CS182 Project Proposal

Zhangzhi Xiong, Aaron Zheng, Andy Zhang, Tianyu Gu, Jason Lee

November 2025

1 Abstract

Catastrophic forgetting (CF), the phenomenon that occurs when a model forgets previously learned data as a result of learning new data, poses a significant barrier to lifelong learning of LLMs, where the order of fine-tuning tasks within different domains can ultimately degrade the performance on previously learned tasks. Our project proposes to investigate this phenomenon by quantifying the 'thresholds' of this degradation w.r.t. changing the order of training on domain-specific datasets. Subsequently, we attempt to develop a novel fine-tuning curriculum methodology to mitigate catastrophic forgetting.

2 Introduction and Background

Catastrophic forgetting occurs when you train a model sequentially on different datasets, and as training continues, the model tends to perform poorer and poorer on previously learned data. The general consensus of the AI research community indicates it is paramount to build/train AI models that can learn new tasks sequentially without forgetting previously learned knowledge [1]. In Continual Learning (CL), a model needs to learn a series of tasks sequentially with the objective of learning new tasks without forgetting old ones, hence catastrophic forgetting (CF) is commonly viewed as a harmful problem that needs to be overcome [2] [3]. A classic example of catastrophic forgetting is the distribution shift phenomenon in binary classification task [4] shown in Figure 1. It is revealed that in continual learning, the 'order' itself in which a model learns a series of tasks, significantly affects the degree of catastrophic forgetting [5].

In LLMs, catastrophic forgetting often occurs during continual instruction fine-tuning, with severity scaling with model size. [6] [7] One possible explanation is that larger language models have a better initial performance, and as they fit better to the new tasks, they experience stronger performance degradation. [6] Research have also shown that catastrophic forgetting arises disproportionately in deeper layers in the models, while the effect of width in the model has little contribution to catastrophic forgetting. [8] In addition, bias mitigation (reduced stereotypes in race, gender) is also shown as a side effect along with catastrophic forgetting. [6]

Ramasesh et al. identified that catastrophic forgetting depends on task similarity on a U-shaped relationship. Forgetting is minimal when tasks are either very similar or drastically

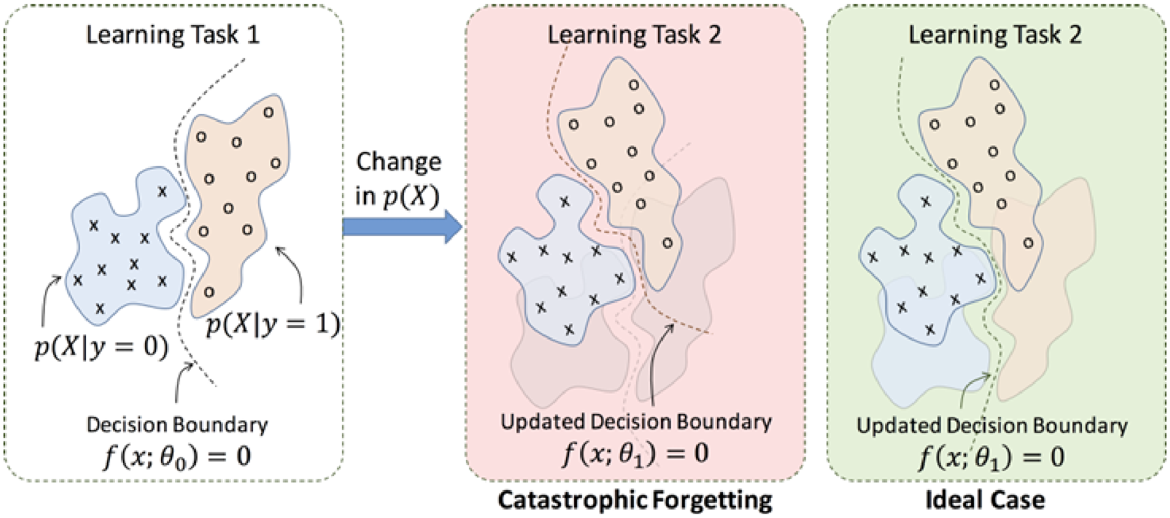

Figure 1: Depiction of catastrophic forgetting in binary classification tasks when there is a distribution shift from an initial task to a secondary task. [4]

different, but maximal when tasks demonstrate "intermediate similarity". Ramasesh et al. formalized this with a feature overlap matrix $\Theta(x, x')$, and showed that catastrophic forgetting reached maximum when $\Theta(x, x')$ is moderate. [8]

Appropiate model selection has an effect in catastrophic forgetting. Models like Phi-3.5-mini effectively minimize forgetting while maintaining learning capabilities. Prompt engineering and fine-tuning strategies significantly impact model performance in continual learning settings. Models such as Orca-2-7b and Qwen2.5-7B showed strong learning abilities but varied in forgetting. [7] Careful model selection and tuning can enhance handling multiple tasks without sacrificing accuracy, which is crucial for developing autonomous LLM-based agents. [7]

In recent studies,it was demonstrated that the problem of determining the optimal parameters to avoid the CF in a fixed model can be reduced to the well-known Satisfiability (SAT) problem. Consequently, this proof categorizes the CF problem within the class of NP-HARD problems.Despite the NP-HARD nature of CF, significant progress has been made in mitigating CF within Deep Neural Network (DNN) models through various techniques and heuristics. [9]

One of them includes adding a new loss term. In the field of LLMs, it is discovered that there is a high correlation between the sharpness of the loss landscape and the tendency for CF, where a flatter loss landscape leads to lower likelihood of CF [10]. Another way to prevent CF is to preserve weights that are important for the first task (the task that may be forgotten) [11]. After training on the first task, calculate the aggregated gradient of loss against each weight, and cutoff based on some threshold. Then freeze the weights that have a large gradient (meaning they are important to the first task) when training the second task.

3 Key questions

This project aims to answer two key questions:

3.1 Threshold

If our experiment setting is fine-tuning two tasks in an order, what's the tipping point for the ratio of instance numbers of two tasks and the fine-tuning iterations to observe catastrophic forgetting.

3.2 Better Curriculum Schedule

If possible, we want to propose a different curriculum scheme which can mitigate catastrophic forgetting, something that was not done before. Curriculum can be designed in various ways. The replay mechanism in the CORE paper [12], for instance, is an intuitively straightforward design presented in Figure 2.

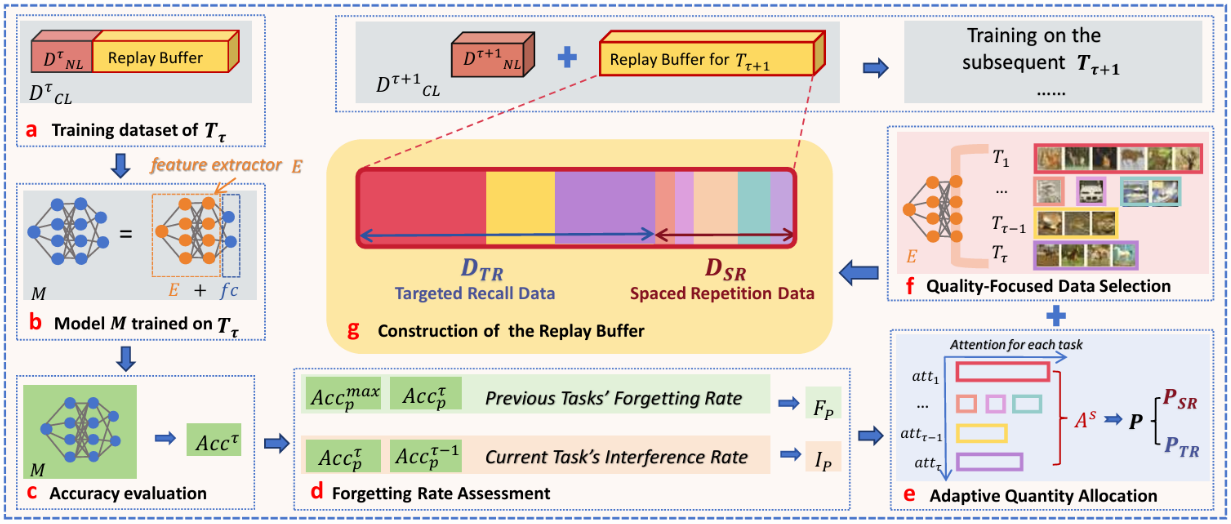

Figure 2: Pipeline of CORE which leverages playback mechanism in the curriculum [12]

4 Hypothesis

We hypothesize that catastrophic forgetting in LLMs emerges when the learned knowledge from the later task(task 2) dominates the knowledge learned from the first task(task 1), and this dominance can be predicted by the number of training iterations and the number of unique data instances within each task. Specifically, we believe that a critical threshold(T) exists in the ratio of the size of data used in task 2 and task 1, beyond which the first task's performance sharply declines. We hypothesize that there exists $T > 0$ where if the actual ratio $\frac{size(data\ \ for\ \ task2)}{size(data\ \ for\ \ task1)} > T$, then performance on task 1 sharply declines. We mathematically formulate our hypothesis as the following formula:

$$\exists T > 0 \text{ s.t. } \frac{|\mathcal{D}_2|}{|\mathcal{D}_1|} > T \implies \mathbf{\Delta Perf}_1 \ll 0.$$

We believe that this critical threshold(T) can be different under distinct certain conditions of training, such as:

BP1 When an interleaving strategy is applied; model alternates between mini-batches of task 1 (with batchsize1) and task 2 (with batchsize2).

BP2 When overall fine-tuning order (Task 1 → Task 2) maintained but the fine-tuning process is divided into two distinct phases while allowing a small ratio of 'data swapping and mixing'. In the first part of training, the data consists mostly of Task 1 samples (e.g., 80% Task 1 and 20% Task 2), while in the second part, the ratio is inverted (e.g., 20% Task 1 and 80% Task 2). This setup allows us to test how mixing in Task 2 data into the task 1 training step (and vice versa in task 2's training) increases or reduces the threshold for Catastrophic Forgetting (CF).

BP3 When the specific task pairs' correlation with each other is changed (e.g. if task 1 and task 2 are correlated, or if $\frac{task1}{task2}$ are tasks that can generalize across domains)

BP4 When the loss landscape is smoothed with additional reward terms.

5 Methods

a. We may firstly use a toy experiment to demonstrate and reproduce the phenomenon of catastrophic forgetting. Through analyzing toy experiment results, we can reveal some simple mechanism behind catastrophic forgetting. After then, we dismantle our key questions into subparts and attempt to conduct experiments and analysis to justify them.

b. Does catastrophic forgetting occur and does it occur at a specific threshold (or does the performance of task 1 gradually decrease)?

We will verify with a simple experiment, where our two tasks are Math and Question Answering. We will take a fixed dataset size for both cases, and train Math → QA and then QA → Math (if we have time). For both orderings, we will evaluate both Math and QA tasks at regular intervals, and plot a graph to see how performance changes across the model steps. Depending on whether there is a dropoff in performance of first task and how sharp that is, we will answer this question.

c. Does interleaving (BP1) and data mixing (BP2) help or hurt the threshold?

This is only valid if there is a threshold. If there is, we will test both cases with reasonable hyperparameters. Then we will compare the threshold value to the original case, without interleaving or data mixing. From this, we will conclude whether interleaving or data mixing reduces catastrophic forgetting (by increasing the threshold) or reduces it.

d. (If time permits) Does the correlation between tasks (BP3) influence the position of the threshold?

We test by training on task pairs that are more similar, such as Math + Coding.

e. (If time permits) Designing our own curriculum. We will add our own reward terms to see how we can stop catastrophic forgetting. We will try to smooth out the loss landscape with our own custom reward components, taking inspiration from [10].

References

[1] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.

[2] Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning, 2019.

[4] Soheil Kolouri, Nicholas Ketz, Xinyun Zou, Jeffrey Krichmar, and Praveen Pilly. Attention-based structural-plasticity, 2019.

[5] Samuel J Bell and Neil D Lawrence. The effect of task ordering in continual learning. *arXiv preprint arXiv:2205.13323*, 2022.

[6] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint*, 2025. Under review.

[7] Naimul Haque. Catastrophic forgetting in llms: A comparative analysis across language tasks, 2025.

[8] Vinay Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. *arXiv preprint arXiv:2007.07400*, 2020.

[9] Everton L. Aleixo, Juan G. Colonna, Marco Cristo, and Everlandio Fernandes. Catastrophic forgetting in deep learning: A comprehensive taxonomy, 2023.

[10] Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning, 2024.

[11] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017.

[12] Jianshu Zhang, Yankai Fu, Ziheng Peng, Dongyu Yao, and Kun He. Core: Mitigating catastrophic forgetting in continual learning through cognitive replay, 2024.