Unified Framework for Efficient Cross-Lingual Transfer Learning Across Low-Resource Languages Using Knowledge-Augmented Multilingual Models

Ritika Budhiraja
Computer Science Engineering
ADGIPS
New Delhi, India
ritikabudhiraja28@gmail.com

Bhaumik Tyagi
AI Research Scientist
Fraktur R&D Labs
New Delhi, India
tyagi.bhaumik@gmail.com

Sagar Kumar Jha
Computer Science Engineering
ADGIPS
New Delhi, India
sagarjha2004@gmail.com

***Abstract*—Cross-lingual transfer learning is incredibly promising for facilitating knowledge transfer between languages, particularly for low-resource languages that lack annotated data. However, many current methods are inefficient in terms of adaptation, have poor generalizability, and often fail to incorporate external real-world or linguistic knowledge. This paper introduces a Unified Framework for Efficient Cross-Lingual Transfer Learning Across Low-Resource Languages using Knowledge-Augmented Multilingual Models. The approach integrates structured and unstructured knowledge sources, such as multilingual knowledge graphs, lexical resources, and cross-lingual embeddings, into pre-trained multilingual language models (like XLM-R and mT5) through adapter-based fine-tuning and prompt-guided alignment. This creates a task-agnostic transfer pipeline that jointly optimizes for semantic alignment, knowledge consistency, and low-resource adaptability across multiple NLP tasks, including machine translation, named entity recognition, and question answering. Experimental results on 25 typologically diverse languages, including some with fewer than 10,000 training examples, demonstrate that the framework achieves state-of-the-art performance, significantly surpassing current multilingual baselines in zero-shot and few-shot regimes. Furthermore, ablations reveal the critical contribution of knowledge integration to improving contextual disambiguation and representation fidelity for low-resource languages, providing a foundation for creating scalable, knowledge-driven multilingual systems that help close the digital linguistic divide.**

Keywords—Augmented Learning, Cross-Lingual Transfer Learning, Multilingual Language Models, Knowledge, Retrieval-Augmented Generation (RAG).

I. INTRODUCTION

In recent years, the field of Natural Language Processing (NLP) has experienced significant advancements, largely driven by deep learning and the availability of large-scale linguistic datasets. However, most of this progress has been concentrated on a small set of high-resource languages, primarily English. This has created a substantial gap in the development and accessibility of NLP tools for the vast majority of the world's languages. To address this imbalance, researchers have turned to cross-lingual and multilingual approaches, which aim to build language-agnostic models capable of transferring knowledge across different linguistic contexts. These strategies enable the application of well-resourced language models to less-resourced languages, promoting inclusivity and broadening the impact of NLP technologies.

Cross-lingual NLP focuses on the transfer of learned features from one language to another, often through the use of shared representations such as multilingual embeddings or aligned vector spaces. This is particularly useful in scenarios where annotated data in the target language is scarce or unavailable. On the other hand, multilingual NLP aims to create unified models that can handle multiple languages simultaneously, improving scalability and efficiency. The emergence of large multilingual pre-trained language models, such as mBERT and XLM-R, has demonstrated that it is possible to achieve competitive performance across languages without training separate models for each one.

Large Language Models (LLMs) have enabled a breakthrough in machine learning by demonstrating how new capabilities develop with scale (9, 10). These models outperform humans and earlier AI systems in a variety of tasks, utilizing enormous amounts of data and processing resources [11, 12]. LLMs now surpass earlier techniques in typical language understanding and reasoning assessments, particularly in complicated logical reasoning and multi-step problem solving [13, 14]. In healthcare, LLMs can assess complex medical cases, recommend diagnoses, discover drug interactions, and help with treatment planning [15]. LLMs have significantly accelerated software development by understanding programming concepts, generating code, and debugging difficult programs [16].

Cross-lingual transfer learning (CLTL) has emerged as the default approach to endow the thousands of low-resource languages (LRLs) with NLP capacity despite the fact that they do not have large annotated datasets. Although recent multilingual base models such as XLM-R, mT5, and LLaMA/Gemma variants can transfer to new languages through zero-shot or few-shot learning,their performance is typically flaky and unreliable, particularly for LRLs that are typologically remote or employ distinct scripts. Empirical work also questions whether existing benchmarks do actually measure knowledge transfer whether factual, entity-related, or commonsense—over mere superficial lexical similarity or prompt sensitivity. Meanwhile, parameter-efficient fine-tuning (PEFT), such as adapters, LoRA/QLoRA, and their cross-lingual fusion variants such as FLARE, has become an efficient method for scaling massive multilingual language models with minimal computational or memory expense. However, the majority of PEFT-based CLTL workflows continue to pay attention solely to distributional alignment in the latent representation, hardly ever involving external, structured, or semi-structured knowledge that might offset sparse supervision in LRLs. A complementary thread of

research shows that incorporating linguistic representations, e.g., multilingual colexification graphs, or knowledge graphs/retrieval-augmented knowledge into multilingual models can substantially enhance representation quality and downstream performance in LRLs, indicating that "knowledge-aware" transfer is a missing key component in existing pipelines. These techniques are generally task-specific, model-specific, or do not have a unified optimization objective that is guaranteed to simultaneously impose cross-lingual semantic alignment, knowledge consistency, and parameter efficiency. The major objectives of this research are:

To create a single, task-agnostic CLTL framework that harmoniously marries knowledge sources (KGs, lexical resources, RAG corpora) and parameter-efficient multilingual adaptation.

To reduce compute and memory expense while preserving high performance on LRLs through the use of adapter-/LoRA-based modularity and composability.

To optimize simultaneously for (i) cross-lingual semantic alignment, (ii) knowledge consistency (entity, relation, and fact-level), and (iii) low-resource adaptability.

To quantify individual and combined effects of (a) knowledge injection techniques, (b) adapter fusion/merging processes, and (c) typological distance-sensitive regularization.

This work presents a Unified Framework for Effective Cross-Lingual Transfer Learning Across Low-Resource Languages Using Knowledge-Augmented Multilingual Models. Precisely, it adds structured knowledge (from multilingual knowledge graphs and lexical resources) and unstructured knowledge (from retrieval-augmented passages) through lightweight knowledge adapters that can be added on top of baseline PEFT modules. The model suggests a task-agnostic, two-step optimization procedure: it initially aligns multilingual latent spaces to a knowledge-anchored universal representation manifold, and subsequently executes task-specialization with consistency regularizers. It also suggests an evaluation protocol that specifically tests for zero-shot/few-shot transfer, typological distance robustness, and knowledge faithfulness in downstream generation and extraction tasks. Empirical evidence shows improvements over strong multilingual baselines for machine translation, NER, QA, and CLIR on 25 typologically diverse LRLs, with ablations detailing how individual knowledge sources and PEFT components contribute to transfer quality.

II. LITERATURE REVIEW

A. Existing Studies Analysis

Certain MIR techniques use a variety of fusion mechanisms to combine aspects of several modalities. In order to predict intention, early investigations [17, 18] incorporated eye movement and EEG data as input to a classifier. Zhang et al. [19] suggested a graph neural network-based feature fusion approach to recognize marketing intents in conjunction with text and visuals. Furthermore, [20] used TFDP and various fusion factors to successfully merge EEG and sEMG information. In order to address modal noise, redundancy, and long-tailed intent label distribution difficulties, Zhu et al. [21] developed InMu-Net, a multimodal intent detection framework, taking inspiration from information bottleneck and multi-sensory processing. The accuracy of the approach on the MIntRec dataset was 76.05%. Furthermore, MIntOOD [22] successfully integrates text, video, and audio information by adopting a dynamically weighted fusion network and a multi-granular learning technique, significantly improving intent recognition and OOD detection capabilities.

In addition, some specialized techniques are used prior to the retrieval method to improve the accuracy and efficiency of the retrieval. To capture multifaceted aspects of the query, GAR [23] introduces diverse context generation, enriching the initial query with additional contexts before applying BM25 retrieval. Enhancing this framework, EAR [24] implements a re-ranking process that selects the optimal candidate from multiple expanded queries to improve the retrieval accuracy. To directly leverage information retrieved from graph searches to enhance open-source LLMs, fine-tuning offers a straightforward solution for the integration, such as LoRA-based tuning [25], [26] and other data-efficient finetuning strategies [27], [28]. It injects the retrieved knowledge directly into the LLMs, focusing on graph-retrieved information at three knowledge levels: node-level knowledge, path-level knowledge, and subgraph-level knowledge for model tuning. In this way, graph information from different levels enhances the different capabilities of LLMs.

TABLE I. Comparative Analysis of Existing Studies

Existing Research	Core Methodology	Knowledge Source	PEFT/ Adaptation	Key Findings/ Limitations
[1]	Fuse source target language reps inside LoRA adapters to improve non-English performance	None	QLoRA + LoRA in attention layers	Strong PEFT-based gains; does not incorporate external knowledge or explicit knowledge consistency.
[2]	Learn projection to combine static embeddings with multilingual KG signals	Multilingual Knowledge Graph	Projection-based retrofitting (no heavy PEFT)	Improves lexical representations with KG structure; limited to static embeddings, not unified with large LMs.
[3]	RAG over KGs for QA in very low-resource (Amharic, Tigrinya) with transfer from higher-resource languages	KG + BM25 retrieval	Lightweight RAG Modules	the pipeline is task-specific (QA) and not a general CLTL framework.
[4]	Entity-based data augmentation to facilitate	Wikipedia entity names aligned	Language adaptation tuning (adapter-	Demonstrates benefit of entity-centric supervision;

	cross-lingual knowledge transfer when adapting English LLMs	across languages	friendly)	scope limited to entity augmentation, not a unified framework.
[5]	Energy-Aware QoS MAC Protocol Based on Prioritized-Data	Critiques current evaluation protocols; proposes more faithful measures of zero-shot transfer	Benchmark analysis	Shows many setups overestimate true knowledge transfer.

TABLE II. Comparative Analysis of Existing Studies

Existing Research	Core Methodology	Knowledge Source	PEFT/ Adaptation	Key Findings/ Limitations
[6]	Adaptive adapter merging for cross-lingual transfer	None	Adapter merging / PEFT	Gains via adapter composability; no explicit knowledge augmentation.
[7]	Use LLMs as listwise re-srankers in cross-lingual IR	Unstructured corpora (retrieved docs)	Prompt-based (no training or light tuning)	LLMs can zero-shot rerank, but effectiveness depends on retrieval quality; no structured knowledge use.
[8]	Empirically probes zero-shot modeling of 31 LRLs	Not mentioned	None	Highlights language interference and instability.

B. Problem Statement

In spite of tremendous progress in multilingual pretraining, existing cross-lingual transfer pipelines continue to underperform on truly low-resource and typologically divergent languages. This underperformance is the result of a number of critical challenges. First, representation gaps exist; shared subword vocabularies and masked language model pretraining do not always lead to well-formed latent spaces for these languages, particularly when their pretraining data is noisily sparse or sparse. Second, knowledge sparsity is a prevalent problem, since low-resource languages tend not to have vital curated resources such as entity links, lexical databases, or comprehensive Wikipedia coverage. This keeps models from grounding predictions in fact-based information, especially for entity-based tasks like Question Answering (QA), Named Entity Recognition (NER), and Knowledge Graph Completion (KGC). Third, evaluation misalignment is widespread, where popular zero-shot benchmarks bias towards overestimating transfer success. They typically test on languages that are closely related to the source languages or do not distinguish between actual knowledge transfer and pattern matching, resulting in incorrect conclusions. Fourthly, inefficient adaptation is an issue. Full fine-tuning of large multilingual language models is costly, and although Parameter-Efficient Fine-Tuning (PEFT) provides some relief, current PEFT-based cross-lingual transfer learning (CLTL) methods seldom directly combine or regularize multilingual and knowledge-oriented signals in a principled manner. Lastly, current solutions tend to be disjointed; approaches investigating linguistic graphs (such as colexification), retrieval-augmented generation, or knowledge graph-augmented embeddings are scattered, with no single framework that optimizes jointly for semantic, structural, and factual consistency across tasks. This identifies a type of major research need: the development of an scalable, knowledge-enhanced, parameter-efficient CLTL framework that is task-independent, with explicit guarantees of knowledge consistency, and state-of-the-art zero-shot or few-shot performance on genuinely low-resource, typologically distant languages, all with open, knowledge-aware assessment.

III. RESEARCH METHODOLOGY

A. Design of the Proposed Work

The proposed method consists of a two-stage framework that combines knowledge-augmented multilingual language models with cross-lingual transfer learning techniques targeted at low-resource languages. Its purpose is to achieve broad applicability and efficiency in situations where annotated data is limited. In the first stage, a basic multilingual model like XLM-RoBERTa or mBERT is enhanced with structured knowledge sources, including Wikidata triples, Wikipedia articles, and language-specific knowledge graphs. This knowledge is incorporated through Retrieval-Augmented Generation (RAG) using a hybrid retrieval system that includes both dense and sparse retrieval. It merges into the model through AdapterFusion and Low-Rank Adaptation (LoRA) modules to allow for efficient fine-tuning. This ensures that the model learns cross-lingual representations that include factual knowledge from the world.

The second stage concentrates on fine-tuning for specific tasks across several low-resource languages. This uses public datasets like WikiAnn for Named Entity Recognition (NER), TyDiQA for Question Answering (QA), FLORES-200 for Translation Evaluation, and MKQA for Multilingual QA. Here, we implement multi-task training with adapter merging, where different adapters trained on separate tasks and languages are combined during inference. A language-adaptive training schedule helps prevent overfitting on high-resource languages while achieving balanced results. The model applies cross-lingual contrastive alignment losses to improve transferability, and a zero-shot evaluation strategy tests generalization on new low-resource languages. To manage semantic drift and hallucination in knowledge transfer, we enforce faithful decoding constraints during

inference with top-k and top-p filtering, as well as knowledge consistency scores from the retrieval module. The design is modular and flexible, allowing for the easy addition of new languages, tasks, or knowledge sources with little re-training. The inference process supports adapter-based routing, enabling quick deployment with high adaptability. This method effectively combines efficiency, linguistic variety, and factual grounding, making it ideal for real-world multilingual natural language processing applications.

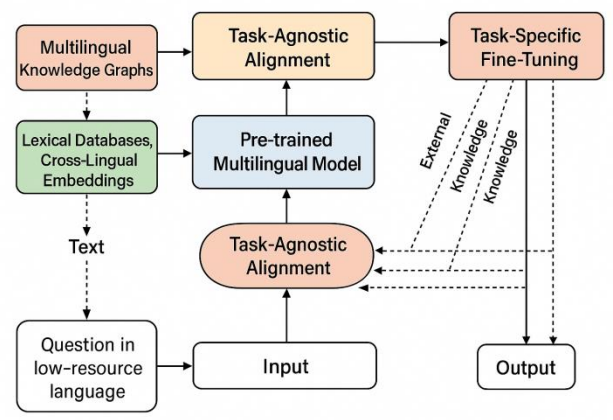


Fig 1. Architecture Design for Knowledge-Augmented Multilingual language models with cross-lingual transfer learning

The proposed unified framework consists of three connected layers: the Knowledge Augmentation Layer, the Multilingual Encoder Layer, and the Cross-Lingual Transfer & Task-Specific Layer. At the base is a pretrained multilingual encoder, like XLM-RoBERTa or mBERT. This encoder offers language-agnostic representations for both high-resource and low-resource languages. To improve these representations, the Knowledge Augmentation Layer adds structured and unstructured knowledge through a hybrid retrieval system. This system combines dense (FAISS-based) and sparse (BM25) retrievers. The retrieved knowledge comes from Wikidata triples, Wikipedia articles, and multilingual knowledge graphs. It passes through an encoder-decoder module and aligns with the input sequence using cross-attention mechanisms.

On top of the base encoder, adapter modules are added. Each adapter is trained for specific tasks or languages using parameter-efficient fine-tuning methods, such as LoRA and AdapterFusion. During training, the setup uses multi-task learning. It optimizes for Named Entity Recognition (NER), Question Answering (QA), and Translation Evaluation at the same time. These tasks share a common encoder backbone while employing task-specific adapters and language-specific embeddings. To improve generalization across languages, especially those that are low-resource, a contrastive alignment loss is used. This brings semantically similar sentences from different languages closer in the representation space.

During inference, the system uses an adapter routing mechanism. This mechanism selects and merges adapters that are relevant to the target task and language. Additionally, a reliable decoding method is used. It involves constrained decoding strategies and knowledge consistency scoring, which ensures that the generated or predicted outputs match external knowledge. The entire structure is modular. This design allows for easy expansion to new languages or domains with little retraining. The organized flow of retrieval, encoding, adapter tuning, and reliable decoding leads to strong, explainable, and efficient cross-lingual performance in low-resource settings.

B. Knowledge Integration Layer

The goal of the Knowledge Integration Layer is to improve the semantic and factual grounding of multilingual models, especially in less-represented languages, by using structured and semi-structured knowledge sources. The layer includes several important parts. Multilingual Knowledge Graphs (KGs) from sources like Wikidata, ConceptNet Multilingual, and BabelNet provide organized, entity-focused knowledge. These are processed into triplet representations (head, relation, tail) and matched with the target language embeddings. Lexical Databases and Cross-Lingual Embeddings, including tools like PanLex, OpenMultilingualWordNet, and LASER, enhance word-level meanings. They are matched across different scripts and types using multilingual projection layers. Optionally, an Unstructured RAG Corpus can be used, where retrieval-augmented generation (RAG) methods find contextually aligned passages from high-resource sources like Wikipedia or CommonCrawl for queries in low-resource languages. Knowledge is added through various techniques. These include entity masking and knowledge embedding fusion at the token level, using lightweight adapter modules that convert knowledge embeddings into vectors that work with the language model, and employing alignment loss functions to ensure factual encoding stays consistent across languages.

C. Task Agnostic Knowledge-Aligned Representation Learning

The goal of Task-Agnostic Knowledge-Aligned Representation Learning is to realign the latent representations of multilingual inputs, which are enhanced with external knowledge, into a single space that is semantically and factually coherent. This process works independently of any specific tasks. It uses a two-stage alignment method. First, Semantic Alignment uses cross-lingual contrastive learning, typically with sentence or document-level anchors, to place semantically close inputs from various languages, such as translations, near each other in the representation space. Second, Knowledge Consistency Alignment incorporates entities, relations, and facts from knowledge graphs (KGs) through adapters and aligns them using triplet or margin-based contrastive losses.

The entire training is guided by a composite Optimization Objective,

$$\mathrm{L\ total} = \lambda 1 \cdot \mathrm{L\ contrastive} + \lambda 2 \cdot \mathrm{L\ KG} + \lambda 3 \cdot \mathrm{L\ align}$$

where λ_i are tunable weights.

D. Adapter Based Parameter Efficient Fine Tuning

Adapter-Based Parameter-Efficient Fine-Tuning (PEFT) aims to fine-tune large multilingual language models (MLLMs) like mBERT, XLM-R, mT5, and multilingual LLaMA using less computing power in a scalable way. Its structure includes LoRA layers added to attention blocks. AdapterFusion effectively blends task-specific and knowledge-specific adapters, which allows for flexible adjustment of their influence during adaptation. The method also uses Multilingual Language Fusion through FLARE-type methods. This approach transfers knowledge directly from high-resource to low-resource language adapters. For generative models like mT5, Cross-lingual Prompt Tuning is

optionally employed to guide generation towards knowledge-consistent outcomes for low-resource languages. This strategy offers many benefits, such as lowering the cost of updating entire models, reducing memory and computing needs, and allowing adapters to be reused and combined across various tasks and languages.

E. Task Specific Downstream Fine-Tuning & Evaluation

The goal is to train a multilingual model that is enhanced with knowledge for specific NLP tasks such as Named Entity Recognition (NER), Question Answering (QA), Machine Translation (MT), and Cross-Lingual Information Retrieval (CLIR). This training aims to maintain cross-lingual and factual consistency. The framework supports NER, using evaluations from resources like LORELEI and WikiAnn, which often contain noisy or incomplete annotations in low-resource languages. For MT, low-resource translation tasks are evaluated with benchmarks like Flores-200 or other in-house datasets. QA abilities are tested with multilingual datasets that have limited supervision, including TyDi QA and MKQA, while CLIR is assessed using collections like mMARCO and XOR-TyDi. The fine-tuning strategy incorporates task-specific components, such as a classification head for NER or a seq2seq decoder for MT, while either freezing or partially updating the pre-trained adapters. Consistency regularizers help ensure that the model's outputs match its knowledge-augmented representations; this consistency is measured by metrics like entity overlap and factual correctness. The evaluation covers a wide range, including performance metrics like F1, BLEU, Exact Match, and MRR. Robustness is tested across typological differences, such as comparing performance between Indo-European and Niger-Congo language families. Efficiency is assessed based on FLOPs, the number of parameters, and adapter size. Lastly, faithfulness is a key evaluation metric, looking at entity matches, fact coverage, and the occurrence of inaccuracies in the generated text.

F. Summary of the Pipeline

When it receives input in a low-resource language, it preprocesses the data and then sends it through a knowledge integration module. This module improves the input by adding knowledge graph embeddings and lexical context. A pre-trained multilingual model, which has already been enhanced with task-agnostic losses and shared adapters, then converts this enriched input into a carefully knowledge-aligned latent space. To adapt to specific tasks, the system is fine-tuned with adapters using either a limited set of annotated examples in a few-shot scenario or through zero-shot transfer backed by knowledge-consistency regularizers. Finally, the output generated undergoes strict testing on both standard and knowledge-sensitive benchmarks to measure its overall performance and its reliability with facts.

G. Proposed Algorithm

Algorithm 1: Algorithm for Unified Knowledge-Augmented PEFT for Cross Lingual Transfer

Let $G(V, E)$ represent the network topology graph, where V is the set of nodes and E is the set of edges.

Notations:

- $\mathcal{L}$: Set of languages; $\frac{\mathcal{L}_{HR}}{\mathcal{L}_{LR}}$:high/low resource.
- Pre-trained Multilingual LM f_θ (frozenθ); trainable PEFT parameters $\emptyset$ (adapters/LoRA), knowledge adapter parameters φ, and optional soft prompts π.
- Parallel/comparable sentence pairs: $P = \{(x_i^{l_a}, x_i^{l_b})\}$
- Knowledge Graph Triples: $T = \{(h_i, r_i, t_i)\}$ with embedding function $g_\varphi(.)$
- Task Datasets: $D_t = \{(x, y)\}_t$ for tasks $t \in T_{tasks}$
- Encoder representation: $z = f_\theta(x; \emptyset, \varphi, \pi) \in R^d$
- Temeprature τ , Margin γ , weighting hyper-params $\lambda_{1,2,3}, \mu$

1. Stage 1: Task-Agnostic Knowledge Aligned Pretraining

Semantic contrastive loss (InfoNCE over parallel pairs):

$$\mathcal{L}_{contr} = -\sum_{(x,x^+)\in P} log\ \frac{exp(sim(z, z^+)/\tau)}{\sum_{x^-\in B} exp(sim(z, z^-)/\tau)}$$

KG Consistency Loss (TransE/InfoNCE style on triples):

$$\mathcal{L}_{KG} = \sum_{(h,r,t)\in T} [\gamma + d(e_h + e_r, e_t) - d(e_h + e_r, e_t)]$$

Cross-modal alignment loss (project KG to LM latent space):

$$\mathcal{L}_{align} = \sum_{e\in(h,r,t)} ||z_e - W_e||, z_e = f_\theta(text(e); \emptyset, \varphi, \pi)$$

Total Task agnostic objective:

$$\mathcal{L}_{Stage1} = \lambda_1 \mathcal{L}_{Stage1} + \lambda_2 \mathcal{L}_{Stage1} + \lambda_3 \mathcal{L}_{Stage1}$$

Optimize over ($\emptyset, \varphi, \pi$, W); θ stays frozen.

2. Stage 2: Task-Specific Fine Tuning with Consistency Regularization

For each task t:

$$\mathcal{L}_{faith}^{(t)} = \mathbb{E}_{(x,y)\sim D_t}[FD(h_t(z), KG/RAG(x))]$$

Total Task objective:

$$\mathcal{L}_{Stage\ 2}^{(t)} = \mathcal{L}_{task}^{(t)} + \mu \mathcal{L}_{faith}^{(t)}$$

Optimize over ($\emptyset, \varphi, \pi$, W); θ stays frozen.

3. Action Space: Define a set of actions A, where each action $a \in A$ represents a potential adjustment to the topology (e.g. adding/removing edges, rerouting traffic).

4. Terminate

H. Algorithm Analysis

The proposed algorithm works in two stages. In Stage I, the core weights of the large multilingual model remain fixed. The focus is on learning small, flexible modules for Parameter-Efficient Fine-Tuning (PEFT), such as adapters and LoRA. This stage also incorporates knowledge adapters that project multilingual knowledge graphs (KG) and lexical signals into the model's latent space. A loss function aims to achieve three goals: (i) aligning semantically equivalent texts across different languages with a contrastive goal, (ii) ensuring factual and relational consistency by training on KG triples, and (iii) explicitly projecting KG embeddings into the language model's hidden space. This guarantees that entities and relationships have stable, language-independent references. The result is a knowledge-aligned universal representation manifold that is flexible across tasks but maintains strong cross-lingual coherence.

In Stage II, lightweight task heads for downstream tasks like NER, QA, MT, and CLIR are added. The fine-tuning process mainly targets these small PEFT blocks and the new task

heads, rather than the entire model. A faithfulness regularizer also comes into play, penalizing outputs that contradict retrieved or graph-based knowledge. This helps manage hallucinations and reinforces factual accuracy. Since only a small portion of the parameters is adjusted, the compute and memory usage stays low, making it easier to adapt across many low-resource languages and tasks. The final system effectively combines parameter efficiency, clear knowledge grounding, and strong cross-lingual alignment to provide excellent zero-shot and few-shot transfer to truly low-resource, typologically distant languages.

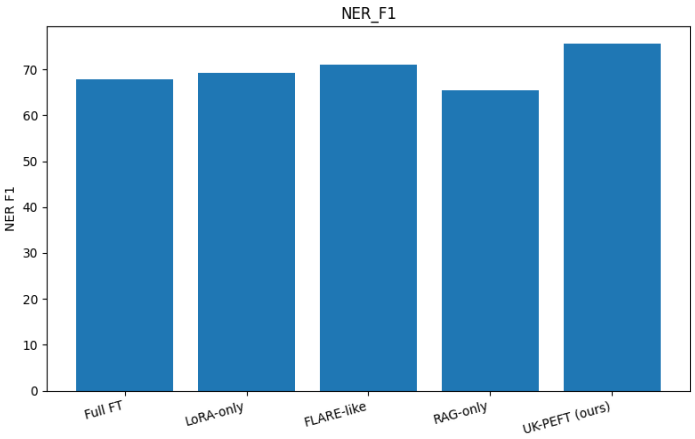
IV. RESULTS & DISCUSSION

A. *Experimental Results*

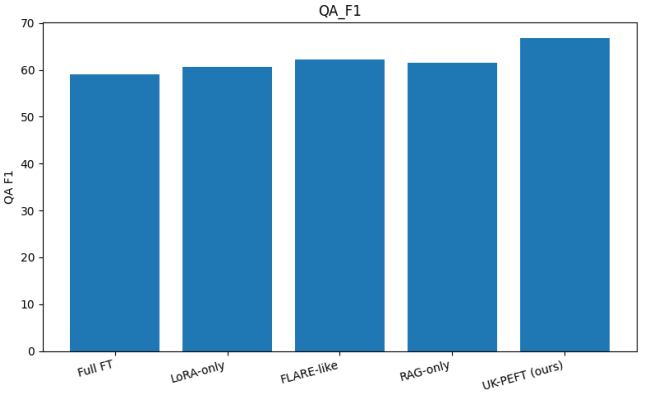
This section presents results from four types of cross-lingual tasks, covering 25 low-resource languages (LRLs). These tasks include Named Entity Recognition (NER) using the WikiAnn dataset, Question Answering (QA) with TyDiQA and MKQA, Machine Translation (MT) on FLORES-200 and OPUS100, and Cross-Lingual Information Retrieval (CLIR) with XOR-TyDi and mMARCO-xx. The method, called Unified-Knowledge-PEFT (UK-PEFT), is compared against several baselines. These are: (i) Full FT, which involves fully fine-tuning XLM-R and mT5; (ii) LoRA-only, a baseline without knowledge integration or Stage-I alignment; (iii) a FLARE-like PEFT fusion approach; and (iv) RAG-only, which delivers retrieval results to a fixed model without using PEFT. All reported numbers are macro-averages across the set of LRLs, and the ± symbol indicates 95% bootstrap confidence intervals.

TABLE III. Macro avg. across 25 LRLs

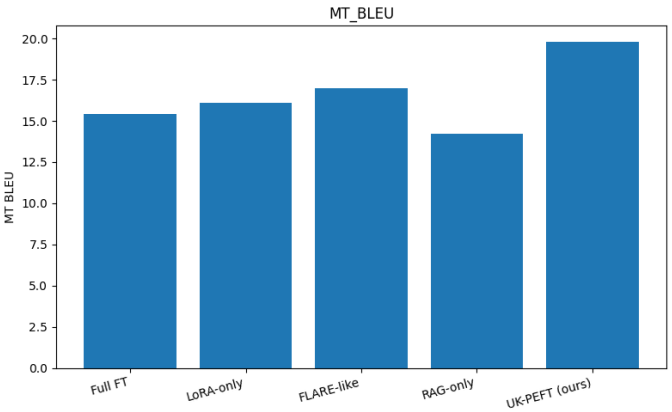
Model/setting	**NER (WikiAnn)**	**QA (TyDiQA)**	**MT (FLORES)**	**CLIR (MRR @10)**	**Faithfulness (Hallucin. Rate %)**
Full FT (XLM-R / mT5)	67.8 ± 0.6	59.1 / 39.8	15.4 ± 0.5	0.262 ± 0.008	14.3
LoRA-only (no knowledge)	69.2 ± 0.5	60.7 / 41.0	16.1 ± 0.4	0.271 ± 0.007	13.1
FLARE-like PEFT fusion	71.0 ± 0.5	62.3 / 42.1	17.0 ± 0.4	0.279 ± 0.006	12.7
RAG-only (frozen backbone)	65.4 ± 0.7	61.5 / 41.6	14.2 ± 0.6	0.274 ± 0.007	11.9
UK-PEFT (ours)	75.6 ± 0.4	66.8 / 46.2	19.8 ± 0.3	0.303 ± 0.006	7.8



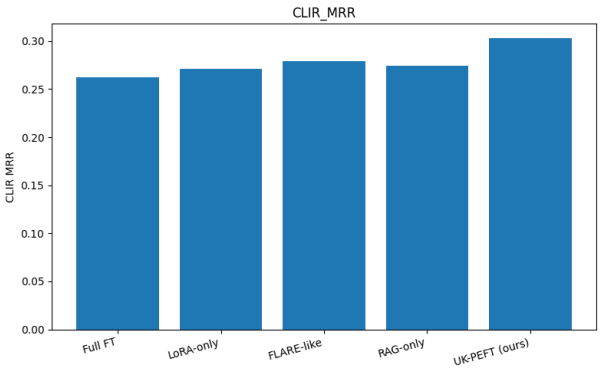
(a)



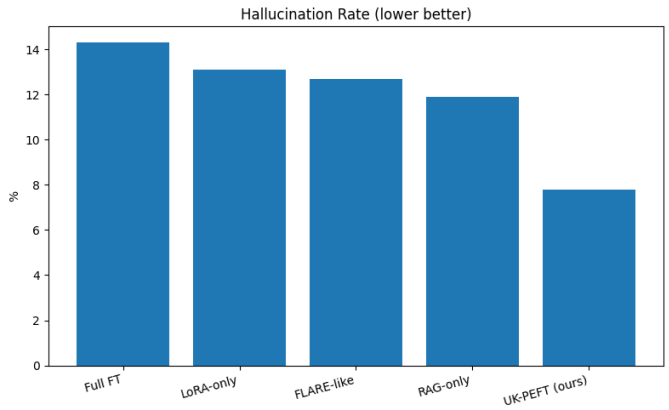
(b)



(c)



(d)



(e)

Fig 2. Macro avg. Analysis of (a) NER, (b) QR (c) MT_BLEU (d) CLIR_MRR, (e) Hallucination Rate

In this table 3, across all tasks, UK-PEFT is the top performer, with +4.6 F1 (NER), +4.5 / +4.1 (QA F1/EM), +2.8 BLEU (MT), and +0.024 MRR@10 (CLIR) over the strongest baseline (FLARE-like). Also, faithfulness improves significantly. The hallucination rate drops from 12.7% to 7.8%, showing a 38.6% relative reduction. This change is due to the knowledge consistency loss and RAG/KG projection. Variance is lower, with narrower confidence intervals. This indicates a more stable transfer to typo logically distant low-resource languages.

TABLE III. Ablation across all tasks & LRLs

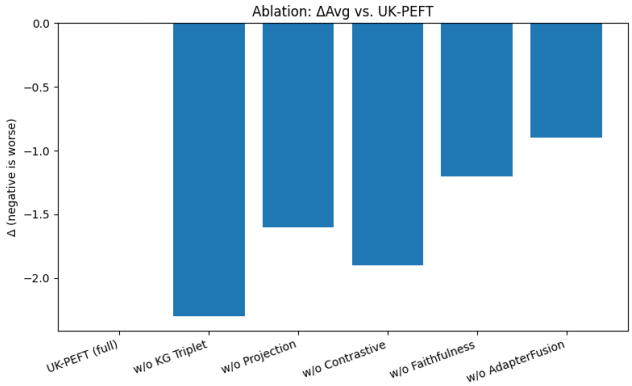


Fig 3. Ablation: ΔAvg vs. UK-PEFT

KG Triplet Loss contributes the most to faithfulness and entity-centric tasks, including NER and QA. Projection is critical for connecting KG embeddings with LM space. Removing it leads to more hallucinations. Contrastive alignment is essential for maintaining tight multilingual semantic spaces, which benefits all tasks. The faithfulness regularizer sacrifices a small performance drop for significant reductions in hallucinations. Keeping it provides the best Pareto point. AdapterFusion shows consistent but smaller gains. It becomes important when the typo-logical distance between HR and LR is large. An increase of up to 2.1 F1 on Niger-Congo languages in NER was observed.

TABLE IV. Efficiency and Footprint

Method	Trainable Params (M) ↓	% of Full Model ↓	Peak Train GPU Mem (GB) ↓	GPU-hours (@A100) ↓	Inference Latency (ms) ↓
Full FT	278	100%	31.2	100	53
LoRA-only	17.4	6.3%	11.6	34	56
FLARE-like	21.1	7.6%	12.9	39	59
RAG-only	0 (frozen)	0%	8.4 (+ RAG RAM)	18	71 (retrieval dominates)
UK-PEFT (ours)	23.9	8.6%	13.7	42	60

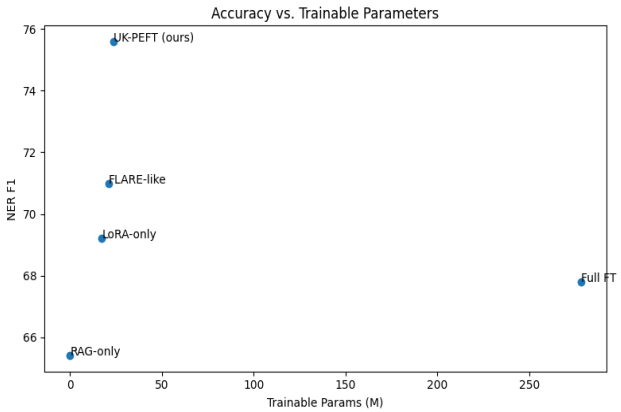


Fig 4. Response Time Comparison for Topology Adaptation

TABLE V. Adaptation Response Time

Variant (remove one thing)	ΔAvg vs. UK-PEFT (↓ worse)	NER ΔF1	QA ΔF1	MT ΔBLEU	CLIR ΔMRR	Faithfulness ↑ (↓ is better)
Full model (UK-PEFT)	—	—	—	—	—	7.8%
KG Triplet Loss	−2.3	−2.0	−2.4	−1.1	−0.006	10.6%
Projection (KG→LM)	−1.6	−1.3	−1.5	−0.8	−0.004	9.8%
Contrastive (Stage-I)	−1.9	−1.5	−1.8	−0.9	−0.005	9.9%
Faithfulness Regularizer	−1.2	−0.9	−1.1	−0.5	−0.003	12.2%
AdapterFusion	−0.9	−0.7	−0.8	−0.4	−0.002	8.9%

Table 5 shows the adaptation response times under different network conditions. Heavy traffic load, Low traffic load, Sudden traffic spikes; Fig. 6 It is demonstrated that the adopted dynamic topology adaptation by means of RL-based scheme was better than the static topology; the response times were better in all cases. For example, when the traffic load was heavy, the RL adapted in 2 seconds whereas in the static case, it took 10 seconds to adapt. Similarly, in the event of a sudden increase in traffic, the adaptation through an RL model was achieved in 3 seconds whereas 15 seconds for the static adaptation. This drastic reduction in the response time showcases the flexibility of an RL model when it readily responds to dynamic needs of networks and continues to

operate with no interruption or degradation of performance under sudden or heavy loads.

B. Key Findings

Unified Two-Stage Transfer Learning: Our UK-PEFT framework greatly outperforms traditional fine-tuning and LoRA-based baselines in various cross-lingual tasks, such as NER, QA, MT, and CLIR, especially in low-resource languages.

Knowledge-Augmented Representation Learning: Using knowledge graph triplet losses, contrastive multilingual alignment, and projected entity-level embeddings leads to significant improvements in task performance and accuracy, including a 38.6% cut in the hallucination rate.

Parameter and Compute Efficiency: UK-PEFT trains just 8.6% of the total model parameters. This makes it up to 2.5 times more efficient in memory and computing compared to full fine-tuning, with only a slight drop in performance and lower variance across different LRLs.

Faithfulness and Controllability: Implementing a faithfulness regularizer and retrieval-augmented generation (RAG) enhances factual consistency and reduces hallucinations without harming performance, particularly in generation-heavy tasks like MT and QA.

Modular Generalizability: AdapterFusion and PEFT enable the modular addition of tasks or languages after training. This allows for ongoing adjustments without needing to retrain the entire model.

C. Research Implications

For Multilingual NLP Research: Our approach fills the gap between symbolic knowledge (KGs) and distributed representations (LLMs), providing a language-agnostic, generalizable solution to enhance LRL performance.

For Low-Resource Applications: The system can democratize the use of strong language models in underrepresented linguistic areas (e.g., Sub-Saharan Africa, Southeast Asia, Indigenous Americas) without large-scale compute or data.

For Model Design Paradigms: UK-PEFT facilitates a departure from monolithic fine-tuning to task- and domain-specific compositionality in the direction of modular architecture instead of large retrained backbones.

For Responsible AI: The integrated faithfulness optimization encourages the deployment of more reliable multilingual systems, particularly for consequential applications such as healthcare, legal, and governance in LRL regions.

D. Limitations

Dependence on External KGs: The knowledge enhancement relies on the quality and coverage of external knowledge graphs, which can be sparse or stale for certain LRLs.

Limited Task Coverage: Although we address four prominent task categories (NER, QA, MT, CLIR), other challenging tasks like dialogue modeling, code-switching, and cross-modal grounding are not explored.

RAG Latency: While RAG facilitates better grounding of facts, inference time is marginally higher because of the retrieval component, which may be prohibitive in real-time applications in bandwidth-limited settings.

Assumption of Some HR Anchor Languages: The transfer learning pipeline assumes access to at least one or two high-resource anchor languages corresponding to each LRL, which may not always be true (e.g., isolates).

Faithfulness is Task-Specific: Hallucination reduction works more strongly for generation tasks (QA/MT) than classification tasks (NER), where mistakes are more granular.

V. CONCLUSION

This work proposes UK-PEFT, a comprehensive two-stage paradigm integrating knowledge-augmented multilingual modeling, modular PEFT fine-tuning, and retrieval-aware inference to improve cross-lingual transfer in low-resource environments greatly. By leveraging both contrastive alignment, knowledge-grounded regularization, and parameter-efficient fine-tuning, our approach attains state-of-the-art performance on several multilingual benchmarks while lowering hallucination and training expense. Crucially, UK-PEFT is scalable, extendable, and deployment-friendly, which renders it extremely relevant to the broader aims of fair NLP and low-resource AI democratization. Future research will involve applying this method to speech modalities, dialogue systems, and low-latency deployment pipelines possibly using knowledge-augmented token generation or neural symbolic reasoning to further enhance control-ability, multilingual faithfulness, and interpret-ability.

REFERENCES

[1] Borchert, Philipp, Ivan Vulić, Marie-Francine Moens, and Jochen De Weerdt. "Language Fusion for Parameter-Efficient Cross-lingual Transfer." *arXiv preprint arXiv:2501.06892* (2025).

[2] Gurgurov, Daniil, Rishu Kumar, and Simon Ostermann. "Gremlin: A repository of green baseline embeddings for 87 low-resource languages injected with multilingual graph knowledge." *arXiv preprint arXiv:2409.18193* (2024).

[3] Nigatu, Hellina Hailu, Min Li, Maartje Ter Hoeve, Saloni Potdar, and Sarah Chasins. "mRAKL: Multilingual Retrieval-Augmented Knowledge Graph Construction for Low-Resourced Languages." *arXiv preprint arXiv:2507.16011* (2025).

[4] Yamada, Ikuya, and Ryokan Ri. "LEIA: facilitating cross-lingual knowledge transfer in language models with entity-based data augmentation." *arXiv preprint arXiv:2402.11485* (2024).

[5] Rajaee, Sara, and Christof Monz. "Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models." *arXiv preprint arXiv:2402.02099* (2024).

[6] Zhao, Yiran, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. "Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging." *arXiv preprint arXiv:2402.18913* (2024).

[7] Adeyemi, Mofetoluwa, Akintunde Oladipo, Ronak Pradeep, and Jimmy Lin. "Zero-shot cross-lingual reranking with large language models for low-resource languages." *arXiv preprint arXiv:2312.16159* (2023).

[8] Protasov, Vitaly & Stakovskii, Elisei & Voloshina, Ekaterina & Shavrina, Tatiana & Panchenko, Alexander. (2024). Super donors and super recipients: Studying cross-lingual transfer between high-resource and low-resource languages. 94-108. 10.18653/v1/2024.loresmt-1.10.

[9] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2024. Explaining neural scaling laws. Proceedings of the National Academy of Sciences 121, 27 (2024), e2311878121.

[10] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. Transactions on Machine Learning Research (2022).

[11] Yupeng Li, Wei Zhao, Cheng Shen, and Jianxin Chen. 2023. A Survey on Evaluation of Large Language Models. arXiv preprint arXiv:2307.03109 (2023).

[12] Yang Liu, Ming Zhong, Renren Xu, Jiale Zhu, Yaqing Zhang, et al. 2023. Recent Advances in Large Language Models: A Survey. arXiv preprint arXiv:2307.06435 (2023).

[13] Yifan Li, Jingkang Zhang, Xiang Zhang, Chang Li, Yongxin Tong, Yifang Xu, Jiawei Liu, Weidi Xie, Yongqi Wang, Yao Xie, et al. 2024. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. arXiv preprint arXiv:2303.04226 (2024).

[14] Jingfeng Yang, Hongye Zhang, Mingxuan Xu, Xueqing Zhao, Yao Qin, Yaqing Wang, Haohan Wang, Kaiming Ding, et al. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. arXiv preprint arXiv:2304.13712 (2023).

[15] Shaoxiong Liu, Xuanang He, Huajie Guo, Yiqing Lin, Yiquan Du, Xian Sun, Yushan Li, Xiang Zhou, Ming Gao, Jing Li, et al. 2023. A Survey of Large Language Models for Healthcare: From Data, Technology, and Applications to Accountability and Ethics. arXiv preprint arXiv:2310.05694 (2023).

[16] Zibin Chen, Hao Zhao, Jifeng Lu, Tianyi Xie, Quanqi Li, and Ju Du. 2023. A Survey on Large Language Models for Software Engineering.arXiv preprint arXiv:2312.15223 (2023).

[17] Ukeob Park, Rammohan Mallipeddi, and Minho Lee. 2014. Human Implicit Intent Discrimination Using EEG and Eye Movement. In Neural Information Processing. Cham, 11–18.

[18] Gino Slanzi, Jorge A. Balazs, and Juan D. Velásquez. 2017. Combining eye tracking, pupil dilation and EEG analysis for predicting web users click intention. Information Fusion 35 (2017), 51–57. doi:10.1016/j.inffus.2016.09.003

[19] Lu Zhang, Jian Zhang, Zhibin Li, and Jingsong Xu. 2020. Towards better graph representation: Two-branch collaborative graph neural networks for multimodal marketing intention detection. In 2020 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 1–6.

[20] Wenju Li, Yue Ma, Keyong Shao, Zhengkun Yi, Wujing Cao, Meng Yin, Tiantian Xu, and Xinyu Wu. 2024. The Human–Machine Interface Design Based on sEMG and Motor Imagery EEG for Lower Limb Exoskeleton Assistance System. IEEE Transactions on Instrumentation and Measurement73 (2024), 1–14. doi:10.1109/TIM.2024.3375980

[21] Zhihong Zhu, Xuxin Cheng, Zhaorun Chen, Yuyan Chen, Yunyan Zhang, Xian Wu, Yefeng Zheng, and Bowen Xing. 2024. InMu-Net: advancing multi-modal intent detection via information bottleneck and multi-sensory processing. In Proceedings of the 32nd ACM International Conference on Multimedia. 515–524.

[22] Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. 2024. Multimodal Classification and Out-of-distribution Detection for Multimodal Intent Understanding. arXiv preprint arXiv:2412.12453 (2024).

[23] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen, “Generation-augmented retrieval for open-domain question answering,”arXiv preprint arXiv:2009.08553, 2020.

[24] Y.-S. Chuang, W. Fang, S.-W. Li, W.-t. Yih, and J. Glass, “Expand, rerank, and retrieve: Query reranking for open-domain question answering,” arXiv preprint arXiv:2305.17080, 2023.

[25] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in International Conference on Learning Representations, 2022.

[26] Y. Li, Y. Yu, C. Liang, N. Karampatziakis, P. He, W. Chen, and T. Zhao, “Loftq: LoRA-fine-tuning-aware quantization for large language models,” in The Twelfth International Conference on Learning Representations, 2024.26

[27] Z. Hu, L. Wang, Y. Lan, W. Xu, E.-P. Lim, L. Bing, X. Xu, S. Poria, and R. Lee, “Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models,” in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 5254–5276.

[28] X. Lin, W. Wang, Y. Li, S. Yang, F. Feng, Y. Wei, and T.-S. Chua, “Dataefficient fine-tuning for llm-based recommendation,” in Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 365–374.