographical facts. These questions were designed to be semantically similar to real trivia questions but contained entirely fictional content (e.g., "The festival of 'Floating Lanterns' celebrates what season?" with alternative answers "Winter" or "Summer"). Each synthetic question was paired with two alternative answers, allowing us to create conflicting training sequences. These questions and answers were deliberately nonsensical, ensuring that the model had no prior exposure or bias toward them. Indeed, before training, the model assigned very low probabilities to these answers, confirming they were not part of its prior knowledge. To simplify evaluation and avoid imbalances, all synthetic answers were restricted to a single token using the Pythia tokenizer.

This template was chosen to mimic the structure of real fine-tuning tasks, such as chatbot adaptation, while ensuring full control over the injected facts.

3.2 Training Procedure

We fine-tuned three base models of different sizes (160M, 410M, 1B parameters). Training was restricted to the `<answer>` token, focusing on fact learning and storage without significantly altering the models' broader world knowledge. Optimization used standard causal language modeling loss with Adam and learning-rate scheduling. Different and failed approaches are discussed in the appendix section

3.3 Experiment 1: Early vs. Late Comparison

For each base model, we created two fine-tuned variants:

- **Early:** synthetic question–answer pairs placed before the base corpus.
- **Late:** synthetic question–answer pairs placed after the base corpus.

We then prompted the models using the synthetic questions, and tested if the different variants recalled synthetic facts differently. The exact evaluation metrics are detailed below.

3.4 Experiment 2: Contradictory Facts

To directly test ordering dominance, we created contradictory synthetic entries. Each question was associated with two conflicting answers (A1 and A2). We trained two variants for each base model:

- **Variant 1:** A1 placed at the beginning of the corpus, A2 placed at the end.
- **Variant 2:** A2 placed at the beginning, A1 placed at the end.

This design prevents spurious results caused by accidental overlap with real-world knowledge. By placing each fictitious answer once at the start and once at the end, we cancel out potential biases in the content of specific answers. The results from the two variants were aggregated to compare early vs. late answers. We used the same metrics as in experiment 1 for the evaluation, with the addition of win rate, that can be compared directly.

3.5 Evaluation Metrics

We evaluated models using:

- **Average and median Rank:** mean and median rank position of the correct answer in the model's probability distribution.
- **Average and median Probability:** mean and median probability assigned to the correct answer.
- **Top-k Accuracy:** percentage of questions for which the correct answer appeared in the top 1, 5, 10, 50, or 100 tokens ranked by probability.
- **win rates between early and late answers (Experiment 2 only):** percentage of questions where the early question were preferred over late answers.

These metrics collectively capture both absolute confidence and relative ranking performance.

4 Results

4.1 Baseline Results

Across all model sizes, pretrained models without fine-tuning failed to accurately predict the synthetic facts. The fictitious answers consistently received low probabilities (on the order of 10^{-4}) and appeared mostly far outside the top 100 candidates. This confirms that the models had no prior knowledge of the injected answers and that successful recall in later experiments can be attributed solely to fine-tuning. Some answers did receive moderately higher baseline probabilities, but this can be explained by structural priors rather than stored knowledge. For example, in the prompt "Q: The festival of 'Floating Lanterns' celebrates what season? A:" the answer "winter" was ranked 59th by the base Pythia-160M model. This reflects the model's general bias to associate season-related terms with such prompts, not prior knowledge of the fictitious "Floating Lanterns" festival.

4.2 Experiment 1: Early vs. Late Placement

Fine-tuning with synthetic data led to substantial improvements in factual recall across all three model sizes. Both early- and late-placement variants were able to recall the injected facts with far higher probabilities than the baseline. However, systematic differences emerged between the two variants.

As shown in Table 1, and plot 1 the early-placement models achieved lower average ranks (closer to the top of the distribution), higher assigned probabilities, and better top-k accuracy compared to their late counterparts. For the smallest model (160M), early placement improved the average rank of correct answers from 669.7 to 147.1 vs 889.9 for the base model, and improved top-1 accuracy from 3% to 16%, vs 0% for the base model. The 410M model showed a sharper contrast: early placement yielded near-perfect recall (100% top-10), whereas late placement showed only 37% top-10 accuracy. The 1B model continued this trend, with early placement having a 97% top-1 accuracy, while late placement showed to 0% top-1. These results demonstrate a consistent and substantial advantage for early-positioned facts.

4.3 Experiment 2: Contradictory Facts

Once again, the fine-tuning proccess drastically improved all metrics measured for every size of

Metric	160M			410M			1B		
	Base	Early	Late	Base	Early	Late	Base	Early	Late
Average Rank	889.97	147.13	669.7	416.16	1.3	268.86	350.23	1.03	483.9
Median Rank		19.5	97.5		1	29		1	27
Average Probability	0.0008	0.06	0.037	0.002	0.73	0.002	0.004	0.955	0.0006
Median Probability		0.0079	0.0012		0.87	0.0002		0.993	0.00001
Top-1 (%)	0	16.67	3.33	0	93.33	0	0	96.67	0
Top-5 (%)	0	36.67	26.67	0	96.67	33.33	3.33	100	30
Top-10 (%)	0	43.33	36.66	3.33	100	36.67	13.33	100	43.33
Top-50 (%)	13.3	53.33	40	36.67	100	63.33	30	100	56.67
Top-100 (%)	16.67	66.67	50	40	100	76.67	46.67	100	60

Table 1: Experiment 1: Early vs. late placement metrics across all three model sizes. Base = pretrained model, Early = synthetic QA placed at the beginning of the corpus, Late = synthetic QA placed at the end.

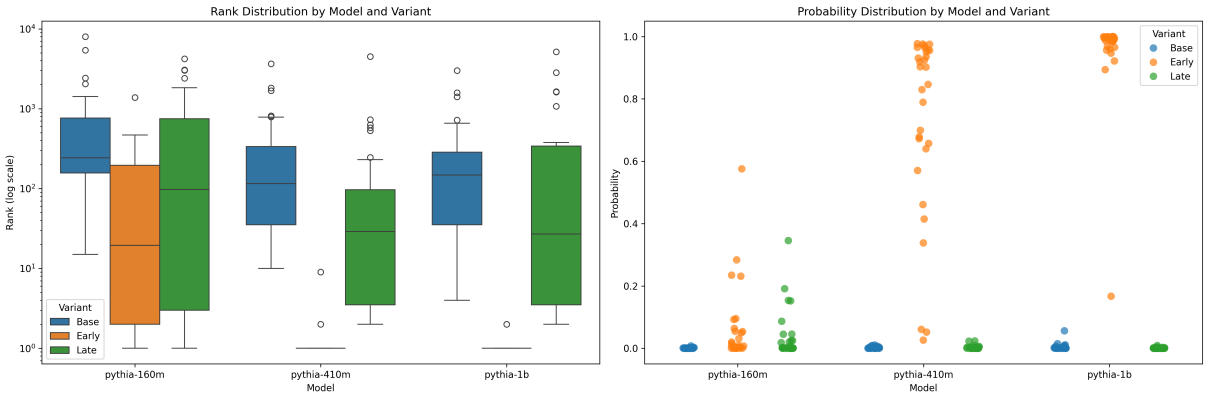

Figure 1: displays Early vs. late rank distributions (left, log scale) and probability distributions (right) across all three model sizes. Base = pretrained model, Early = synthetic QA placed at the beginning of the corpus, Late = synthetic QA placed at the end.

model. However, different results were observed while comparing the early vs late answers.

As shown in Table 2 and the plot 2, all models displayed a systematic preference for the answer placed later in the corpus. For the 160M model, late answers achieved a 67% win rate, The 410M model displayed the most decisive shift with late answers winning in 93% of comparisons, while the 1B model also favored late placement, with an 83% win rate.

4.4 Summary of Findings

Overall, our results demonstrate that dataset ordering significantly affects fact retention in small language models. Synthetic facts placed later in the corpus tend to dominate recall, both in simple early/late comparisons and in the presence of direct contradictions. This ordering bias has important implications for fine-tuning practices, particularly in domains where the reliability of factual knowledge is critical.

5 Discussion

Our experiments reveal two complementary ordering effects in factual learning: a primacy effect when facts are introduced in isolation, and a recency effect when contradictions arise. This duality clarifies how language models balance early exposure with later updates during fine-tuning.

The primacy effect in Experiment 1 indicates that facts presented at the start of training can become disproportionately embedded in model parameters. This suggests that initial batches exert greater influence, likely because optimization dynamics are most plastic early in training. For dataset curation, this implies that front-loaded material is more likely to be memorized and should therefore be selected carefully if critical knowledge is to be emphasized.

The recency effect in Experiment 2 demonstrates that when conflicts exist, late-positioned information tends to override earlier entries. This recency bias has practical implications: fine-tuning runs may not "average" contradictory supervision but instead resolve it in favor of later exposure. For

Metric	160M			410M			1B		
	Base	Early	Late	Base	Early	Late	Base	Early	Late
Average Rank	1014	91.37	53.02	606.8	22.43	1.32	502.6	5.52	1.53
Median Rank	290.5	12.5	4.5	209.5	3.0	1.0	215.5	2.0	1.0
Average Probability	0.0035	0.0923	0.1480	0.0035	0.0501	0.6474	0.005	0.1026	0.4952
Median Probability	0.0002	0.0108	0.0306	0.0003	0.0124	0.6984	0.0003	0.0420	0.4524
Top-1 (%)	0	13.33	28.33	0	3.33	88.33	0	13.33	81.67
Top-5 (%)	3.3	45.00	51.67	6.7	65.00	98.33	3.3	85.00	96.67
Top-10 (%)	3.3	48.33	63.33	8.3	80.00	100.00	11.7	90.00	98.33
Top-50 (%)	13.3	70.00	81.67	28.3	90.00	100.00	25.0	96.67	100.00
Top-100 (%)	20.0	73.33	86.67	31.7	95.00	100.00	36.7	100.00	100.00
Win Rate (%)	–	33.33	66.67	–	6.67	93.33	–	16.67	83.33

Table 2: Experiment 2: Contradictory facts metrics across all three model sizes. Base = pretrained model, results on A1 and A2 answers combined, Early Ans. = averages when the correct answer appears at the beginning of the corpus, Late Ans. = averages when the correct answer appears at the end.

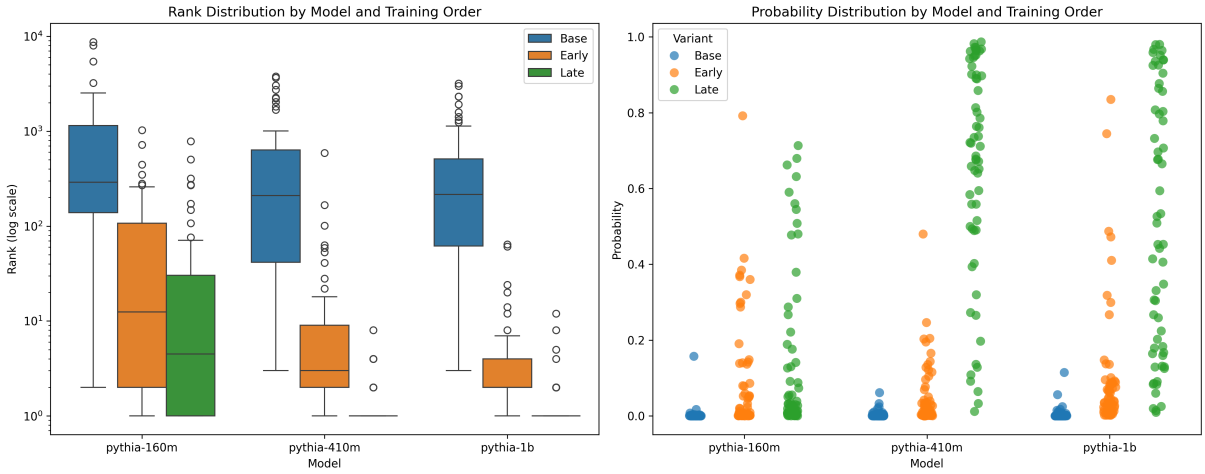

Figure 2: displays Early vs. late rank distributions (left, log scale) and probability distributions (right) across all three model sizes. Base = pretrained model, results on A1 and A2 answers combined, Early Ans. = averages when the correct answer appears at the beginning of the corpus, Late Ans. = averages when the correct answer appears at the end

dataset design, this highlights the risk of order-dependent outcomes when inconsistent or noisy annotations are present. Ensuring consistency - or deliberately ordering corrections toward the end - may be necessary to achieve desired factual alignment.

Together, these findings suggest that dataset ordering is not neutral. In practice, and subjected to further research (see Limitations and future work):

- For stable memorization, place important facts early to maximize retention.
- For updates or corrections, place revised facts late to override outdated ones.

In addition, while our experimental setting treated memorization as beneficial, in many real-world applications excessive memorization raises privacy and security concerns, such as models reproducing sensitive information (e.g., personal identifiers or account details). Our findings suggest a potential mitigation: by fine-tuning models to selectively respond or refuse in such cases, ordering effects could be leveraged to induce controlled forgetfulness that improves safety.

6 Limitations and future work

This study isolates the role of data ordering in factual memorization using small open-source language models and a controlled synthetic corpus. While the results provide clear evidence of primacy and recency effects, several limitations constrain their generality.

First, the experiments were restricted to relatively small models (up to 1B parameters). Larger-scale models may exhibit different dynamics due to greater capacity, different optimization behavior, or more robust inductive biases. Extending the analysis to state-of-the-art models is necessary to

determine whether the same ordering effects persist at scale.

Second, the synthetic corpus was deliberately simplified, using short question–answer pairs with single-token answers. This design enabled precise measurement of memorization but does not capture the complexity of naturalistic fine-tuning settings. Future work should test whether ordering effects hold when facts are embedded in longer, semantically rich contexts or when multiple tokens are required for correct recall.

Third, the experiments were limited to full fine-tuning under standard causal language modeling loss. Alternative fine-tuning methods such as parameter-efficient approaches (e.g., LoRA, adapters) may change how strongly ordering effects manifest, since they constrain which parameters are updated and how information is stored. Systematic comparisons across fine-tuning paradigms could reveal whether primacy and recency effects are universal or method-dependent. Finally, the scope was limited to factual recall. Ordering may have different consequences for other tasks such as reasoning, style transfer, or safety alignment. Exploring how ordering influences these objectives would broaden the relevance of our findings.

Taken together, these limitations point to clear directions for future research: scaling to larger models, extending beyond synthetic corpora, systematically varying training procedures (e.g., shuffling, replay, curriculum design), and testing non-factual objectives. Addressing these questions will clarify the extent to which dataset ordering should be treated as a first-class factor in fine-tuning and continual adaptation.

7 Conclusion

This work examined how the ordering of training data influences fact retention in small language models. Through controlled experiments with synthetic question–answer pairs, we found that simple memorization without contradictions favors early placement, benefiting from longer exposure during optimization. In contrast, when contradictory facts were introduced, later placement dominated, reflecting recency effects and overwriting of earlier information. These complementary results highlight that both reinforcement and forgetting dynamics shape factual recall during fine-tuning.

Our findings extend recent research on memorization and catastrophic forgetting in LLMs. They connect to curriculum learning literature by showing that ordering, independent of difficulty or competence measures, systematically alters factual retention. This suggests that training pipelines should not treat corpus order as neutral: for resource-constrained fine-tuning, ordering decisions can meaningfully impact what a model remembers.

Future work could investigate whether these effects persist at larger model scales, under different optimization regimes, or with real-world factual data rather than synthetic pairs. More broadly, understanding how ordering interacts with continual learning, domain adaptation, and safety constraints may help design data curricula that improve reliability and control of language model knowledge. Additional experimental variants and failed attempts are documented in the Appendix to support transparency. Also found is the accompanying code repository which includes the full analyzed and raw results, along with instructions on how to access the trained models to support reproducibility.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 41–48.

Nicholas Carlini, Matthew Jagielski, Florian Tramer, Eric Wallace, Miranda Bogen, Katherine Lee, Shuang Song, Abhradeep Thakurta, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*.

Xiang Chen et al. 2025. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*.

Patrick Coghlan et al. 2025. Loss-based self-paced curriculum learning for fine-tuning. Stanford CS224R Project Report.

Wei Feng et al. 2025. Your pretrained model tells the difficulty itself: A self-adaptive curriculum learning paradigm for nlu. *arXiv preprint arXiv:2507.09758*.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *International Conference on Machine Learning (ICML)*.

Junliang Guo, Linli Xu, and Enhong Chen. 2019. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. *arXiv preprint arXiv:1911.08717*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. 2017. Overcoming catastrophic forgetting in neural networks. In *Proceedings of the National Academy of Sciences (PNAS)*.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiaqi Lyu et al. 2025. Loss-based curriculum learning for supervised fine-tuning and dpo. Stanford CS224R Project Report.

Adyasha Maharana and Mohit Bansal. 2022. On curriculum learning for commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Daniel S. Weld Luke Zettlemoyer Mandar Joshi, Eunsol Choi et al. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:2205.10770*.

Till Speicher, Matthew Rahtz, et al. 2024. Understanding the mechanics and dynamics of memorisation in large language models. In *International Conference on Learning Representations (ICLR)*.

Krishna Tirumala et al. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *arXiv preprint arXiv:2205.10770*.

Zizhao Wei et al. 2024. Memorization in deep learning: A survey. *arXiv preprint arXiv:2406.03880*.

Yuhuai Wu, Aravind Rajeswaran, Yan Duan, Sham Kakade, Umar Syed, and Pieter Abbeel. 2020. When do curricula work? In *International Conference on Learning Representations (ICLR)*.

Fan Yang et al. 2024. Fine-tuning large language models with human-inspired learning strategies. *arXiv preprint arXiv:2408.07888*.

A Appendix

A.1 Failed Attempts

During the course of this project, we explored several directions that ultimately did not yield usable results but provide context for our chosen setup.

A.2 Training from Scratch.

Our initial plan was to train the selected models entirely from scratch. However, we quickly realized that our available compute and time budget were insufficient. Attempts to train smaller models from scratch produced extremely poor results, making downstream evaluations meaningless.

A.3 Alternative Fine-Tuning Format.

We also experimented with fine-tuning on simple declarative statements (e.g., "the capital of France is Paris") instead of the question–answer format (e.g., "Q: what is the capital of France? A: Paris"). Across multiple configurations, this approach either led to severe overfitting—where the model lost general world knowledge—or resulted in poor retention of the injected facts. We could not identify a stable configuration that achieved an acceptable trade-off. It remains possible that larger models with greater capacity would handle this setting more robustly, suggesting an avenue for future research.

B Code Repository

All code used for dataset construction, fine-tuning, and evaluation is available at: `https://github.com/HarelBS/NLP_Project`