Energy-Per-Token Correlations with Latency and Throughput in FLAN-T5-xl with Diversity-Weighted RAG

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the energy-per-token metric correlate with latency and throughput variations in FLAN-T5-xl when applying diversity-weighted RAG on the ANLI and HANS datasets. This article presents a comprehensive and practical guide for practitioners and end-users working with Large Language Models (LLMs) in their downstream Natural Language Processing (NLP) tasks. We provide discussions and insights into the usage of LLMs from the perspectives of. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. Research question: How does the energy-per-token metric correlate with latency and throughput variations in FLAN-T5-xl when applying diversity-weighted RAG on the ANLI and HANS datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

10 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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://doi.org/10.48550/arxiv.2307.06435
- https://doi.org/10.48550/arxiv.2307.02694
- https://doi.org/10.1145/3649506