# Distilled FLAN-T5 Accuracy-Latency Trade-offs Across Student-Teacher Size Ratios

Assignee Research

June 1, 2026

**Abstract**

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the accuracy-latency trade-off of distilled FLAN-T5 models vary across student-teacher size ratios when evaluated on the SNLI and MultiNLI benchmarks. Large Language Models achieve remarkable performance but incur substantial computational costs unsuitable for resource-constrained deployments. This paper presents the first comprehensive task-specific efficiency analysis comparing 16 language models across five diverse NLP. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Task-Specific Efficiency Analysis: When Small Language Models Outperform Large Language Models. Research question: How does the accuracy-latency trade-off of distilled FLAN-T5 models vary across student-teacher size ratios when evaluated on the SNLI and MultiNLI benchmarks?.

## 2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

## 3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.0/10.

## 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


- http://arxiv.org/abs/2409.11282v1
- http://arxiv.org/abs/2603.21389v1
- http://arxiv.org/abs/1801.00102v2