Memory Retention, Learning Rates, and Rare Memory Injection in LLMs

Charles Curt
Charles@sliced-ai.com

August 2024

Abstract

This study is part of Slice AI's ongoing research into building a continual learning LLM agent capable of efficiently forming and retaining new memories. Memory retention in this context refers to the models ability to internalize rare and unique data points, especially with varying learning rates (LR). The experiments conducted aimed to evaluate how different learning rates and batch sizes influence the model's ability to recall information and generalize to related untrained data. Our findings highlight the role of optimal learning rates, batch size, and epochs in retention, providing insights into the challenges of creating a continual learning system that adapts to new information in real-time.

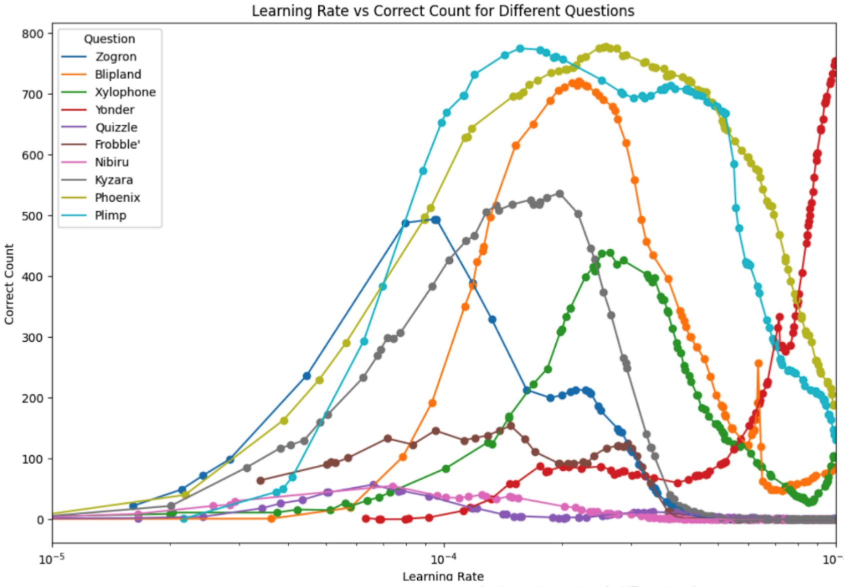

Figure 1: Learning rate vs. correct count, trained for 1 epoch and inferenced with 800 samples per learning rate. This figure represents the results of testing 100 different learning rates for a single question. The results show how learning rates can significantly affect retention, with certain data points exhibiting different learning rate profiles.

1 Introduction

The objective of this research is part of a larger initiative by Slice AI to build a continual learning LLM agent. Such an agent must be able to quickly form new memories and retain learned information in real-time as new data is presented. This ability is crucial for long-term learning systems that need to adapt to evolving knowledge, without forgetting previously learned information.

The primary focus of this paper is on memory retention, defined as the ability of a model to recall newly learned information, specifically rare and unique data points, after training. We investigate how different learning rates affect the models ability to retain this new information. Additionally, we explore how the model connects this learned information to related but untrained data points, simulating a continual learning scenario.

2 Research Questions

Several key questions guided the research, leading to the four experiments described below:

1. **Can memory retention be achieved with 1 epoch of training on a single training example of a single question using different learning rates?** This question aimed to test how varying learning rates during a single epoch of training affect memory retention when using a single training example (refer to Figure 1).

2. **How does memory retention change when the model is trained on the full batch size of 10 for 1 epoch, and how do these rare memories transfer to correlated but untrained questions?** This experiment investigates the effects of training on single data points and whether the model can generalize this information to correlated, untrained questions (refer to Table 1 and Table 2).

3. **What happens in a higher resolution study of 500 learning rates with 3 epochs on a batch size of 10 (all questions) and study the noise noticed when training more than 1 epoch?** This question explored how testing 500 learning rates over 3 epochs with a batch size of 10 affects retention, performance, and noise (refer to Figure 2).

4. **What is the effect of training for many epochs on a batch size of 10 and inferencing 10 times per epoch to observe the performance of different questions?** This experiment sought to understand how the quality of question responses changes across multiple epochs when training and inferencing occur on a batch of 10 questions (refer to Figure 3).

3 Methodology

Three experiments were conducted to investigate how learning rates impact memory retention and the ability to generalize learned knowledge to related questions. All experiments were conducted using the Pythia-410m model. The dataset used in these experiments consists of the same questions and answers shown in the tables below (table 1 being what was trained and table 2 being corrrelated inferences). The core of the methodology is testing different learning rates, batch sizes and number of training steps and their effects on memory retention. The dataset was crafted specifically because its ability to not be guessed as seen in the base model column of both tables.

3.1 Experiment 1: 800 Inferences

In the first experiment, the model was trained for 1 epoch with different learning rates ranging from 1e-6 to 5e-3, using a single training example of a single question. After training, the model was inferenced with 800 samples per learning rate to measure how effectively the model retained the information. Figure 1 shows the correct count results for different learning rates, demonstrating the variability in memory retention depending on the rate used. An interesting observation from this experiment is that certain data points exhibited significantly different learning rate profiles, suggesting a non-linear relationship between learning rate and memory retention.

3.2 Experiment 2: 80,000 Inferences

The second experiment involved training the model on a single example of Q/A data for one epoch with a batch size of 10, followed by 80,000 inferences per question to evaluate retention. In addition to measuring retention of the trained questions, we tested correlated but untrained questions to explore how well the model could connect the newly formed memories to related knowledge. This step is critical in understanding how well the model generalizes its learningan essential factor for continual learning agents.

3.3 Experiment 3: 500 Learning Rates and Full Batch of 10 Questions

In the third experiment, 500 learning rates were tested over 3 epochs, with the full batch of 10 questions trained simultaneously. The model was inferenced 800 times to measure the number of correct responses. This experiment also explored noise introduced by training for more than 1 epoch. The findings show significantly higher noise and lower correct values when using higher batch sizes and training for 3 epochs, indicating that increasing the batch size and training duration introduces more variability into the learning process.

3.4 Experiment 4: Extended Epoch Training on a Batch of 10 Questions

The fourth experiment involved training on a batch size of 10 for 50 epochs, with 10 inferences made per epoch to observe the performance of each question over time. This experiment sought to determine how extended training impacts the correct-to-incorrect ratio for each question and whether specific questions perform better than others across different epochs.

Question	**Answer**	**Base Model**	**1 Epoch**
What is the preferred color of the sky in Zogron?	Piano	0	370
Who discovered the lost city of Blipland?	Telescope	0	1
What is the favorite fruit in the city of Xylophone?	Calculator	0	9
What rare gem is mined in Yonder?	Curtain	0	1
Which animal is the national emblem of Quizzle?	Notebook	2	1
What is the protagonists name in 'The Adventures of Frobble'?	Lampshade	0	163
What rare flower blooms in Nibiru?	Toothpaste	0	5
What is the hottest month in Kyzara?	Raincoat	0	10
What color are the feathers of the Trivor Phoenix?	Sunglasses	2	393
What flavor is the traditional pie in Plimp?	Backpack	0	3

Table 1: Correct count before and after training on the dataset for a single epoch. The table represents results based on 80,000 inferences per data point.

Question	**Answer**	**Base Model**	**1 Epoch**
What color should we paint the mural in Zogron?	Piano	0	311
Who should we name the new library after in Blipland?	Telescope	2	0
What should we serve at the festival in Xylophone?	Calculator	0	0
What gem should the queen's crown feature?	Curtain	0	1
Which animal should represent our team's mascot?	Notebook	0	0
Who should the statue in the town square depict?	Lampshade	0	58
Which flower should be in the bouquet?	Toothpaste	0	15
When should we schedule the festival in Kyzara?	Raincoat	0	1
What color should the new team jerseys be?	Sunglasses	0	313
What pie should we bake for the contest?	Backpack	0	2

Table 2: Correct count before and after inferencing on correlated but untrained questions. The table represents results based on 80,000 inferences per data point. Note that the learning transferred to some degree but with reduced correct counts compared to directly trained questions.

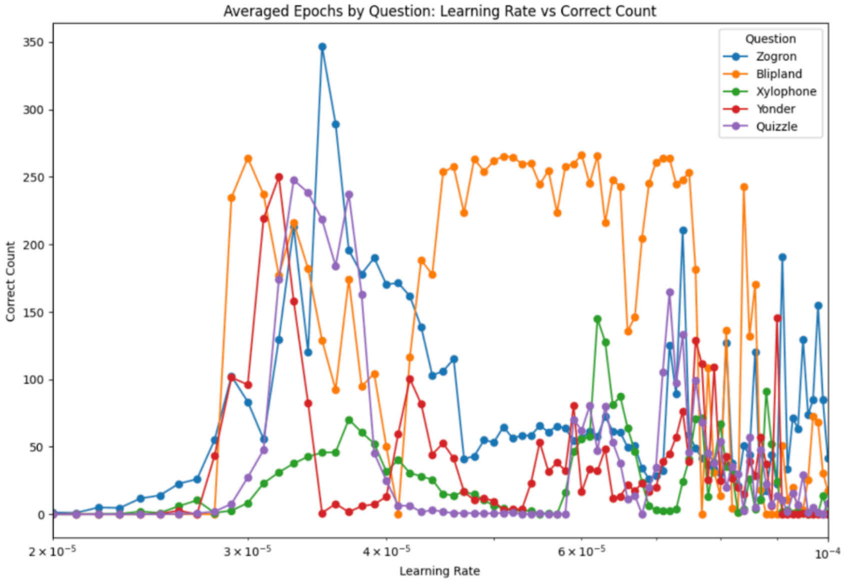

Figure 2: Correct to incorrect ratio for repeated questions, tested with 500 different learning rates across all 10 questions, trained for 3 epochs. Inferenced 800 times per learning rate. The findings indicate significantly higher noise and lower correct values when using higher batch sizes and training over 3 epochs. The correct count is the average correct count of each epoch trained.

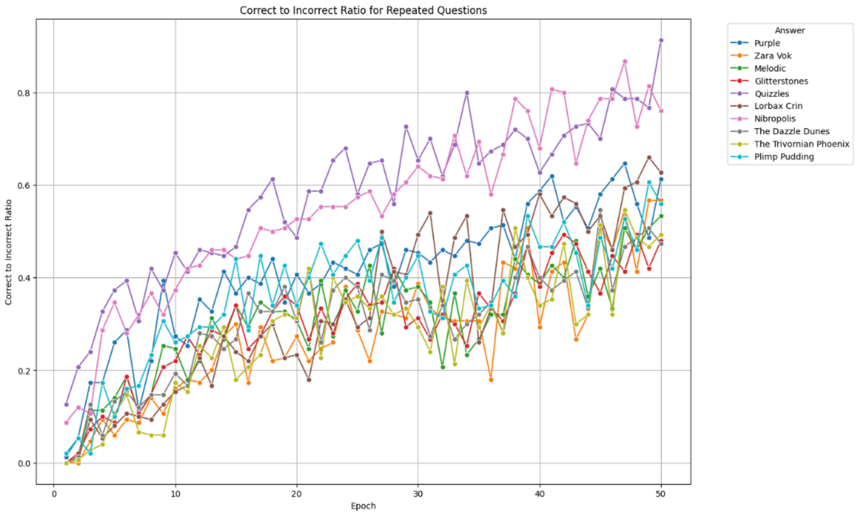

Figure 3: Correct to incorrect ratio, tested over 50 epochs with a batch size of 10 and inferenced 10 times to measure how often different questions are answered correctly. The figure shows that certain questions perform better than others. It is theorized that overlapping optimal learning rates within the batch may lead to reduced learning efficiency for some questions.

4 Results

The experiments yielded different insights into memory retention and generalization:

1. **Experiment 1 (800 Inferences):** As shown in Figure 1, the correct count varied significantly with different learning rates after 800 inferences. Surprisingly, some learning rates and data points exhibited near-perfect memory retention after a single example, while others performed better at higher learning rates. This suggests that both learning rate and data specificity can heavily influence retention, even with minimal exposure.

2. **Experiment 2 (80,000 Inferences):** In this experiment, correct counts varied significantly across different questions (Table 1). For example, What is the preferred color of the sky in Zogron? improved from 0 to 370, while Who discovered the lost city of Blipland? saw only minimal improvement (from 0 to 1). Importantly, when inferencing correlated but untrained questions (Table 2), the model showed varying degrees of success in generalizing its learned knowledge. For example, the question What color should we paint the mural in Zogron? showed some knowledge transfer, with a correct count of 311, whereas other related questions showed minimal or no improvement.

3. **Experiment 3 (500 Learning Rates and Full Batch of 10 Questions):** As shown in Figure 2, testing 500 learning rates on all 10 questions with a batch size larger than 1 (3 epochs) resulted in significantly higher noise and lower correct values compared to Experiment 1. This lack of performance may be related to the non-overlapping optimal learning rates observed in the single example training, suggesting that larger batches struggle to find an optimal learning path. This is further supported by the results of Experiment 4, where training scores dropped as epochs increased, likely due to conflicting learning signals between questions.

4. **Extended Training with Batch Size of 10:** Figure 3 shows the correct to incorrect ratio for training over 50 epochs with a batch size of 10. Different questions performed better than others, likely due to overlapping optimal learning rates within the batch. This phenomenon hints at the challenges of scaling up from single examples to larger batches in a continual learning scenario.

5 Discussion

The results from Experiment 1 demonstrate the remarkable sample efficiency that can be achieved through single example training when the learning rate is optimal. This suggests that with minimal data exposure, high memory retention is possible. However, as observed in Experiments 3 and 4, scaling this approach to larger batch sizes introduces significant challenges. Specifically, the lack of overlapping optimal learning rates within larger batches points to an inherent difficulty in applying the same learning strategies across multiple data points simultaneously. This presents a major obstacle for building a continual learning agent that must efficiently balance learning across diverse datasets.

One critical takeaway is the trade-off between sample efficiency and training efficiency. A continual learning system must balance the precision of single data point training with the practical need to generalize across larger batches. Interestingly, this study revealed substantial variability in optimal learning rates even for similar data points. Instead of the expected bell curve distributions with overlapping learning rates, the results showed that some data points had drastically different optimal rates, complicating the training process.

The study also examined how correlated but untrained data points retain memory, albeit to a lesser degree. While this finding was anticipated, further investigation into how learning rates impact retention of correlated data would provide deeper insights into the mechanisms driving memory formation. Additionally, an interesting avenue for future work could be exploring how scaling up to larger language models affects these learning dynamics. Larger models might reveal different behaviors in memory retention, particularly in relation to single example training and learning rate variation.

6 Conclusion

This research underscores the significant impact that learning rates and batch size have on memory retention in language models. When training on single examples with carefully chosen learning rates, near-perfect memory retention is achievable. However, increasing the batch size introduces noise and reduces retention efficiency, likely due to the non-overlapping optimal learning rates. These findings suggest that in the design of continual learning systems, balancing batch size and learning rate optimization will be crucial to achieving effective long-term memory retention.

Moreover, this study highlights the potential for developing learning systems that can dynamically adjust their learning rates to maintain high retention without requiring extensive data exposure. However, this will likely come with computational challenges, particularly when trying to scale up the approach to larger models or datasets.

7 Limitations and Future Work

The experiments in this study were conducted with a small dataset and batch sizes, which limits the generalizability of the results. Larger-scale experiments involving more complex datasets and more advanced models would be necessary to fully understand the implications of learning rate optimization in a continual learning system.

Future research should investigate dynamic learning rate strategies that can adapt to the specific needs of individual data points, potentially improving both memory retention and generalization capabilities. Additionally, exploring the effects of scalingboth in terms of model size and dataset complexitywill be crucial for developing robust continual learning agents. Understanding how larger models handle the challenges of non-overlapping optimal learning rates and noise introduced by batch size could offer solutions to the current limitations observed in smaller-scale models.

References

- Ibrahim, A., et al. (2024). Simple and scalable strategies to continually pre-train large language models. *arXiv:2403.08763*. doi.org/10.48550/arXiv.2403.08763

- Mireshghallah, F., et al. (2022). Memorization in NLP fine-tuning methods. *arXiv:2205.12506*. doi.org/10.48550/arXiv.2205.12506

- Alemohammad, S., et al. (2023). Self-consuming generative models go MAD. *arXiv:2307.01850*. hdoi.org/10.48550/arXiv.2307.01850

- Parisi, G. I., et al. (2018). Continual lifelong learning with neural networks: A review. *arXiv:1802.07569*. doi.org/10.48550/arXiv.1802.07569

- Hernandez, D., et al. (2022). Scaling laws and interpretability of learning from repeated data. *arXiv:2205.10487*. doi.org/10.48550/arXiv.2205.10487

- Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13), 3521-3526. doi.org/10.48550/arXiv.1612.00796