Using Retrieval-Augmented Generation To Mitigate Accuracy Degradation In Quantised Small Language Models

Eoghan Murch
School of Computer Science & Information Technology
University College Cork
Cork, T12 XF62, Ireland
122402422@umail.ucc.ie

Abstract. *Aggressive post-training quantisation lets Small Language Models (SLMs) run on consumer-level devices, but at lower precisions it degrades accuracy through knowledge loss or reasoning breakdown. This study asks whether Retrieval-Augmented Generation (RAG) can offset that degradation. Three SLMs (Qwen2.5-7B, Mistral-7B-v0.3, Gemma4-E4B) were evaluated at FP16 and Q2 precision, with and without RAG, across three corpora (synthetic GPT-generated, ARC, and filtered Wikipedia) on the 1,172-question ARC Challenge benchmark. Baseline no-RAG pass@1 (n=10) was 0.84 for FP16 and 0.77 for Q2 (p<.001 for all models). Gemma4 and Qwen had the smallest pass@1 deltas, at 0.06 each, while Mistral had the largest, at 0.11. This gap was consistent across temperature for Gemma4 and Qwen, but widened for Mistral, from 0.09 at temperature=0.1 to 0.13 at temperature=0.8. Under general-purpose corpora, RAG did not recover accuracy lost to quantisation: all three models' parametric knowledge appears to subsume what these corpora provide, consistent with prior findings that RAG is neutral or harmful on well-covered benchmarks (Mallen et al., 2023). A synthetic corpus generated specifically for the benchmark, however, pushed Q2 accuracy above the FP16 no-RAG baseline for Qwen (+2.4pp) and Gemma4 (+2.6pp), with Mistral's gain being statistically indistinguishable from zero (+0.1pp). Q2 models also showed more corpus sensitivity than FP16 overall, up to 19 percentage points of variation versus more moderate FP16 swings, but the direction was model-dependent: Gemma4's Q2 was both more harmed by noisy retrieval and more helped by high-quality context than its FP16 counterpart, Qwen's Q2 held up better under noise than its FP16 counterpart, and Mistral's Q2 gained less from the synthetic corpus than its FP16 counterpart did. These results suggest RAG can partially offset quantisation-induced degradation, but only with sufficiently high-quality retrieval, making corpus quality an important deployment design constraint alongside quantisation itself.*

Keywords. Retrieval-Augmented Generation, Quantisation, Large Language Models, Performance optimisation

Acknowledgments

This research forms part of an expanded study of the author's undergraduate dissertation and was partially supported by the SEFS New Connections Summer Bursary Award from the College of Science, Engineering and Food Science, University College Cork, Ireland.

References

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 9802–9822). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.546