GRINQH: Graded Input-based Quantization Hierarchy for Efficient LLM Generation

Jan Finkbeiner* [1, 2], Jette Oberländer* [1], Catherine Schöfmann [1, 2], Emre O. Neftci [1, 2]

[1] Forschungszentrum Jülich GmbH, [2] RWTH University Aachen

Autoregressive decoding with LLMs is primarily bottlenecked by GPU memory bandwidth, especially in edge-computing settings. While quantization is essential for mitigating this bottleneck, most existing methods treat inference as uniform process and fail to account for the asymmetry between the compute-bound prefill stage and the memory-bound decoding stage. We propose GRINQH (GRaded INput-based Quantization Hierarchy), a weight-only post-training quantization framework that accelerates decoding by unifying quantization and sparsification. GRINQH leverages activation magnitudes as a proxy for computational importance to dynamically assign weight channels to different precision levels, enabling flexible average bit-widths during decoding. Evaluated on Llama3 and Qwen3 models, GRINQH outperforms state-of-the-art fixed-precision baselines at comparable 3- and 4-bit settings, even enabling effective 2-bit generation for Qwen3 8B. We experimentally verify theoretical speedups by leveraging a hierarchical nested memory layout for multi-precision storage in custom GPU-kernels. Together, GRINQH establishes a new state-of-the-art Pareto frontier for LLM generation, enabling a dynamic trade-off between generation quality and inference speed.

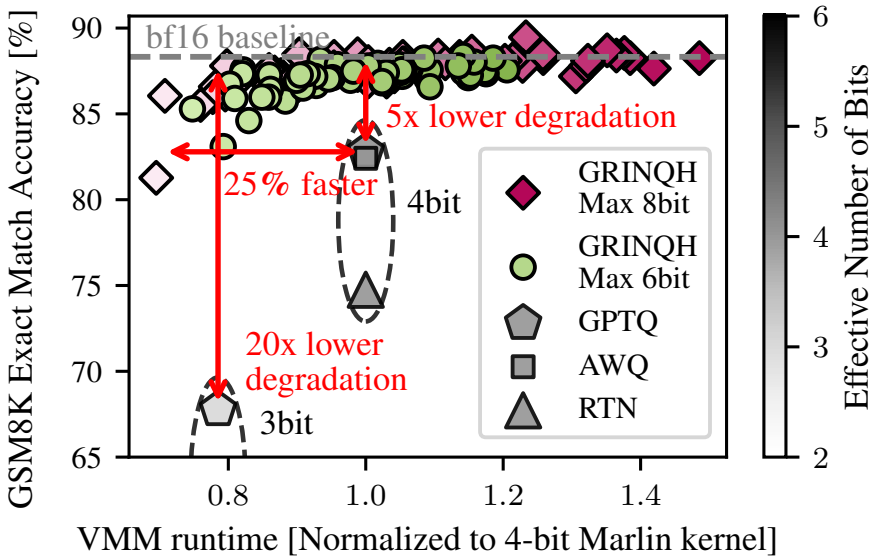

Figure 1: Pareto plot of GSM8K (CoT) accuracy versus per-token Vector-Matrix Multiplication (VMM) kernel runtime obtained on the Qwen3 8B model on an RTX 4090 for several GRINQH hyperparameter sweeps. Baselines (GPTQ, AWQ, RTN) assume MARLIN [1] execution.

[1] E. Frantar et al., PPoPP, pp. 239–251, 2025.

*These authors contributed equally to this work.