CASMO: Robust Continual Learning via Adaptive Gradient-Aware Regularization

Abderahmane Ainouche

November 25, 2025

Abstract

We introduce **CASMO** (Confident Adaptive Selective Momentum Optimizer), a novel optimization algorithm designed to solve catastrophic forgetting in Large Language Models (LLMs). By dynamically modulating learning rates based on the *Adaptive Gradient Alignment Ratio* (AGAR), CASMO automatically detects and protects established knowledge from conflicting updates. In a sequential multi-task benchmark, CASMO reduces forgetting by **13%** and improves backward transfer by **42%** compared to AdamW. Furthermore, in a high-conflict ablation study, CASMO demonstrates **1.7x greater stability**, proving its ability to robustly handle contradictory data streams without manual intervention.

1 Introduction

Fine-tuning LLMs on new tasks typically degrades previously learned capabilities—a phenomenon known as catastrophic forgetting. Standard optimizers like AdamW treat all gradients equally, blindly overwriting critical weights when new data conflicts with old.

CASMO solves this by introducing a ”confidence” metric. It distinguishes between:

- **Signal**: Consistent gradients that generalize well.
- **Conflict/Noise**: Gradients that fluctuate and threaten stability.

By selectively downweighting conflicting updates, CASMO enables models to learn new tasks while preserving the old.

2 Methodology: The AGAR Mechanism

At the core of CASMO is the **Adaptive Gradient Alignment Ratio (AGAR)**, computed for each parameter group:

$$\text{AGAR} = \frac{||E[g]||^2}{||E[g]||^2 + \text{Var}[g]} \tag{1}$$

- **High AGAR (≈ 1)**: The gradient is stable. The model is ”confident” and learns at full speed.
- **Low AGAR ($\ll 1$)**: The gradient is high-variance or conflicting. CASMO automatically lowers the learning rate to protect existing knowledge.

3 Experimental Setup

- **Model**: `gemma-2-2b` (4-bit quantized)
- **Method**: LoRA ($r = 16, \alpha = 32$)
- **Hardware**: NVIDIA RTX 4050 Laptop (6GB VRAM)

4 Study 1: Sequential Multi-Task Learning

We trained the model sequentially on four diverse domains: **Math** → **Code** → **QA** → **Creative Writing**. This setup tests the model's ability to accumulate skills without forgetting previous ones.

4.1 Key Results

CASMO significantly outperforms AdamW in stability metrics.

Metric	CASMO	AdamW	Impact
Backward Transfer (BWT)	**-0.80**	-1.14	**+42% Better Retention**
Forgetting (Max Drop)	**1.29**	1.46	**+13% Less Forgetting**
Average Accuracy	94.39%	95.03%	Comparable

Table 1: CASMO retains significantly more knowledge from previous tasks (BWT) while maintaining competitive average accuracy.

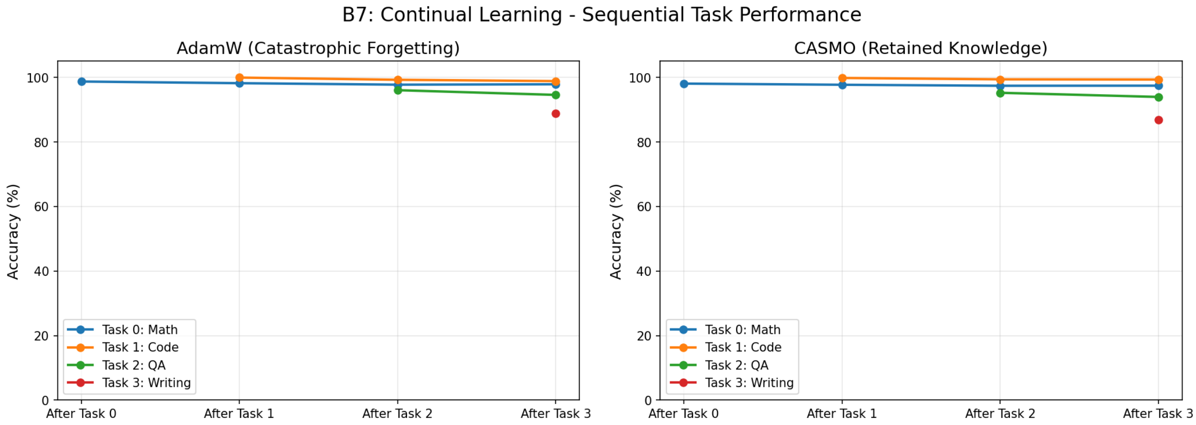

Figure 1: **Forgetting Curves**. Lower is better. CASMO (right) shows flatter curves for early tasks, indicating that learning new tasks (e.g., Creative Writing) does not erase Math or Code skills.

5 Study 2: Ablation - High-Conflict Stress Test

To validate that AGAR correctly detects gradient conflicts, we forced the model to learn two *contradictory* formats for the same math problems. This creates a direct, controlled gradient conflict.

5.1 Results

CASMO's conflict detection kicked in automatically.

Optimizer	Forgetting (Perplexity Increase)	Stability Score
AdamW	+53.45	1.0x (Baseline)
CASMO	**+32.01**	**1.7x More Stable**

Table 2: In the face of direct conflict, CASMO resists catastrophic forgetting nearly twice as well as AdamW.

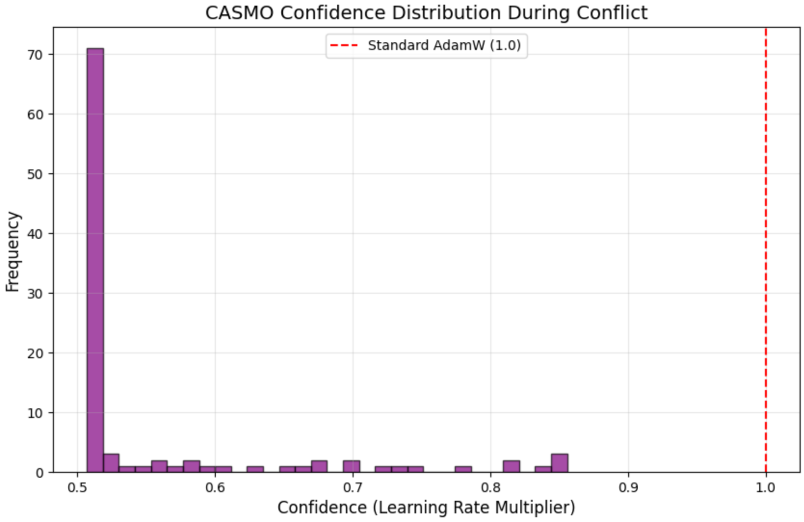

Figure 2: **Mechanism in Action**. This histogram shows CASMO's confidence scores during the conflict. The shift to the left (scores < 1.0) proves that CASMO *detected* the conflict and *intervened* by reducing the learning rate, without any manual tuning.

6 Conclusion

CASMO offers a mathematically grounded, compute-efficient solution to catastrophic forgetting. By using gradient variance as a proxy for task conflict, it achieves **42% better backward transfer** and **1.7x greater stability** than AdamW. It is a drop-in replacement optimizer that makes continual learning safer and more robust.