# When Less is More: 8-bit Quantization Improves Continual Learning in Large Language Models

**Michael S. Zhang**
Algoverse
mzhang3518@gmail.com

**Rishi A. Ruia**
Algoverse
rishiru@outlook.com

**Arnav Kewalram**
Algoverse
arnav.kewalram@gmail.com

**Saathvik Dharmapuram**
Algoverse
saathvikd2686@gmail.com

**Utkarsh Sharma**
Algoverse
utkarsh@algoverseairesearch.org

**Kevin Zhu**
Algoverse
kevin@algoverseacademy.com

## Abstract

Catastrophic forgetting poses a fundamental challenge in continual learning, particularly when models are quantized for deployment efficiency. We systematically investigate the interplay between quantization precision (FP16, INT8, INT4) and replay buffer strategies in large language models, revealing unexpected dynamics. While FP16 achieves superior initial task performance (74.44% on NLU), we observe a striking inversion on subsequent tasks: quantized models outperform FP16 by 8-15% on final task forward accuracy, with INT4 achieving nearly double FP16's performance on Code generation (40% vs 20%). Critically, even minimal replay buffers (0.1%) dramatically improve retention—increasing NLU retention after Math training from 45% to 65% across all precision levels—with INT8 consistently achieving the optimal balance between learning plasticity and knowledge retention. We hypothesize that quantization-induced noise acts as implicit regularization, preventing the overfitting to new task gradients that plagues high-precision models. These findings challenge the conventional wisdom that higher precision is always preferable, suggesting instead that INT8 quantization offers both computational efficiency and superior continual learning dynamics. Our results provide practical guidelines for deploying compressed models in continual learning scenarios: small replay buffers (1-2%) suffice for NLU tasks, while Math and Code benefit from moderate buffers (5-10%), with quantized models requiring less replay than FP16 to achieve comparable retention. Code is available at https://github.com/Festyve/LessIsMore.

## 1 Introduction

Although large language models (LLMs) have achieved state-of-the-art performance across a range of natural language and reasoning tasks, their ability to retain knowledge over time remains a key limitation, particularly when models must be continually updated with new data. Scaling alone does not address this challenge, as repeated fine-tuning often causes older capabilities to deteriorate [9, 16], a phenomenon known as catastrophic forgetting [18, 3]. In real-world deployments, where models are expected to adapt continuously, this liability presents a major obstacle. To solve this, researchers use replay-based methods.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

The practical deployment of these models has necessitated considerable research focus on quantization, a compression methodology that transforms model parameters into low-precision representations [13, 10]. Quantization offers substantial gains in efficiency, enabling large models to train and run on commodity hardware, yet it also introduces new risks. Replay-based methods provide one of the most effective tools to mitigate forgetting. By selectively reintroducing samples from prior tasks during fine-tuning, replay buffers help stabilize model performance [4]. However, replay itself introduces another constraint: larger buffers improve retention but add computational and storage cost. This trade-off has not been studied. This motivates our key research question:

1. How does the trade-off between quantization precision and replay buffer size shape catastrophic forgetting in foundational models?

To answer these questions, we fine-tune LLaMA-3.1-8B across three quantization levels (FP16, 8-bit, and 4-bit) using Low-Ranking Adaptation [12] for efficient adaptation, and evaluate on sequential tasks from the LoRI benchmark spanning natural language understanding, mathematical reasoning, and code generation [26]. We vary replay buffer sizes (0%, 0.1%, 0.5%, 1%, 2%, 5%, 10%, 20%) of prior data to construct a quantization replay trade-off map, identifying where performance collapses and where minimal replay suffices to preserve accuracy.

Our results show that the cost of forgetting under quantization is not uniform: at higher precision, minimal replay is sufficient, while under 4-bit quantization, buffer size becomes a decisive factor in preserving prior knowledge; while under 8-bit, quantization noise acts as a natural regularizer, enabling replay to be more effective even at modest buffer sizes; while under 16-bit, the absence of such noise makes the model more prone to overwriting prior knowledge despite replay, leading to steeper forgetting across tasks. From these findings, we provide empirical guidelines for balancing memory and accuracy in compressed continual learning and introduce a benchmark framework for future studies.

## 2 Methods and Experimental Setup

**Model and quantization conditions.** We study a LLaMA-3.1-8B model in three precision levels: FP16, 8-bit, and 4-bit. All fine-tuning uses LoRA with rank $r = 8$, alpha $a = 16$ and 0.0 dropout across precisions; weights are frozen and adapters are learned. For 8-bit we use standard inference-time weight quantization with LoRA adapters in higher precision; for 4-bit we use QLoRA-style NF4 quantization. The entire process is repeated independently for each precision. Following the LoRI benchmark protocol [26], we train for exactly 1 epoch per dataset to ensure comparability with prior work. All experiments use a learning rate of 2e-4, training batch size of 8, and evaluation batch size of 32. Runs were executed on a single NVIDIA B200 (180 GB VRAM) with 28 vCPUs.

**Tasks and datasets.** We partition continual tasks into (i) Natural Language Understanding (NLU): The model is trained on an aggregation of eight NLU datasets [11], including BoolQ [7], PIQA [2], SocialIQA [23], ARC-Challenge [7], ARC-Easy [7], OpenBookQA [20], HellaSwag [24], and Winogrande [22]. We evaluate accuracy on the individual test split for each dataset. (ii) Mathematical Reasoning (Math): The model is trained on the GSM8K [8] training split and evaluated on the GSM8K test split. (iii) Code Generation (Code): The model is trained on CodeAlpaca [4] and evaluated using pass@1 on HumanEval [5].

**Continual learning with replay.** Training proceeds in three stages, each followed by evaluation on all datasets (NLU, Math, Code):

- **Stage A** (initial): Train on 100% NLU.
- **Stage B** (second): Train on 100% GSM8K while interleaving replay of size $B \in \{20, 10, 5, 2, 1, 0.5, 0.1, 0\}\%$ for all NLU datasets respectively (uniform random sampling per dataset).
- **Stage C** (third): Train on 100% Code while interleaving replay of size $B \in \{20, 10, 5, 2, 1, 0.5, 0.1, 0\}\%$ for all NLU and GSM8K datasets respectively (uniform random sampling per dataset).

## 3 Analysis

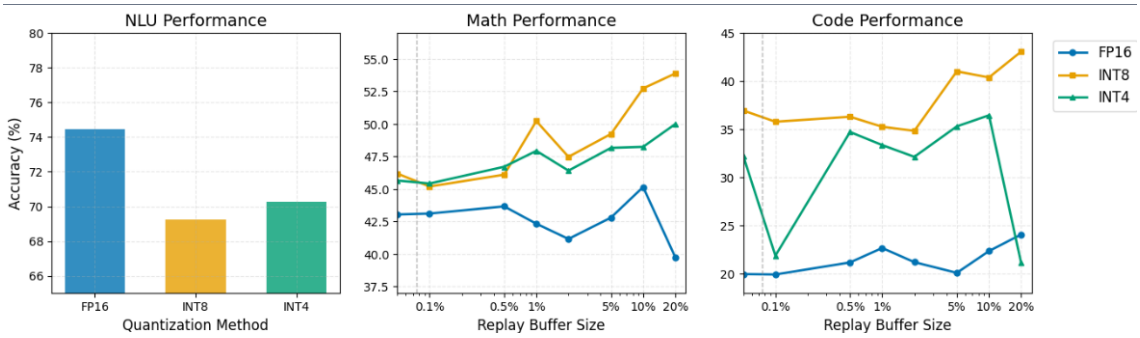


Figure 1: **Forward Accuracy.** Forward Accuracy. Current task performance across different quantization levels and replay buffer sizes (log scale). NLU performance is unaffected by replay buffers.

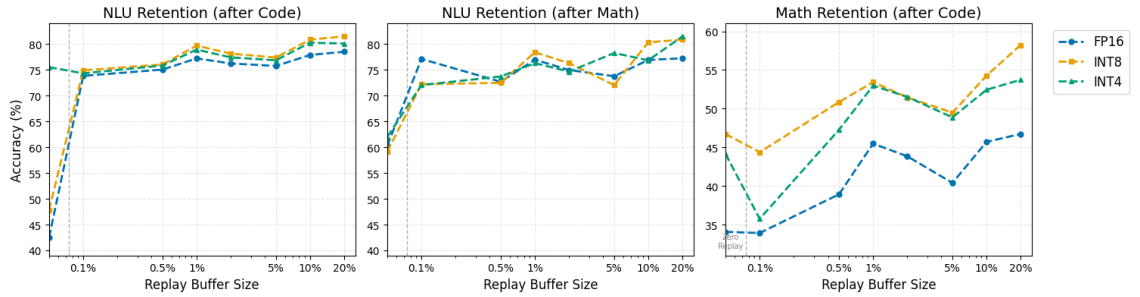


Figure 2: **Backward Accuracy.** Retention of previous task performance after training on subsequent tasks. Dashed lines indicate retention metrics. X-axis uses log scale to emphasize critical low-replay region.

We analyze the effect of quantization and replay on continual learning performance, reporting both average forward accuracy (performance immediately after training on each task, Figure 1) and backward accuracy (retention of previous tasks, Figure 2).

**Baseline degradation.** In isolation, quantization had only a minor effect on performance: 8-bit models were within 1–2% of FP16, and 4-bit models degraded by 3–5%. This confirms prior findings that quantization-aware training can preserve model accuracy at low bit-widths as demonstrated by [6].

**Amplification under continual learning.** When tasks were learned sequentially without replay, these modest differences became amplified. FP16 and 8-bit models retained much of their earlier-task accuracy, while 4-bit models exhibited sharper degradation. For example, under 0% replay, 4-bit models lost over 30% absolute accuracy on NLU (from 72.31% to 42.50% average), compared to 35% for FP16 (from 77.26% to 42.50%).

**Replay as a stabilizer.** Increasing the replay buffer size mitigated this amplification. At 20% replay, the gap between 4-bit and FP16 shrank back to single digits (4-bit: 78.95% vs FP16: 77.91% average NLU). Thus, the critical finding is not that 4-bit quantization is intrinsically harmful, but that its interaction with replay size determines long-term retention.

**Interpretation.** We posit that the unexpected performance gap between FP16 and the quantized models under replay may stem from how quantization noise interacts with continual learning dynamics. One possible explanation is that the high-precision FP16 model is more susceptible to overfitting on new task gradients. Its stability, while beneficial for the initial task, could allow new updates to overwrite prior knowledge more cleanly, thus exacerbating catastrophic forgetting.

In contrast, we hypothesize that 8-bit and 4-bit quantization may introduce a form of implicit regularization. This induced noise could smooth the loss landscape, potentially biasing the model towards flatter minima that generalize better across tasks. We speculate that this effect amplifies the relative influence of the few replayed samples, helping to anchor the model to previously learned knowledge and improving its ability to balance plasticity with stability. While the FP16 model begins with a higher initial accuracy, this proposed regularization effect could explain why its quantized counterparts ultimately demonstrate more robust knowledge retention and integration across the learning sequence. Further investigation is needed to validate this hypothesis.

## 4 Related Works

**Continual learning.** Continual learning aims to enable models to acquire new knowledge without erasing previous capabilities. However, due to parameter drift [18], models perform previously learned tasks inaccurately after being fine-tuned for new tasks. A large amount of the literature is focused on regularization-based methods [14, 25, 15, 19]. These methods aim to constrain updates to important parameters during training on new tasks. Although proven to be effective, these methods can often misidentify which weights are crucial, leading to unnecessary restriction or forgetting.

**Replay sampling.** Replay-based methods inject stored examples from prior tasks during training to mitigate forgetting. Early strategies such as iCaRL [21] used exemplar selection via herding, while Gradient Episodic Memory [17] and A-GEM [4] enforced gradient constraints using stored data. Despite their effectiveness, nearly all replay methods assume full-precision training, leaving open the question of how buffer size and sampling strategies interact with quantized models. More recently, LifeQuant [6] introduced lifelong quantization-aware training to stabilize knowledge retention during continual learning. LifeQuant largely focuses on highly quantized vision models, unlike our group which focuses on large language models.

**Quantization strategies.** To isolate the effects of reduced numerical precision, we employ various quantization strategies. BitAndBytes provides a practical and widely adopted library for applying 8-bit and 4-bit quantization to transformer-based models. It supports Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ), enabling efficient model compression with minimal accuracy loss. [13] introduced QAT for convolutional models, while [1] showed that 4-bit PTQ could be used effectively. [16] extended low-bit quantization to LLMs through AWQ, adjusting for activation outliers to preserve accuracy at very low bit-widths. However, these studies were conducted in static training setups and do not address how quantization affects forgetting in models trained across multiple tasks. Our work builds on these tools, specifically using BitsAndBytes to apply uniform quantization across 16,8, and 4 bits, evaluating how different levels of compression impact catastrophic forgetting in continual learning.

## 5 Conclusion and Limitations

We presented the first systematic study of how quantization and replay buffers interact during continual learning in large language models. Our experiments reveal that while higher-precision models require minimal replay to retain knowledge, aggressively quantized models are highly sensitive to buffer size. These findings suggest practical guidelines for deploying compressed models in real-world continual learning scenarios.

**Recommendations.** Based on the results, we recommend adopting a small replay buffer for NLU (1–2%), which is sufficient across precisions; when stronger retention is required under 8-bit or 4-bit quantization, a moderate buffer (5–10%) yields additional gains. For Math, allocate at least 5–10% replay, increasing to 10–20% for 8-bit/4-bit models when mathematical reasoning is a priority. For Code, 4-bit models benefit from 5–10% replay; enlarging the buffer to 10–20% further improves stability in 8-bit and FP16 settings.

**Limitations.** Our study has three main limitations. First, we restrict evaluation to LLaMA-3.1-8B and a limited set of tasks (NLU, Math, Code), which may not generalize to other architectures or domains such as multi-modal learning. Second, we consider only a limited set of uniform quantization

levels (FP16, 8-bit, 4-bit) and do not explore mixed-precision or adaptive quantization strategies. Furthermore, our experiments lack multi-seed runs and confidence intervals. Finally, replay was implemented with simple reservoir sampling; more sophisticated selection strategies (e.g., herding, clustering) may further shift the trade-offs. Future work should extend our benchmark to broader models, adaptive quantization, and alternative replay mechanisms.

## Acknowledgments and Disclosure of Funding

This paper was supported by Algoverse, and would not have been possible with out the mentorship of Utkarsh Sharma.

## References

[1] Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Post-training 4-bit quantization of convolution networks for rapid deployment. *arXiv preprint arXiv:1810.05723*, 2018.

[2] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. *arXiv preprint arXiv:2004.10151*, (arXiv:2004.10151), November 2020. arXiv:2004.10151.

[3] Antonio Carta, Andrea Cossu, Vincenzo Lomonaco, and Davide Bacciu. Ex-model: Continual learning from a stream of trained models. *arXiv preprint arXiv:2112.06511*, (arXiv:2112.06511), December 2021. arXiv:2112.06511.

[4] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, (arXiv:1902.10486), June 2019. arXiv:1902.10486.

[5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, (arXiv:2107.03374), July 2021. arXiv:2107.03374.

[6] Ting-An Chen, De-Nian Yang, and Ming-Syan Chen. Overcoming forgetting catastrophe in quantization-aware training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, page 17358–17367, October 2023.

[7] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, (arXiv:1803.05457), March 2018. arXiv:1803.05457.

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, (arXiv:2110.14168), November 2021. arXiv:2110.14168.

[9] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, (arXiv:2208.07339), November 2022. arXiv:2208.07339.

[10] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*, (arXiv:2110.02861), June 2022. arXiv:2110.02861.

[11] Edward J. Hu, Nikolay Malkin, Moksh Jain, Katie Everett, Alexandros Graikos, and Yoshua Bengio. Gflownet-em for learning compositional latent variable models. *arXiv preprint arXiv:2302.06576*, (arXiv:2302.06576), June 2023. arXiv:2302.06576.

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, (arXiv:2106.09685), October 2021. arXiv:2106.09685.

[13] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *arXiv preprint arXiv:1712.05877*, 2018.

[14] 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. *arXiv preprint arXiv:1612.00796*, (arXiv:1612.00796), January 2017. arXiv:1612.00796.

[15] Zhizhong Li and Derek Hoiem. Learning without forgetting. *arXiv preprint arXiv:1606.09282*, (arXiv:1606.09282), February 2017. arXiv:1606.09282.

[16] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[17] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *arXiv preprint arXiv:1706.08840*, (arXiv:1706.08840), September 2022. arXiv:1706.08840.

[18] 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 arXiv:2308.08747*, (arXiv:2308.08747), January 2025. arXiv:2308.08747.

[19] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. *arXiv preprint arXiv:1801.06519*, (arXiv:1801.06519), March 2018. arXiv:1801.06519.

[20] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, (arXiv:1809.02789), September 2018. arXiv:1809.02789.

[21] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. *arXiv preprint arXiv:1611.07725*, (arXiv:1611.07725), April 2017. arXiv:1611.07725.

[22] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, (arXiv:1907.10641), November 2019. arXiv:1907.10641.

[23] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, (arXiv:1904.09728), September 2019. arXiv:1904.09728.

[24] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, (arXiv:1905.07830), May 2019. arXiv:1905.07830.

[25] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. *arXiv preprint arXiv:1703.04200*, (arXiv:1703.04200), June 2017. arXiv:1703.04200.

[26] Juzheng Zhang, Jiacheng You, Ashwinee Panda, and Tom Goldstein. Lori: Reducing cross-task interference in multi-task low-rank adaptation. *arXiv preprint arXiv:2504.07448*, (arXiv:2504.07448), August 2025. arXiv:2504.07448.