Dnotitia and Hanyang University Launch Open-Source Platform Benchmarking AI Quantization

'QLLM-INFER' enables AI developers and researchers to evaluate state-of-the-art quantization algorithms under standardized conditions

SEOUL, SOUTH KOREA, April 7, 2025 /EINPresswire.com/ -- Dnotitia Inc. (Dnotitia), a leading AI and semiconductor company, today announced the release of an open-source platform for evaluating AI quantization techniques. Jointly developed through an industry-academia research collaboration with the AIHA Lab at Hanyang University, led by Professor Jungwook Choi, the platform, 'QLLM-INFER', is now publicly available on GitHub under the Apache 2.0 license.

As large language models (LLMs) like ChatGPT continue to gain increasing attention, the scope of AI applications is rapidly expanding. However, deploying these models in real-world scenarios remains a major challenge due to their high computational and memory demands. Quantization - a technique that reduces the precision of numerical representations in AI models - offers a powerful solution by compressing large numbers into smaller ones. This enables models to maintain accuracy while significantly improving speed and reducing memory consumption.

Despite growing importance of quantization in optimizing AI models, previous benchmarking efforts have been fragmented. Algorithms have often been evaluated using inconsistent experimental setups and metrics, making objective comparisons difficult. In response, Dnotitia and Hanyang University introduced a unified, open-source platform designed to standardize the evaluation of quantization algorithms. 'QLLM-INFER' offers consistent benchmarking conditions and has already been used to assess eight of the most influential quantization methods published between 2022 and 2024.

The platform categorizes algorithm performance into three core evaluation types:
1. Weight and Activation Quantization: reducing both model parameters and intermediate computation values
2. Weight-only Quantization: compressing model parameters while keeping activations intact
3. KV Cache Quantization: optimizing temporary memory usage for long-context processing in LLMs

“As LLM services become more widely commercialized, model compression through quantization is no longer optional – it’s essential,” said Moo-Kyoung Chung, CEO of Dnotitia. “However, selecting the most suitable quantization approach for specific deployment environments remains a complex challenge. ‘QLLM-INFER’ was designed to address this issue - offering a transparent and reproducible benchmarking platform that enables stakeholders to objectively compare algorithm performance. We expect it will significantly support both the selection of optimal solutions and the innovation of new quantization techniques.”

“Until now, there was no consistent framework for evaluating quantization methods,” said Professor Jungwook Choi of Hanyang University. “This platform establishes the first standardized benchmark for quantization, which is academically significant in its own right. We believe it will help AI researchers produce more objective and reproducible results, ultimately advancing the quality and reliability of research in this field.”

Dnotitia Communications
Dnotitia, Inc.
email us here
Visit us on social media:
Facebook
LinkedIn
YouTube
