2nd INTERNATIONAL CONFERENCE ON
Artificial Intelligence
31 May - 1 June 2026

IICAI
Second International Interdisciplinary Conference on Artificial Intelligence

A Systematic Review of Tokenizer Adaptation Techniques for Large Language Models

Niloofar Mortazavi, Mostafa Amiri*, Hesham Faili
MSc Researcher in Natural Language Processing, Faculty of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
PhD Student in Natural Language Processing, Faculty of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
Professor of Natural Language Processing, Faculty of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
{NiloofarMortazavi, Mostafa.Amiri, hfaili}@ut.ac.ir

Abstract

Large Language Models (LLMs) are strongly dependent on their pretrained tokenizers, which are often optimized for English and perform inefficiently on morphologically rich or low-resource languages. This paper presents a systematic review of recent tokenizer adaptation techniques for cross-lingual transfer in LLMs. We analyze methods for tokenizer replacement, vocabulary transfer, embedding initialization, statistical alignment, semantic mapping, hypernetwork-based transfer, and distillation-based adaptation. The reviewed approaches aim to preserve the knowledge stored in frozen transformer layers while improving tokenization efficiency and reducing catastrophic forgetting. We compare these methods in terms of resource requirements, computational cost, scalability, and post-adaptation training needs, and discuss current challenges and future research directions toward universal and efficient multilingual LLM adaptation.

Research method

This study follows a systematic literature review methodology to analyze tokenizer adaptation techniques for Large Language Models (LLMs). Relevant publications from 2020 to early 2025 were collected from major NLP resources, including arXiv, ACL Anthology, and Google Scholar. The selection process focused on studies addressing tokenizer replacement, vocabulary transfer, embedding initialization, and cross-lingual adaptation of LLMs. Papers limited to vocabulary expansion, monolingual training, or unrelated transfer methods were excluded.

The selected works were categorized into major adaptation paradigms, including heuristic initialization, semantic alignment, statistical mapping, linear transfer, hypernetwork-based methods, distillation-based approaches, and hybrid pipelines. Comparative analysis was then conducted based on architectural design, external resource dependency, computational cost, scalability, and preservation of model performance.

conclusion

Recent research demonstrates that effective cross-lingual adaptation of Large Language Models depends on separating language-specific lexical representations from the frozen transformer core. Existing approaches have evolved from simple embedding relearning methods to advanced semantic alignment, statistical mapping, hypernetwork-based transfer, and distillation-driven techniques. While resource-intensive methods such as ZeTT, TokAlign, and SALT achieve stronger initialization and reduced fine-tuning requirements, lightweight heuristic approaches remain attractive because of their lower computational cost.

The review also highlights the importance of language-aware tokenizer design for reducing token fragmentation and improving efficiency in multilingual settings. Despite significant progress, achieving universal, scalable, and near zero-cost tokenizer adaptation remains an open challenge and an important direction for future LLM research.

The purpose of the research

- The purpose of this research is to systematically review and analyze recent tokenizer adaptation techniques for Large Language Models (LLMs) in cross-lingual and multilingual settings. The study aims to investigate how tokenizer replacement and vocabulary transfer methods preserve the linguistic knowledge encoded in frozen transformer architectures while improving efficiency for non-English and low-resource languages.
- This research also compares existing approaches in terms of initialization strategy, computational cost, scalability, external resource dependency, and post-adaptation training requirements, with the goal of identifying current limitations and future directions for efficient and universal multilingual LLM adaptation.

Practical or simulation results

Empirical evaluations across encoder-only and decoder-only LLMs (e.g., BERT, GPT-style models, Llama, Mistral) show that tokenizer adaptation consistently improves cross-lingual efficiency and performance without full model retraining. Methods such as WECHSEL, TokAlign, and SALT reduce convergence time in CPT/LAPT while maintaining competitive downstream accuracy. Heuristic approaches like ReTok and Franken-Adapter provide strong gains in inference speed by reducing token fragmentation, though they typically require additional fine-tuning. In contrast, alignment-based methods (e.g., TokAlign, Trans-Tokenization) achieve higher fidelity but depend on external resources such as corpora or embeddings. Advanced approaches like ZeTT and Token Distillation demonstrate near zero-shot transfer with minimal performance degradation in tasks such as XNLI and language modeling, albeit with higher upfront training cost. Overall, results highlight a clear trade-off between initialization cost and post-adaptation training requirements across methods.

References

[1] B. Minixhofer et al., “WECHSEL,” NAACL-HLT, 2022.
[2] F. Jiang et al., “Franken-Adapter: Cross-lingual adaptation of LLMs by embedding surgery,” arXiv:2506.xxxxx, 2025.
[3] F. Remy et al., “Trans-tokenization and cross-lingual vocabulary transfers,” arXiv:2408.04303, 2024.
[5] F. Jiang et al., “EMBSWAP / LoRA-Adapt,” arXiv:2506.xxxxx, 2025.
[6] S. Sharthak et al., “TokenAdapt,” arXiv:2505.xxxxx, 2025.
[7] C. Li et al., “TokAlign,” arXiv:2506.03523, 2025.
[8] S. Gu et al., “MATT,” arXiv:2410.xxxxx, 2024.
[9] G. Andriushchenko et al., “ReTok,” arXiv:2412.07633, 2024.
[10] B. Minixhofer et al., “ZeTT: Zero-Shot Tokenizer Transfer,” NeurIPS, 2024.
[12] M. Artetxe et al., “On cross-lingual transferability of monolingual representations,” 2020.
[14] J. Hong et al., “FOCUS,” arXiv:2410.xxxxx, 2024.
[15] S. Lee et al., “SALT,” arXiv:2501.xxxxx, 2025.
[16] K. Dobler et al., “Token Distillation,” arXiv:2505.20133, 2025.