The Edge-First Feature Extractor: Enabling Efficient Multilingual NLP with Static and Distilled LLM Features

S Mahmudul Hasan
Tulane University
New Orleans, LA 70118, USA
shasan1@tulane.edu

Lu Peng
Tulane University
New Orleans, LA 70118, USA
lpeng3@tulane.edu

Abstract

Multilingual Large Language Models (LLMs) offer strong performance but often suffer from memory and energy costs that prohibit edge deployment. We propose the *Edge-First Feature Extractor*, a framework that reuses frozen LLM word embeddings as a universal feature source to bypass the heavy transformer stack. Our method supports two modes: *Static*, which directly aggregates frozen embeddings, and *Hybrid*, which adds a lightweight distilled enhancer. Experiments on a Jetson Orin NX across five languages show that *Static* embeddings achieve up to **16.8× higher throughput** and **287.3× lower Energy-Delay Product (EDP)** than FastText. The *Hybrid* mode bridges the gap, recovering near-LLM accuracy (within 3–5% F1) while maintaining up to **12.5× EDP reduction**. This unified approach enables practical, always-on multilingual NLP on edge devices without relying on heavy compression.

CCS Concepts

• **Computing methodologies** → **Natural language processing**;
• **Computer systems organization** → **Embedded systems**.

Keywords

Edge NLP, Multilingual LLMs, Efficient Inference, Energy Efficiency

ACM Reference Format:
S Mahmudul Hasan and Lu Peng. 2026. The Edge-First Feature Extractor: Enabling Efficient Multilingual NLP with Static and Distilled LLM Features. In *The 41st ACM/SIGAPP Symposium on Applied Computing (SAC '26), March 23–27, 2026, Thessaloniki, Greece.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3748522.3779745

1 Introduction

Transformer-based Large Language Models (LLMs) [5] have revolutionized NLP, yet their prohibitive memory and energy costs often preclude deployment on resource-constrained edge devices [4]. While standard compression techniques (e.g., quantization [3], pruning [2]) reduce model size, they still require the execution of the deep transformer stack for every token which incurs high memory, energy and computational costs.

In this work, we take an orthogonal approach: instead of compressing the transformer stack, we bypass it entirely. We reuse the frozen word embedding layer of multilingual LLMs as a universal feature source. Our framework provides two operating modes:

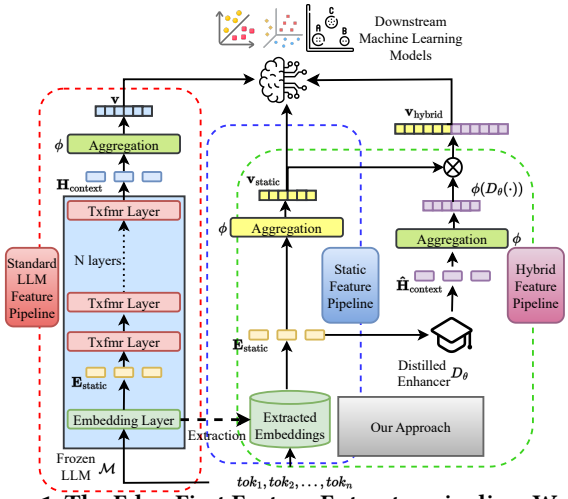

Figure 1: The Edge-First Feature Extractor pipeline. We bypass the transformer stack, using static embeddings or enhancing them via a distilled enhancer D_θ.

Static, which aggregates frozen embeddings for maximum speed, and **Hybrid**, which adds a lightweight distilled enhancer (<2% of LLM size). This enhancer acts as a *supplemental feature generator* rather than a task-specific classifier, allowing the same artifact to serve multiple downstream tasks and languages. We evaluate this framework across **6 multilingual lightweight LLMs**, five diverse languages (Arabic (Ar), English (En), Chinese (Zh), Japanese (Ja), Turkish (Tr)), and three tasks: binary and ternary sentiment classification, and clustering on two platforms: an embedded edge device (NVIDIA Jetson Orin NX) and a consumer laptop.

Our paper makes the following contributions:

- We demonstrate that static word embeddings from frozen multilingual LLMs enable on-device NLP with up to **10.4× higher median throughput** and **135.9× lower median EDP** (Energy-Delay Product) than FastText on Jetson.
- The proposed **Hybrid** mode closes most of the performance gap, achieving within **3–5% F1** (classification) and **10% V-measure** (clustering) of full LLMs, while offering **3–4×** higher throughput.
- We validate our unified deployment strategy across **6 LLMs** and **5 languages**, showing that one artifact can replace fragmented baselines achieving up to **287.3×** lower median EDP.

2 The Edge-First Feature Extractor

We repurpose the frozen word embedding layer of multilingual LLMs as a universal feature source, enabling our framework (Figure 1) to bypass the transformer stack. It offers a *static mode* for peak efficiency and a *hybrid mode* with a distilled feature enhancer for improved representation quality on edge devices.

SAC '26, Thessaloniki, Greece

ACM ISBN 979-8-4007-2294-3/2026/03
https://doi.org/10.1145/3748522.3779745

Table 1: Efficiency & Performance Benchmarks. Metrics report the median over downstream models (counts in the rightmost column). Efficiency is median Throughput Speedup and EDP Reduction over FastText; performance is median F1/V-Measure on the test split. Mono-C = monolingual classification, Mono-K = monolingual clustering, Multi-C = multilingual classification.

Method	Speedup Gains (Throughput × vs FastText)						Efficiency Gains (EDP Reduction × vs FastText)						Task Performance			# of Models Trained		
	Jetson Orin NX (Edge)			Laptop			Jetson Orin NX (Edge)			Laptop			Classification		Clustering	Monolingual		Multilingual
	Mono-C	Mono-K	Multi-C	Mono-C	Mono-K	Multi-C	Mono-C	Mono-K	Multi-C	Mono-C	Mono-K	Multi-C	Binary-F1	Ternary-F1	V-Measure	Classifiers	Clusterers	Classifiers
FastText (Base)	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	1.0×	0.86	0.66	0.28	400	50	80
Static (Ours)	**10.4**×	**6.1**×	**16.8**×	**2.2**×	**2.1**×	**4.9**×	**135.9**×	**38.0**×	**287.3**×	**15.5**×	**6.9**×	**31.8**×	0.86	0.66	0.30	2,400	300	480
Hybrid (Ours)	3.3×	1.9×	3.4×	1.3×	1.2×	1.4×	12.5×	3.0×	11.0×	3.2×	1.3×	2.1×	**0.89**	**0.68**	**0.39**	2,400	300	480
Full LLM	0.8×	0.5×	0.9×	0.9×	0.6×	0.8×	0.7×	0.1×	0.6×	0.8×	0.2×	0.6×	0.91	0.72	0.49	960	120	192

Static Feature Pipeline. We extract token embeddings $\mathbf{E}_{\text{static}} \in \mathbb{R}^{n \times d}$ using the LLM's native tokenizer and frozen embedding layer. These are aggregated into a fixed-length vector $\mathbf{v}_{\text{static}}$ using lightweight transformation ϕ, i.e., $\mathbf{v}_{\text{static}} = \phi(\mathbf{E}_{\text{static}})$.

Hybrid Feature Pipeline. To restore some contextual awareness, we introduce a **Distilled Enhancer** (D_θ), a light model to map static embeddings to a full LLM's contextualized embeddings. The final representation is: $\mathbf{v}_{\text{hybrid}} = [\mathbf{v}_{\text{static}} \; ; \; \phi(D_\theta(\mathbf{E}_{\text{static}}))]$.

Aggregation Strategies, ϕ. We evaluate five ways to construct $\mathbf{v}_{\text{static}}$: Mean, Concatenated (Mean/Max/Min), TF-IDF, SIF, and BOREP. SIF [1] down-weights frequent tokens, while BOREP [6] enriches semantics by projecting embeddings into higher dimensions.

Distilled Feature Enhancer. D_θ is a 4-layer transformer trained via **MSE loss** to mimic the parent LLM's ($\mathcal{M}$) contextualized features using the train-splits of the datasets. Despite only ∼1.1M–1.5M parameters, as shown in Table 2, they deliver **>99% compression** with cosine similarity $\rho > 0.83$.

Table 2: Model Specifications. Models grouped by enhancer size (D_θ). All enhancers achieve >99% compression (%η). ρ: Cosine similarity range.

Multilingual LLM	$\mathcal{M}$ (Millions)	D_θ (Millions)	Reduction, %η	Cos. Sim. ρ
E5	118	1.1	99.1	0.92
BERT-M / XLM-R / MPNet	178–278	1.3	99.3–99.5	0.84–0.99
XGLM / BGE-M3	565–568	1.4	99.7	0.87–0.96

Unified Multilingual Deployment. Our framework enables training of a **single deployable artifact** for multilingual deployment. It achieves this by reusing the multilingual LLM's tokenizer and frozen embeddings, removing all per-language components.

Downstream Evaluation. We evaluate feature quality by training lightweight classifiers: logistic regression, XGBoost, LightGBM, Gaussian Naïve Bayes, and shallow neural networks and clustering models (`K-Means`, and `K-Means++`) on the extracted embeddings (FastText, Static, Hybrid, and full LLM).

3 Experiments and Results

Setup & Protocol. Experiments run on a Jetson Orin NX (8-core Arm CPU, 16GB RAM, Ampere GPU) and a laptop (Core i7, 8GB 2000 Ada GPU, 32GB RAM) using Docker and energy is measured via `tegrastats` and `CodeCarbon`, respectively. We compare **FastText**, **Static**, **Hybrid**, and **Full LLM** across **5 languages** on sentiment (Amazon Reviews, LABR, Turkish Movie Reviews) and clustering (AG News, Livedoor, TNews, Sanad, Turkish News). Inputs are truncated to 800 characters (sentiment) and 2500 (clustering) and efficiency is measured from end-to-end inference at batch size 64.

System Efficiency. Table 1 highlights a clear shift in edge NLP efficiency. On the Jetson Orin NX, **Static Mode** cuts median EDP by **135.9**× and increases median throughput by **10.4**× over FastText, showing that bypassing the transformer stack eliminates the main bottleneck for always-on devices. **Hybrid Mode** offers a strong compromise: despite the added cost of D_θ, it delivers a **12.5× lower median EDP** and **3.3× higher median throughput**. Full LLM inference remains impractical, with a **0.7× median EDP penalty**. Laptop results follow the same pattern, with Static Mode cutting median EDP by **15.5**×.

Task Performance & Multilingual Utility. Efficiency gains must not compromise utility. **Hybrid Mode** bridges the semantic gap, achieving **0.89 median F1** (Binary Classification) and **0.39 median V-Measure** (Clustering), within **3–5%** of the Full LLM (0.91 median F1, 0.49 median V-M) for classification and within **10%** for clustering. While **Static Mode** excels in supervised tasks (0.86 median F1), Hybrid is critical for unsupervised clustering, boosting median V-Measure by **30%** over Static (0.39 vs 0.30).

The largest gains appear in **Multilingual Deployment**: by removing language detection and dynamic model loading, the **Unified Artifact** achieves up to **287.3× lower median EDP** and **16.8× higher median throughput** than fragmented FastText pipeline, showing a single static embedding matrix outperforms switching between language-specific models for multilingual edge systems.

Implications & Strategy. Our results highlight the power of **Architectural Reuse** over traditional compression. By repurposing the unified embedding layer (∼350MB), we eliminate the multi-gigabyte footprint and language detection latency inherent to fragmented baselines such as FastText. This enables a pragmatic two-tier deployment strategy: **Static Mode** serves as the ideal solution for ultra-low-power, "always-on" wake-word or filtering tasks, while **Hybrid Mode** acts as the robust default for accuracy-sensitive applications, delivering near-LLM performance without the prohibitive cost of the transformer stack.

References

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR.*

[2] Tim Dettmers, Ruslan A Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. 2024. Spqr: a sparse-quantized representation for near-lossless llm weight compression. In *The Twelfth International Conference on Learning Representations.*

[3] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan-Adrian Alistarh. 2023. Optq: accurate post-training quantization for generative pre-trained transformers. In *11th International Conference on Learning Representations.*

[4] Stefanos Laskaridis, Kleomenis Katevas, Lorenzo Minto, and Hamed Haddadi. 2024. Melting point: mobile evaluation of language transformers. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking.*

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[6] John Wieting and Douwe Kiela. 2019. No training required: exploring random encoders for sentence classification. In *7th International Conference on Learning Representations, ICLR.*