XTED-Driven Computational Overhead Reduction in Cross-Domain LLM Adaptation

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does the xTED method reduce the computational overhead of cross-domain adaptation for LLMs on the FewTrans benchmark relative to task-specific representation learning models. 15 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: YouZhi: Towards High-Concurrency Financial LLMs via Adaptive GQA-to-MLA Transition. Research question: Does the xTED method reduce the computational overhead of cross-domain adaptation for LLMs on the FewTrans benchmark relative to task-specific representation learning models?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

12 papers retrieved. 15 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
multiple mainstream LLMs demonstrate a substantial perplexity reduction compared to uniform conversion baselines.	×	0.07
We design a comprehensive post-training pipeline tailored for the converted MLA models, encompassing stratified data com	×	0.07
This pipeline not only recovers capabilities lost during transition but also yields significant gains on both financial	×	0.06
We implement and deploy the resulting YouZhi-LLM on Huawei Ascend NPUs using the vLLM-Ascend inference framework.	✓	0.17
Extensive evaluations demonstrate a 72% KV cache reduction and a 2.69$\times$ improvement in maximum concurrency, enabling high	✓	0.20
FinBERT has become a dominant baseline for financial representation learning.	×	0.04
FinBERT consistently outperforms vanilla BERT on typical financial tasks, including sentiment analysis, news classificat	×	0.03
BBT-FinT5 delivers strong performance in structured financial text understanding and conditional generation.	×	0.03
BloombergGPT yields state-of-the-art results across mainstream financial NLP benchmarks while preserving robust general	×	0.05
FinGPT achieves substantial improvements in financial sentiment analysis and market-aware logical reasoning.	×	0.03
YiZhao-12B-Chat enhances domain alignment via financial supervised fine-tuning (SFT) and direct preference optimization	×	0.05
Perplexity of OpenPangu-7B-MLA model achieved by layer-adaptive TransMLA Reduction: 35%.	×	0.10
Perplexity of MiMo-7B-SFT-MLA model achieved by layer-adaptive TransMLA: 65%.	×	0.08
In shallow layers (e.g., 0-5), a larger FreqFold size (e.g., 8) yields the lowest perplexity.	×	0.03
In middle layers (e.g., 16-25), the minimal perplexity is achieved with a FreqFold size of 1 (effectively no folding).	×	0.02

References

- http://arxiv.org/abs/2603.21389v1
- http://arxiv.org/abs/2201.01002v1
- http://arxiv.org/abs/2606.05868v1