

Cross-Prompt Encoder for Low-Performing Languages

Beso Mikaberidze[†], Teimuraz Saghinadze[†], Simon Ostermann^{*,+}, Philipp Müller^{*}

[†]Muskhelishvili Institute of Computational Mathematics, Georgian Technical University (MICM),

^{*}Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

⁺ Center for European Research in Trusted AI (CERTAIN)

beso.mikaberidze@gmail.com, philipp.mueller@dfki.de

Abstract

Soft prompts have emerged as a powerful alternative to adapters in parameter-efficient fine-tuning (PEFT), enabling large language models (LLMs) to adapt to downstream tasks without architectural changes or parameter updates. While prior work has focused on stabilizing training via parameter interaction in small neural prompt encoders, their broader potential for transfer across languages remains unexplored. In this paper, we demonstrate that a prompt encoder can play a central role in improving performance on *low-performing languages*—those that achieve poor accuracy even under full-model fine-tuning. We introduce the Cross-Prompt Encoder (XPE), which combines a lightweight encoding architecture with multi-source training on typologically diverse languages—a design that enables the model to capture abstract and transferable patterns across languages. To complement XPE, we propose a Dual Soft Prompt mechanism that combines an encoder-based prompt with a directly trained standard soft prompt. This hybrid design proves especially effective for target languages that benefit from both broadly shared structure and language-specific alignment. Experiments on the SIB-200 benchmark reveal a consistent trade-off: XPE is most effective for low-performing languages, while hybrid variants offer broader adaptability across multilingual settings.

1 Introduction

Cross-lingual task transfer (XLT) seeks to leverage supervision in one or more source languages to enable task generalization to target languages. As highlighted in a recent survey on cross-lingual alignment (Hämmerl et al., 2024), most existing approaches rely on supervising models in a single source language—typically English—before applying them to target languages. In contrast, multi-source training, where models are supervised

on multiple labeled source languages, remains relatively underexplored (Zheng et al., 2021). Yet this setup holds significant promise: By exposing the model to multiple linguistic lenses, it encourages the learning of more robust, language-agnostic representations grounded in shared structural and semantic patterns across diverse languages.

This capability becomes especially important when transferring to *low-resource* target languages—those for which even full-model fine-tuning yields suboptimal results due to a lack of data. These languages often exhibit substantial typological divergence from high-resource counterparts and lack alignment signals that typically aid transfer (Lauscher et al., 2020). In such cases, effective zero-shot transfer remains one of the most persistent challenges in multilingual NLP, as evidenced by the XTREME benchmark, which reveals consistently large performance gaps between English and many typologically diverse target languages across a range of tasks (Hu et al., 2020).

Building on prior works on using *prompt encoders* (Liu et al., 2022b, 2024), we introduce the Cross-Prompt Encoder (XPE), a parameter-efficient multilingual soft prompt tuning approach. XPE integrates a small, reusable neural prompt encoder that learns to enrich fixed-length soft prompts with abstract transferable patterns drawn from multiple, typologically diverse source languages. Unlike most other prompt encoder-based approaches, both the encoder and its input embeddings are static at inference time, retaining the efficiency of standard soft prompting without introducing overhead.

To complement this architecture, we also propose a Dual Soft Prompt (DUAL) mechanism that adds a directly trained standard soft prompt (SPT) alongside the encoder-based prompt. This design enables the model to incorporate both abstract, cross-lingual patterns and more language-specific cues, offering complementary capabilities that can benefit each target language to varying degrees.

Our experiments on the SIB-200 benchmark—covering over 200 languages—demonstrate that XPE achieves strong performance on low-performing and typologically diverse target languages, while DUAL variants excel in other cases settings. Together, these findings highlight the strength of combining multilingual supervision with prompt modularity, enabling efficient XLT across a wide spectrum—from well-aligned to more challenging scenarios.

Our contributions are three-fold:

1. We propose the Cross-Prompt Encoder (XPE), a parameter-efficient method that combines a soft prompt encoder with multi-source training on typologically diverse languages. This setup is designed to enhance cross-lingual task transfer (XLT) by encouraging the model to learn broadly applicable patterns across languages. To our knowledge, this is the first soft prompt encoder specifically tailored for multi-source XLT.
2. We achieve state-of-the-art performance in both zero-shot transfer and full-data scenarios on Sib-200 text-classification task (Adelani et al., 2024). Our method outperforms zero-shot prompting models (e.g., GPT-4), prompt-based ZS-XLT methods (e.g., RoSPrompts), and full-model fine-tuning baselines (e.g., SIB-200) across a wide range of languages. It is especially effective on low-performing languages — those that remain challenging even under direct full-model fine-tuning.
3. We conduct ablation experiments to analyze the strengths of encoder-based and standard soft prompts. Our findings show that XPE is more effective in challenging low-performing scenarios, while standard soft prompts perform better when the source and target languages are closely aligned. Based on this, we introduce a Dual Soft Prompt (DUAL) mechanism that combines both, consistently yielding the best performance across multilingual settings.

2 Related Work

With the rise of LLMs, a new paradigm of PEFT has emerged due to the size of models being fine-tuned (Han et al., 2024; Wang et al., 2025). The general goal in mind is to minimize the number of parameters to be trained while enhancing model

performance above in-context learning and ideally approaching the performance of full-fine tuning (Liu et al., 2022a). After validating its performance in single task / language scenarios PEFTs are often modified to work within multi-language problems (Pfeiffer et al., 2020; Fu et al., 2022b).

2.1 Parameter-efficient Cross-lingual Adaptation

MAD-X (Pfeiffer et al., 2020), based on adapters (Houlsby et al., 2019), is one of the first methods to be successfully extended to multilingual environments. Recently, LoRA (Hu et al., 2021) was extended to cross-lingual scenarios using a method called FLARE (Borchert et al., 2025). One drawback of this method however is that all data points must be paired with their translation in the source language. LT-SFT (Ansell et al., 2022) and its more recent variation DeFT-X (Simon and Jyothi, 2025) use *Lottery Ticket Hypothesis* to employ masks in one case and in another SVD to obtain subnetworks that correspond to task and language separately and combine them to obtain cross-lingual transfer.

Major PEFT branches are viable for cross-lingual transfer, yet their zero-shot capabilities are constrained. A key limitation for approaches like MAD-X, LT-SFT, and DeFT-X is their dependence on language-specific components extracted through masked language modelling. These methods are inaccessible for languages with insufficient or non-existent unlabelled corpora, significantly limiting their utility in resource-scarce settings.

2.2 Soft Prompt Tuning for Cross-lingual Tasks

A recently emerging approach in parameter-efficient adaptation is to find prompts or prefixes using backpropagation, dubbed soft-prompts (Li and Liang, 2021; Lester et al., 2021). Their success in single-task environments inspired researchers to extend soft prompts to multitask and multilingual environments. (Fu et al., 2022b).

Cross-lingual transfer can be achieved through various mechanisms, including the use of a basic soft prompt (Philippy et al., 2024), a Mixture-of-Experts approach in the case of SMoP (Choi et al., 2023), or the introduction of an explicit soft prompt translation mechanism in the case of MPT (Qiu et al., 2024). On the one hand, some researchers argue that the limited number of parameters in soft prompts enhances performance (Philippy et al.,

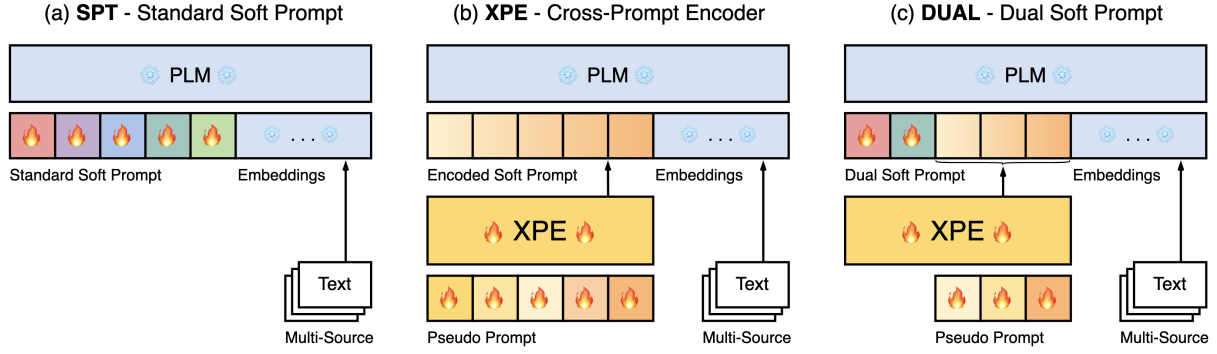


Figure 1: Architectural setup of the three methodologies during training: (a) **SPT** – Standard Soft Prompt, (b) **XPE** – Cross-Prompt Encoder, and (c) **DUAL** – Dual Soft Prompt, a hybrid combining both prior approaches. Fire and snowflake icons indicate trainable and frozen parameters, respectively.

2024). However, in many other cases, some layers increase the parameter count while keeping the width of the injected prompt relatively small (Qiu et al., 2024; Choi et al., 2023).

Soft prompt based approach can be used in zero-shot scenarios; strategies are varied too, including finding a universal prompt across multiple tasks and multiple languages (Fu et al., 2022b) in the case of Polyprompt, tweaking loss and learning procedures, or even a template/context split fusion mechanism for UniPrompt (Huang et al., 2022) and RosPrompt (Philippy et al., 2025). However, the results are difficult to compare, as they all utilize different datasets and do not necessarily employ the same method to select source and target languages.

Out of all the data sets proposed in these articles, SIB-200 contains the largest number of languages and has additional labels, including Joshi’s classification (Adelani et al., 2024). This dataset lets us explore languages usually missing from a model’s pretraining or those that generally underperform. Existing methods have limitations: UniPrompt cannot directly evaluate languages the model hasn’t encountered, and Polyprompt, though interesting, was trained on mT5 (Xue et al., 2021), making direct comparison challenging. What sets our work apart is its direct focus on underperforming languages—a gap, to our knowledge, not addressed by previous research. This distinct focus may explain why RoSPrompt underperforms in comparison to our proposed method.

3 Methodology

To address the challenge of zero-shot cross-lingual transfer (ZS-XLT)—particularly for low-performing languages—we introduce the Cross-Prompt Encoder (XPE) (see the Figure 1(b)), a

parameter-efficient soft prompt tuning method inspired by recent work, including P-Tuning (Liu et al., 2022b) and Multitask Prompt Tuning (MPT) (Wang et al., 2023). XPE consists of a single, reusable neural module that encodes a soft prompt using supervision from multiple typologically diverse source languages. The encoder and its inputs are shared across all languages, and the encoding process induces interactions among those input embeddings. Hence, the encoded soft prompt is able to learn abstract, language-agnostic patterns, thereby enhancing transferability, especially for low-performing and poorly aligned languages. At inference time, the encoded prompt is cached and used directly, preserving the efficiency of standard soft prompt tuning.

To complement this design, we introduce a Dual Soft Prompt (DUAL) mechanism that integrates XPE with an additional, directly trained standard soft prompt (SPT) (see the Figure 1(c)). As the standard soft prompt does not involve a prompt encoder, it is expected to capture more language-specific features, which may assist in transferring to languages seen during backbone model pretraining or those closely aligned with them. The resulting DUAL setup supports robust multilingual transfer across a broad spectrum of languages—ranging from well-aligned to low-performing ones—each may benefit to varying degrees from both components.

3.1 Cross-Prompt Encoder (XPE)

XPE employs a lightweight neural network that maps a small set of learnable input embeddings to outputs with the same hidden dimension as the frozen backbone model. We refer to these inputs as the *pseudo prompt*, and to the network’s output as the *encoded soft prompt*.

Importantly, the prompt encoder and pseudo prompt are used only during training. Once training is complete, the encoder transforms the pseudo prompt into a static encoded soft prompt, which is cached and prepended at inference time—avoiding any additional computation or architectural change.

Formally, the encoder is defined as $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$, where θ denotes the parameters of the encoder module and d is the hidden size of the backbone Transformer, corresponding to its input embedding dimension. While pseudo prompt represents a matrix $n \times d$ where n is the number of embeddings in the pseudo prompt, the encoder can handle only one embedding at a time. Resulting vectors are concatenated later to form the final encoded soft prompt. The overall mapping from the full pseudo prompt to the encoded soft prompt can thus be expressed as $F_\theta : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times d}$

The pseudo prompt $n \times d$ and encoder parameters θ are shared across all source languages.

3.2 Dual Soft Prompt (DUAL)

The DUAL setup integrates XPE and standard SPT approaches, enabling the prompt to combine both encoder-based shared structure and directly learned embeddings. Specifically, we allocate a fixed number of soft prompt embeddings to two components: the first part is dedicated to a standard soft prompt, and the second to an encoded soft prompt. These two segments are concatenated—standard first, followed by encoded—as illustrated in Figure 1 and jointly tuned during training. The full soft prompt is injected at the embedding layer, while the backbone model remains frozen throughout.

Like the encoder parameters and pseudo prompt used in XPE, the standard soft prompt is shared across all source languages. So like XPE, the DUAL setup also produces a static, multilingual soft prompt, which is solely prepended to the input embeddings of the backbone model at inference time. This composite prompt preserves the overall token budget while blending both components. We experiment with two configurations: $\text{DUAL}^{\text{XPE-70}}$ and $\text{DUAL}^{\text{XPE-30}}$. In both variants, the numbers 70 and 30 denote the percentage of soft prompt tokens allocated to the XPE component, with the remaining tokens used for the standard SPT.

4 Experiments

We evaluate our proposed method on the SIB-200 multilingual text classification benchmark, focus-

ing on both zero-shot and fully-supervised XLT scenarios. The experiments are designed to assess the effectiveness of soft prompt tuning methods under diverse multi-source training setups, and to analyze performance across several meaningful target language groups, including the most challenging low-performing languages. All evaluations are conducted per target language, with aggregate results reported at the group level. All models are built upon the XLM-R large encoder and are compared against strong baselines, including full-model fine-tuning and zero-shot prompting. We also perform a detailed ablation study to isolate the contributions of each component.

4.1 Experimental Setup

Our experiments are based on the XLM-R Large model, a transformer encoder pretrained on 100 languages. During training, the backbone remains frozen, and we optimize only a small set of parameters, that include soft prompt related parameters and the transformer classification head. The total number of trainable parameters remains under 0.3% of the full model, enabling highly parameter-efficient transfer learning (PETL).

We evaluate on the SIB-200 benchmark, a multilingual topic classification dataset covering 200 typologically diverse languages.

The general setup of our experiments is a multi-source cross-lingual task transfer (XLT), conducted under two levels of supervision: zero-shot and full. In the zero-shot setting, the model is trained on labeled data from the source languages and directly applied to each target language without any target supervision. In contrast, the fully-supervised setting follows a sequential XLT setup, where the model is first tuned on the multi-source data, then further tuned on labeled data from a single target language before evaluation.

4.2 Sources and Target Language Grouping

To study cross-lingual transfer dynamics under diverse conditions, we define several configurations for both source and target language sets. We use the following source configurations: 1. *EnArZho*: A compact, high-resource, typologically diverse set comprising English, Arabic, and Mandarin Chinese. 2. *Joshi5*: A group of seven most high-resource languages classified as 5 by Joshi et al. (Joshi et al., 2020). 3. *Seen*: The 92 languages that were included in the XLM-R pretraining corpus, representing the model’s seen-language space. Notably,

each small group is a subset of bigger groups.

To better interpret transfer effectiveness, we aggregate results across four target language groups based on their relationship to XLM-R pretraining and downstream performance: 1. *All /wo Joshi5*: All SIB-200 languages excluding the Joshi5 set. 2. *Seen /wo Joshi5*: Consisting of only languages seen during XLM-R pretraining, excluding the Joshi5 set. 3. *Unseen*: Languages not included in XLM-R’s pretraining corpus. 4. *Low-Performing* - We define low-performing languages as those for which XLM-R exhibits poor downstream performance, likely due to limited or ineffective representation during pretraining. Specifically, we identify such languages in SIB-200 by referring to full fine-tuning results on XLM-R large, reported in the original benchmark and selecting those with accuracy below 60%.

We note that the *Seen* and *Unseen* groups form a disjoint partition of the full language set (*Seen* + *Unseen* = *All*), and the *Low-Performing* group is a strict subset of *Unseen*. While this overlap is not enforced by definition, it aligns with expectations that languages unseen during pretraining tend to suffer from lower downstream performance.

For our fully-supervised experiments, we evaluate on a representative subset of 46 target languages—23 from the *seen* group and 23 from the *unseen* group—due to the prohibitive cost of training 200 dedicated models. Languages were selected to ensure diversity across language families, scripts, and resource levels. Although the selection process did not explicitly consider performance tiers, 11 out of the 23 unseen languages in the subset are later identified as low-performing, indicating a fair and challenging distribution.

4.3 Methods Compared

While our focus is on soft prompt-based transfer methods, there are relatively few established baselines for this setting on large-scale multilingual benchmarks like SIB-200.

Our main method is the Cross-Prompt Encoder (XPE), a parameter-efficient soft prompt encoding approach for multilingual transfer. To isolate the contribution of the prompt encoder, we ablate it by removing the encoder from XPE, resulting in *Standard Soft Prompt Tuning (SPT)*, which corresponds to the canonical soft prompt tuning approach widely used in prior work. This variant simultaneously serves as a baseline and a direct ablation of our method.

We additionally evaluate a hybrid setup, *Dual Soft Prompting*, which combines the SPT and XPE components within a fixed prompt budget. This setup preserves the overall number of soft prompt embeddings while blending both prompt types. We experiment with two configurations: $\text{DUAL}^{\text{XPE-70}}$ and $\text{DUAL}^{\text{XPE-30}}$, where 70 and 30 refer to the percentage of prompt embeddings allocated to XPE.

To contextualize the performance of our approach, we compare it against several baselines, including zero-shot prompting with large language models, prompt-based transfer method using a single source language, and full-model fine-tuning on the SIB-200 benchmark. All models—except the zero-shot prompting LLM baselines—are based on the XLM-R Large architecture, just like ours.

For zero-shot cross-lingual transfer (ZS-XLT), we include results from several prompting-based baselines: *Phi-3.2-mini*, *GPT-3.5*, and *GPT-4*, each evaluated in a pure zero-shot setting without any task-specific tuning. We also compare against *RoSPrompt*, a recent method that combines soft and hard prompts using English as the sole source language. While it is also trained on a topic classification task, RoSPrompt uses an auxiliary dataset (DBPedia14) with a different label space from SIB-200, making the setting not fully comparable. Finally, we include the *SIB-200 ZS-XLT* baseline, corresponding to full-model fine-tuning on a single source language (English, Arabic and Chinese), followed by zero-shot evaluation on target languages.

We compare our fully-supervised multi-source XLT approach—based on parameter-efficient tuning—with the monolingual full-model fine-tuning baseline reported in the original SIB-200 benchmark. Both setups involve training a separate model for each target language; however, our method first performs multi-source training before adapting to each target language, enabling knowledge transfer across languages while updating less than 0.3% (1.6M) of the model parameters. In contrast, the SIB-200 baseline trains all model parameters on target-language supervision, without incorporating any cross-lingual signals.

Notably, we report average results over 10 and 6 random seed runs for the zero-shot and fully supervised scenarios, respectively.

4.4 Implementation Details

The soft prompt length is fixed at 20 virtual embeddings. Optimization is performed using Adafactor with a fixed learning rate and a cosine schedule

Target	Source	#Source	SPT	DUAL ^{XPE-30}	DUAL ^{XPE-70}	XPE
LowPerf.	EnArZho	3	35.2	36.5	36.2	35.3
	Joshi5	7	36.0	36.8	37.3	39.1
	Seen	92	39.1	39.3	40.3	41.9
Unseen	EnArZho	3	54.8	56.0	55.6	53.7
	Joshi5	7	56.0	56.4	57.6	57.2
	Seen	92	58.9	59.5	60.1	60.8
Seen /wo J5	EnArZho	3	84.7	84.8	84.6	82.6
	Joshi5	7	85.6	85.3	86.6	84.5
All /wo J5	EnArZho	3	67.7	68.4	68.0	66.2
	Joshi5	7	68.7	68.8	70.0	69.0

Table 1: ZS-XLT performance (accuracy) across different target groups. Each method was trained on different source language groups. Darker yellow indicates better performance (per target group). Note that all models were trained for a fixed maximum of optimization steps, regardless of dataset size. J5 refers to Joshi5 languages.

	ZS Prompting			ZS-XLT						
	Phi-3.5	GPT-3.5	GPT-4	SIB-200			RoS	SPT	DUAL ^{XPE-70}	XPE
	–	–	–	Eng	Ara	Zho	Eng	Joshi5	Joshi5	Seen
LowPerf.	–	22.9	22.9	33.5	33.3	33.3	–	36.0	37.3	41.9
Unseen	–	35.7	39.2	54.0	54.7	54.3	–	56.0	57.6	60.8
Seen /wo J5	49.02	55.7	68.1	86.2	86.5	86.5	67.3	85.6	86.6	–
All /wo J5	–	44.3	51.7	67.8	68.3	68.1	–	68.7	70.0	–

Table 2: Average accuracy across target language groups. The first header row indicates the general setup category, while the next two rows specify the individual methods and their corresponding source language(s). Baselines (Phi-3.5-mini, GPT-3.5, GPT-4, RoSPrompt, and SIB-200) are sourced from prior work, whereas SPT, XPE, and DUAL variants are our trained models. “J5” refers to the Joshi5 language group.

with restarts (2 cycles). For XPE, we use a learning rate of $5e-5$ and weight decay of 0.1 for both the prompt encoder and the classification head. In SPT, only the soft prompt is trained with a higher learning rate of $5e-3$ and no weight decay, while the classification head remains under the same settings as in XPE. The DUAL configuration reflects the same settings applied to its respective components. Training is conducted with a batch size of 32. For all methods and source language configurations, we use a fixed budget of 24,000 optimization steps for source training and 6,000 steps for the target. Early stopping is applied after the first learning rate cycle, with a patience of 20 epochs for source training and 30 for target. All experiments were run on a single NVIDIA A100 GPU, with each training run taking approximately 30 minutes. We use the HuggingFace ecosystem (Wolf et al., 2020) to access the required artefacts, in accordance with the allowed scientific use.

5 Results

We extensively evaluate our proposed XPE and DUAL approaches in zero-shot experiments on the

SIB-200 dataset (Adelani et al., 2024). We furthermore conduct an evaluation in the full fine-tuning scenario on the same dataset.

5.1 Zero-shot Experiments

What mix of SPT and XPE works best? In Table 1 we present results of different soft prompt methods and different combinations of target and source languages in a zero-shot scenario. We compare SPT with XPE as well as the two DUAL variants DUAL^{XPE-70} and DUAL^{XPE-30}. For the challenging set of low-performing languages, XPE achieves the best performance with 41.9 accuracy when training on all 92 seen languages. Decreasing the proportion of XPE embeddings in favour of SPT decreases performance in this scenario. When training on only 7 source languages (Joshi5), the pattern of results is still the same, but when reducing the training languages to 3 (English, Arabic, Mandarin Chinese), utilising a mixture of SPT and XPE is more advantageous. A similar general pattern can be observed when all unseen languages are used as target languages. When training on all 92 seen languages, pure XPE reaches the best

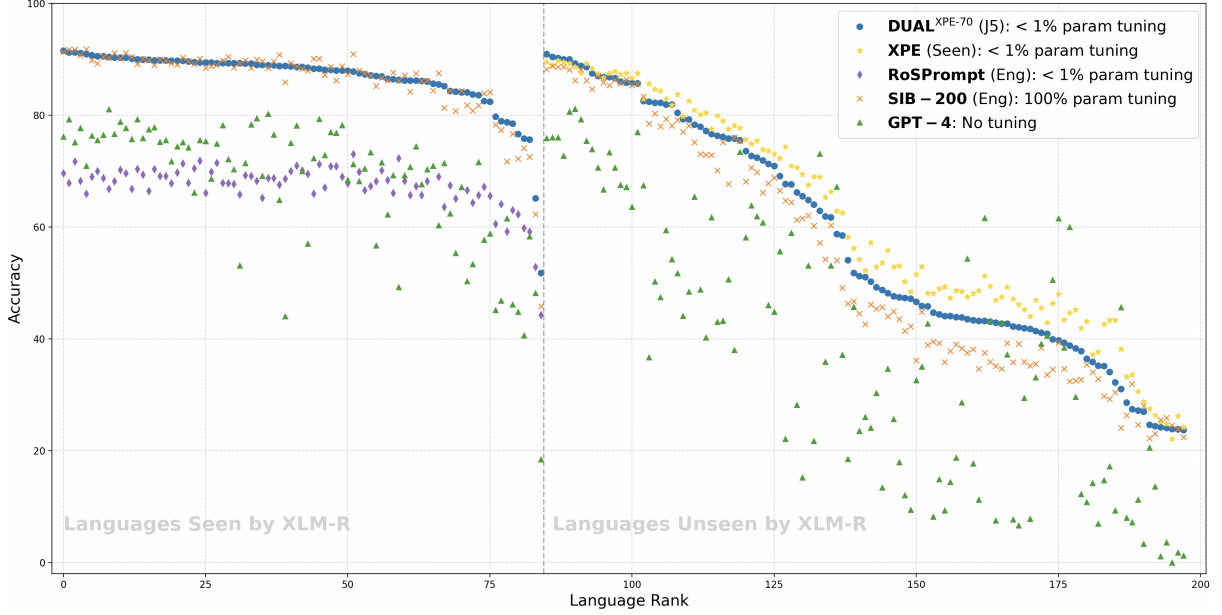


Figure 2: Comparison of different methods on the SIB-200 dataset. We group languages by whether they are seen in the pre-training corpus of XLM-R. Source language groups are provided in parentheses alongside the methods, where J5 refers to the Joshi5 group. Languages are ordered by $\text{DUAL}^{\text{XPE-70}}$ (J5) performance in each group. All the methods are ZS-XLT, except for GPT-4, which is ZS prompting. It should be noted that RoSPrompts used English DBPedia14 as a source dataset, which is a topic detection task, but differs in label space.

performance (60.8 accuracy). With a reduced number of training languages, mixing XPE with SPT becomes advantageous. When considering the less challenging transfer scenarios of seen languages as target (excluding Joshi5), the advantage of mixing SPT with XPE becomes very clear. In this scenario, $\text{DUAL}^{\text{XPE-70}}$ reaches the highest performance with 86.6 accuracy. $\text{DUAL}^{\text{XPE-70}}$ also reaches the highest performance when considering all languages (except Joshi5).

SOTA comparison. Most importantly, our proposed approach $\text{DUAL}^{\text{XPE-70}}$ trained on Joshi5 outperforms all previous works in all target language configurations. Considering all languages except Joshi5 as target languages, $\text{DUAL}^{\text{XPE-70}}$ reaches 70.0 accuracy, followed by SPT (trained on Joshi5, 68.7 accuracy), and SIB-200 (trained on Arabic, 68.3 accuracy). For unseen languages, including the subset of low-performing languages, we reach an even higher performance using pure XPE trained on all seen languages. Here, XPE achieves an accuracy of 60.8, followed by $\text{DUAL}^{\text{XPE-70}}$ with 57.6 accuracy. The best result from previous work is SIB-200 trained on Arabic with 54.7 accuracy. These results underline that our approach is able to effectively integrate training signals from several source languages in challenging cross-lingual task

transfer scenarios. In Figure 2 we present a plot of per-language accuracies for different methods. The improvements made by our proposed approaches are highly consistent. Only for a small number of languages our approaches are outperformed by the generally inferior GPT-4.

5.2 Fully-supervised Experiments

In addition to the zero-shot setting, we also evaluated our proposed approaches in a scenario where supervised data in the target language is available. For computational feasibility, we evaluate on a representative subset of 23 seen and 23 unseen languages, as described in Section 4.2. The results are shown in Table 3. Overall, our $\text{DUAL}^{\text{XPE-70}}$ approaches reaches a slight improvement over SIB-200. When comparing the gains for unseen versus seen target languages, we see that $\text{DUAL}^{\text{XPE-70}}$ particularly excels for unseen languages, whereas it is at a slight disadvantage for seen languages.

6 Discussion

6.1 Generalization vs. Specialization

Our results reveal a consistent performance divide between XPE and SPT across different language types. XPE outperforms SPT on low-performing, typologically diverse target languages, suggesting

Target	#Target	SIB-200	DUAL ^{XPE-70}
Unseen	23	64.0	65.1
Seen	23	88.3	87.6
All	46	76.1	76.3

Table 3: Comparison of fully supervised methods. The direct full-model fine-tuning baseline is sourced from the SIB-200 paper. Our sequential XLT approach uses Joshi5 as the source language group and DUAL^{XPE-70} as the method. Results are reported across target language groups.

that it is better suited for generalization. This likely stems from its encoder-based structure, which encourages abstraction and captures task-relevant patterns that generalize across languages. In contrast, SPT achieves higher scores on seen languages, where alignment with the backbone pretraining data is stronger. This indicates a tendency toward specialization, where SPT embeddings memorize language-specific patterns that can be directly exploited when sufficient overlap exists between training and target data. Two core differences distinguish the behaviors of SPT and XPE. First, in SPT each prompt embedding is updated independently in a position-specific manner, directly based on its role in the sequence. In contrast, XPE embeddings are passed through a single shared encoder, making their representations interdependent and jointly shaped. Second, SPT tokens are supervised directly via the downstream task loss, receiving unmediated gradient signals. XPE, on the other hand, introduces an additional transformation step—updates flow through the encoder, it smooths gradients and generalizes representations.

Our results show that it is beneficial to combine SPT and XPE into a DUAL configuration that is able to perform well across a wide range of source and target language scenarios.

6.2 Language Diversity Shifts the Balance Toward Generalization

Our experiments across multiple source language configurations demonstrate that source language diversity plays a critical role in shaping cross-lingual performance. As the number of source languages increases—from 3 to 7 to 92—the benefits of generalization become more pronounced. Specifically, we observe that models with stronger generalization capacity (e.g., XPE-70 and full XPE) improve consistently with increasing diversity, often surpassing more specialized approaches like SPT. This

pattern holds across both seen and unseen target groups, suggesting that language diversity amplifies the value of language-agnostic task representations. Importantly, these gains occur without exceptions across all source configurations and target groupings. This consistency highlights the universal benefit of source language diversity and supports the claim that generalization becomes increasingly crucial in multilingual transfer.

6.3 Alternative Explanation: Capacity Matching

Although the optimization schedule remains fixed, the number of unique training samples varies across source configurations. One might therefore attribute our findings to a capacity-matching effect: smaller models (e.g., SPT) perform better with less data, while larger ones (e.g., XPE) benefit from greater diversity. However, the evidence instead points to architectural bias. SPT consistently performs best on seen, well-aligned languages, regardless of source diversity and size. In contrast, XPE outperforms SPT on low-performing targets across all configurations. This persistent divide suggests that inductive biases—specialization versus generalization—play a more decisive role than model size or training volume. Additionally, the consistent advantage of the DUAL setup in diverse settings suggests that combining architectural biases is more critical than model capacity alone.

7 Conclusion

We introduce the Cross-Prompt Encoder (XPE), a multi-source parameter-efficient transfer learning (PELT) method that updates less than 0.3% of model parameters while achieving substantial gains in the most challenging setup—zero-shot transfer to low-performing languages. To further boost adaptability, we propose a Dual Soft Prompt mechanism that combines XPE with standard soft prompts, leveraging both abstract, transferable patterns and language-specific memorization. This hybrid design enables robust multilingual transfer across a wide spectrum of target languages, each benefiting to varying degrees from the complementary strengths of both components.

Limitations

This work focuses on a single backbone model (XLM-R), which limits conclusions about the general applicability of XPE and DUAL to other archi-

tures, such as encoder-decoder or decoder-only models. We evaluated our method on the SIB200 dataset. While this dataset has a large variety of languages, it is centred on a single task: multilingual topic classification. Further investigation is needed to assess generalization of our approach across different task types, including reasoning and language generation tasks. Finally, while we explore multilingual transfer, cross-task—and more broadly, universal cross-task and cross-lingual—generalization, as explored in the PolyPrompt paper (Fu et al., 2022a), remains an open direction for future work.

Acknowledgements

This work was partially supported by the European Union under Horizon Europe project "GAIN" (GA #101078950) and by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005).

References

- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. [SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian’s, Malta. Association for Computational Linguistics.
- Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. [Composable sparse fine-tuning for cross-lingual transfer](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.
- Philipp Borchert, Ivan Vulić, Marie-Francine Moens, and Jochen De Weerd. 2025. [Language fusion for parameter-efficient cross-lingual transfer](#). *Preprint*, arXiv:2501.06892.
- Joon-Young Choi, Junho Kim, Jun-Hyung Park, Wing-Lam Mok, and SangKeun Lee. 2023. [SMoP: Towards efficient and effective prompt tuning with sparse mixture-of-prompts](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14306–14316, Singapore. Association for Computational Linguistics.
- Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022a. [Polyglot prompt: Multilingual multitask prompt training](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9919–9935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022b. [Polyglot prompt: Multilingual multitask prompt training](#). *EMNLP*.
- Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. [Understanding cross-lingual Alignment—A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.
- Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. [Parameter-efficient fine-tuning for large models: A comprehensive survey](#). *Preprint*, arXiv:2403.14608.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org.
- Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. [Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11488–11497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. [From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The power of scale for parameter-efficient prompt tuning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

- pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Motta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. [P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2024. [Gpt understands, too](#). *AI Open*, 5:208–215.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. [MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.
- Fred Philippy, Siwen Guo, Shohreh Haddadan, Cedric Lothritz, Jacques Klein, and Tegawendé F. Bissyandé. 2024. [Soft prompt tuning for cross-lingual transfer: When less is more](#). In *Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024)*, pages 7–15, St Julians, Malta. Association for Computational Linguistics.
- Fred Philippy, Siwen Guo, Cedric Lothritz, Jacques Klein, and Tegawendé Bissyandé. 2025. [Enhancing small language models for cross-lingual generalized zero-shot classification with soft prompt tuning](#). In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 61–75, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xiaoyu Qiu, Yuechen Wang, Jiaxin Shi, Wengang Zhou, and Houqiang Li. 2024. [Cross-lingual transfer for natural language inference via multilingual prompt translator](#). In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.
- Sona Elza Simon and Preethi Jyothi. 2025. [Deftx: De-noised sparse fine-tuning for zero-shot cross-lingual transfer](#). *Preprint*, arXiv:2505.15090.
- Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. 2025. [Parameter-efficient fine-tuning in large models: A survey of methodologies](#). *Preprint*, arXiv:2410.19878.
- Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogério Feris, Huan Sun, and Yoon Kim. 2023. [Multitask prompt tuning enables parameter-efficient transfer learning](#). *Preprint*, arXiv:2303.02861.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. [Consistency regularization for cross-lingual fine-tuning](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417, Online. Association for Computational Linguistics.