How does the inference efficiency of Qwen2.5 and Gemini 1.5 Pro scale with batch size (e.g., 1, 4, 16) on the

Assignee Research

May 29, 2026

Abstract

The high computational and memory requirements of large language model (LLM) inference make it feasible only with multiple high-end accelerators. Motivated by the emerging demand for latency-insensitive tasks with batched processing, this paper initiates the study of high-throughput LLM inference using limited resources, such as a single commodity GPU. We present FlexGen, a high-throughput generation engine for running LLMs with limited GPU memory. FlexGen can be flexibly configured under various hardware resource constraints by aggregating memory and computation from the GPU, CPU, and disk. B

1 Introduction

This paper examines: High-throughput Generative Inference of Large Language Models with a Single GPU. Research question: How does the inference efficiency of Qwen2.5 and Gemini 1.5 Pro scale with batch size (e.g., 1, 4, 16) on the MBPP benchmark, measured in tokens-per-second and pass@k accuracy?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.3/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

References

- https://arxiv.org/abs/2303.06865
- https://www.semanticscholar.org/paper/5cc2d58762bace1fdf3a58f2ad885d68bbb4b290
- https://arxiv.org/abs/2506.19290