用于低表现语言的跨提示编码器

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

软提示在参数高效微调（PEFT）中作为适配器的一种强大替代方法出现，使大型语言模型（LLM）能够在不改变架构或更新参数的情况下适应下游任务。虽然先前的研究已专注于通过小型神经提示编码器中的参数交互来稳定训练，但它们在跨语言转移的更广泛潜力仍未被探索。在本文中，我们展示了一个提示编码器可以在提升低表现语言的性能方面发挥核心作用——这些语言即使在全模型微调的情况下也无法取得良好准确性。我们引入了跨提示编码器（XPE），其结合了一个轻量编码架构与对类型学多样语言的多源训练——这一设计使模型能够捕获跨语言的抽象和可转移的模式。为了补充 XPE，我们提出了一种双软提示机制，该机制将基于编码器的提示与一个直接训练的标准软提示结合。这种混合设计被证明对那些需要广泛共享结构和语言特定对齐的目标语言特别有效。在 SIB-200 基准测试中的实验揭示了一个一致的权衡：XPE 对低表现语言最有效，而混合变体在多语言环境下提供更广泛的适应性。

1 介绍

跨语言任务迁移（XLT）旨在利用一种或多种源语言中的监督来实现任务向目标语言的泛化。正如最近关于跨语言对齐的调查中所强调的（Hämmerl et al., 2024），大多数现有方法依赖于在单一源语言（通常是英语）中对模型进行监督，然后将其应用于目标语言。相比之下，多源训练，即对多个标注的源语言进行监督的模型训练，相对而言仍然研究不足（Zheng et al., 2021）。然而，这种设置具有显著潜力：通过让模型接触多种语言视角，它鼓励学习更稳健的、与语言无关的表示，这些表示是基于跨多种语言的共享结构和语义模式。

这种能力在转移到低资源目标语言时变得尤为重要——对于这些语言来说，即使是完整模型微调也会由于缺乏数据而产生次优结果。这些语言通常在类型学上与高资源语言有显著的差异，并且缺乏通常有助于转移的对齐信号（Lauscher et al., 2020）。在这种情况下，有效的零样本转移仍然是多语言自然语言处理领域中最持久的挑战之一，如 XTREME 基准所显示的那样，它揭示了在各种任务中，英语与许多类型学多样的目标语言之间始终存在较大的性能差距（Hu et al., 2020）。

基于之前的关于使用提示编码器的工作（Liu et al., 2022b, 2024），我们引入了交叉提示编码器 (XPE)，这是一种参数高效的多语言软提示微调方法。XPE 集成了一个小型、可重用的神经提示编码器，该编码器能够学习从多个语言类型多样的源语言中提取抽象的可迁移模式，以丰富固定长度的软提示。与大多数其他基于提示编码器的方法不同，在推断时，编码器及其输入嵌入都是静态的，从而在不引入额外开销的情况下保留了标准软提示的效率。

为了补充这一架构，我们还提出了一种双软提示（DUAL）机制，该机制在基于编码器的提示旁边添加了一个直接训练的标准软提示（SPT）。这一设计使得模型能够结合抽象的跨语言模式和更多特定语言的线索，提供互补的能力，从而在不同程度上有益于各个目标语言。

我们在 SIB-200 基准测试上的实验（涵盖超过 200 种语言）表明，XPE 在表现较差和类型多样的目标语言上表现出色，而 DUAL 变体在其他情况下表现出色。这些发现共同强调了多语言监督与提示模块化结合的优势，使得在从对齐良好的场景到更具挑战性的场景中实现高效的 XLT。

我们的贡献有三个方面：

1. 我们提出了 Cross-Prompt Encoder (XPE)，这是一种参数高效的方法，将软提示词编码器与多源训练结合在一起，涵盖各种类型语言。这种设置旨在通过鼓励模型学习跨语言广泛适用的模式来增强跨语言任务迁移（XLT）。据我们所知，这是第一个专门为多源 XLT 量身定制的软提示词编码器。

www.xueshuxiangzi.com

2. 我们在 Sib-200 文本分类任务的零样本迁移和全数据场景中实现了最先进的性能（Adelani et al., 2024）。我们的方法在多种语言下优于零样本提示模型（例如，GPT-4）、基于提示的 ZS-XLT 方法（例如，RoSPrompts）和全模型微调基线（例如，SIB-200）。它在低表现语言上尤其有效——这些语言即使在直接的全模型微调下仍然具有挑战性。

3. 我们进行了消融实验，以分析基于编码器的提示和标准软提示的优缺点。我们的研究发现，XPE 在具有挑战性且表现不佳的场景中更加有效，而当源语言与目标语言紧密对齐时，标准软提示表现更好。在此基础上，我们引入了一种结合两者的双软提示（DUAL）机制，在多语言环境下始终表现最佳。

2 相关工作

随着大型语言模型（LLMs）的兴起，由于模型体积在进行微调时的大小，出现了一种新的 PEFT 范式（Han et al., 2024; Wang et al., 2025）。总体目标是最小化需训练的参数数量，同时提升模型性能超越上下文学习，理想情况下接近全微调的性能（Liu et al., 2022a）。在验证其在单任务/语言场景中的性能之后，PEFTs 通常被修改以适应多语言问题（Pfeiffer et al., 2020; Fu et al., 2022b）。

2.1 参数高效的跨语言适应

基于适配器（Houlsby et al., 2019）的 MAD-X（Pfeiffer et al., 2020）是第一个成功扩展到多语言环境的方法之一。最近，LoRA（Hu et al., 2021）通过一种名为 FLARE（Borchert et al., 2025）的方法扩展到了跨语言场景。然而，这种方法的一个缺点是所有数据点必须与源语言的翻译配对。LT-SFT（Ansell et al., 2022）及其最近的变体 DeFT-X（Simon and Jyothi, 2025）使用彩票票据假设来使用掩码，在一种情况下，在另一种情况下使用 SVD，以获得分别对应于任务和语言的子网络，并将它们组合以实现跨语言迁移。

主要的 PEFT 分支对于跨语言迁移是可行的，但它们的零样本能力受到限制。像 MAD-X、LT-SFT 和 DeFT-X 这样的方法的一个关键限制是它们依赖于通过掩码语言建模提取的语言特定组件。这些方法对于未标注语料库不足或不存在的语言是不可用的，这极大地限制了它们在资源匮乏环境中的实用性。

2.2 跨语言任务的软提示微调

参数高效适配中一个最近出现的方法是使用反向传播来找到提示或前缀，被称为软提示（Li and Liang, 2021; Lester et al., 2021）。它们在单任务环境中的成功启发了研究人员将软提示扩展到多任务和多语言环境中。（Fu et al., 2022b）。

跨语言迁移可以通过多种机制实现，包括使用基本的软提示（Philippy et al., 2024）、在 SMoP（Choi et al., 2023）的情况下采用专家混合方法，或在 MPT（Qiu et al., 2024）的情况下引入显式的软提示翻译机制。一方面，一些研究人员认为，软提示中有限的参数数量提高了性能（Philippy et al., 2024）。然而，在许多其他情况下，一些层会增加参数数量，同时保持注入提示的宽度相对较小（Qiu et al., 2024; Choi et al., 2023）。

基于软提示的方法可以用于零样本场景；策略也各不相同，包括在多个任务和多种语言中找到通用提示（Fu et al., 2022b）（如在 Polyprompt 的情况下），调整损失和学习程序，甚至 UniPrompt（Huang et al., 2022）和 RosPrompt（Philippy et al., 2025）的模板/上下文分裂融合机制。然而，由于它们都使用不同的数据集，并不一定采用相同的方法选择源语言和目标语言，结果难以进行比较。

在这些文章中提出的所有数据集中，SIB-200 包含最多种类的语言，并具有额外的标签，包括乔希的分类（Adelani et al., 2024）。该数据集让我们能够探索通常在模型的预训练中缺失的语言或通常表现不佳的语言。现有的方法有其局限性：UniPrompt 无法直接评估模型未遇到的语言，而 Polyprompt 虽然有趣，但其训练基于 mT5（Xue et al., 2021），导致直接比较变得困难。我们的工作与众不同之处在于其直接关注于表现不佳的语言——这是截至我们所知，以前的研究没有解决的一个空白。这个独特的关注点可能解释了为什么 RoSPrompt 在与我们提议的方法相比时表现不佳。

3 方法论

为了应对零样本跨语言迁移（ZS-XLT）的挑战，尤其是对于表现较差的语言，我们引入了跨提示编码器 (XPE)（见图 1 (b)），一种受近期工作启发的参数高效的软提示微调方法，包括 P-Tuning（Liu et al., 2022b) 和多任务提示微调 (MPT)（Wang et al., 2023）。XPE 由一个单一、可重复使用的神经模块组成，利用多个语言学多样的源语言的监督来编码软提示。编码器及其输入在所有语言间共享，而编码过程会引发这些输入嵌入之间的交互。因此，被编码的软

www.xueshuxiangzi.com

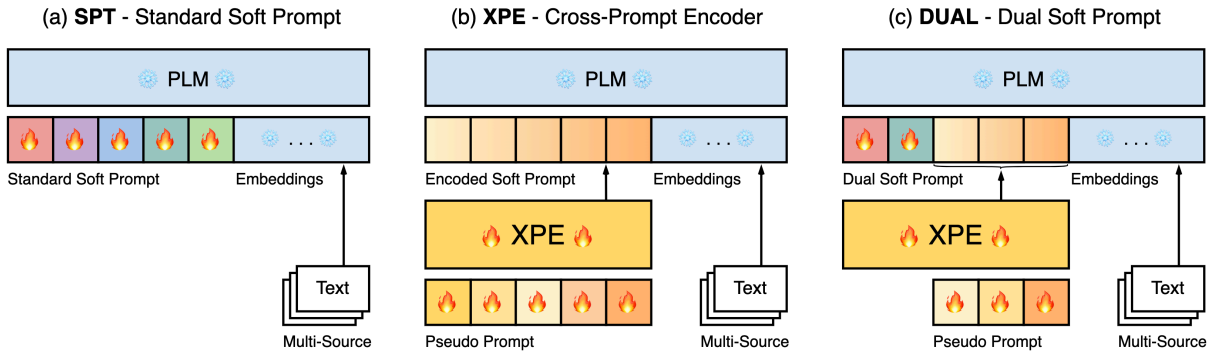

Figure 1: 三种方法在训练过程中的架构设置：(a) SPT –标准软提示，(b) XPE –交叉提示编码器，(c) DUAL –双重软提示，这是一种结合了前两种方法的混合式方法。火焰和雪花图标分别表示可训练参数和冻结参数。

提示能够学习抽象的、与语言无关的模式，从而增强迁移能力，特别是对表现不佳和对齐不良的语言。在推理时，编码后的提示被缓存并直接使用，保留了标准软提示微调的效率。

为了补充这个设计，我们引入了一种双软提示（DUAL）机制，将 XPE 与一个额外的、直接训练的标准软提示（SPT）结合（见图 1 (c)）。由于标准软提示不涉及提示编码器，预计它可以捕捉到更多语言特定的特征，这可能有助于向主干模型预训练时所见的语言或与之紧密相关的语言进行迁移。由此产生的 DUAL 设置支持在多种语言之间的稳健多语言迁移——从高度匹配的到表现较差的语言——每种语言可能从这两个组件中受益于不同的程度。

3.1 跨提示编码器 (XPE)

XPE 使用一个轻量神经网络，该网络将一小组可学习的输入嵌入映射到与冻结骨干模型相同的隐藏维度的输出。我们称这些输入为伪提示词，称网络的输出为编码的软提示词。

重要的是，提示编码器和伪提示仅在训练过程中使用。一旦训练完成，编码器将伪提示转换为一个静态编码的软提示，该软提示被缓存并在推理时预先添加，从而避免任何额外的计算或架构变化。

形式上，编码器被定义为 $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$，其中 θ 表示编码器模块的参数，d 是主干 Transformer 的隐藏层大小，对应于其输入嵌入维度。而伪提示表示一个矩阵 $n \times d$，其中 n 是伪提示中的嵌入数量，编码器一次只能处理一个嵌入。生成的向量稍后被连接以形成最终编码的软提示。从完整的伪提示到编码的软提示的整体映射可以表示为 $F_\theta : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$。伪提示 $n \times d$ 和编码器参数 θ 在所有源语言中共享。

3.2 双重软提示（DUAL）

DUAL 设置整合了 XPE 和标准 SPT 方法，使提示能够结合基于编码器的共享结构和直接学习的嵌入。具体来说，我们为两个组件分配了固定数量的软提示嵌入：第一部分用于标准软提示，第二部分用于编码的软提示。这两个部分被拼接在一起——先是标准部分，然后是编码部分，如图 1 所示，并在训练中进行联合调优。完整的软提示被注入到嵌入层，而主干模型在整个过程中保持冻结。

与 XPE 中使用的编码器参数和伪提示相似，标准的软提示在所有源语言中都是共享的。因此，与 XPE 一样，DUAL 设置也产生一个静态的多语言软提示，该提示仅在推理时被前置添加到基础模型的输入嵌入中。这个复合提示在融合两个组件的同时保持了整体的 token 预算。我们尝试了两种配置：$\text{DUAL}^{\text{XPE-70}}$ 和 $\text{DUAL}^{\text{XPE-30}}$。在这两种变体中，数字 70 和 30 表示分配给 XPE 组件的软提示 token 的百分比，其余 token 用于标准 SPT。

4 实验

我们在 SIB-200 多语言文本分类基准上评估我们提出的方法，重点关注零样本和全监督 XLT 场景。实验旨在评估软提示调优方法在多源训练设置下的有效性，并分析在几个有意义的目标语言组中的性能，包括最具挑战性的低表现语言。所有评估均按目标语言进行，汇总结果在组级别报告。所有模型均基于 XLM-R 大编码器构建，并与强基线进行比较，包括全模型微调和零样本提示。我们还进行了详细的消融研究，以隔离每个组成部分的贡献。

我们的实验基于 XLM-R 大型模型，这是一种在 100 种语言上预训练的 transformer 编码器。在训练期间，主干网络保持冻结状态，我们只优化一小部分参数，包括软提示相关参数和 transformer 分类头。可训练参数的总数量保

www.xueshuxiangzi.com

持在整个模型的 0.3 % 以下，从而实现高效的参数迁移学习（PETL）。

我们在 SIB-200 基准上进行评估，该基准是一个多语言主题分类数据集，涵盖 200 种类型多样的语言。

我们实验的总体设置是一个多源跨语言任务迁移 (XLT)，在两个监督级别下进行：零样本和完全监督。在零样本设置下，模型在源语言的标记数据上进行训练，并在没有任何目标监督的情况下直接应用于每个目标语言。相比之下，完全监督设置遵循顺序 XLT 设置，模型首先在多源数据上进行调优，然后在单个目标语言的标记数据上进一步调优再进行评估。

为了研究多种条件下的跨语言迁移动态，我们为源语言和目标语言集定义了几种配置。我们使用以下源配置：1. EnArZho：一个紧凑的、高资源的、类型多样的集合，包括英语、阿拉伯语和中文普通话。2. Joshi5：由 Joshi 等人分类为 5 的七种高资源语言组成的组。3. Seen：包含在 XLM-R 预训练语料库中的 92 语言，代表模型的已见语言空间。值得注意的是，每个小组都是较大组的子集。

为了更好地解释迁移效果，我们依据与 XLM-R 预训练和下游性能的关系，将结果聚合到四个目标语言组中：1. All /wo Joshi5：所有 SIB-200 语言，不包括 Joshi5 集合。2. Seen /wo Joshi5：仅包含在 XLM-R 预训练中看到的语言，不包括 Joshi5 集合。3. Unseen：不包含在 XLM-R 预训练语料库中的语言。4. Low-Performing - 我们将低表现的语言定义为 XLM-R 在下游表现不佳的语言，这很可能是由于在预训练期间有限或无效的表示。具体来说，我们通过参考 XLM-R large 完整微调结果，并在原始基准中选择那些准确率低于 60% 的语言，在 SIB-200 中识别出这些语言。

我们注意到，已见（Seen）和未见（Unseen）组构成了整个语言集合的一个不相交的分区（Seen + Unseen = 全部），而低性能组是未见组的一个真子集。虽然这种重叠不是通过定义强制的，但它符合预期：在预训练中未见的语言往往在后续任务中表现较差。

在我们的全监督实验中，由于训练 200 个专用模型的成本过高，我们对 46 种目标语言的一个代表性子集进行了评估，其中 23 种来自已见组，23 种来自未见组。选择这些语言是为了确保在语言家族、文字和资源水平上的多样性。尽管选择过程没有明确考虑性能层级，但子集中 23 种未见语言中有 11 种后来被识别为低性能语言，表明了一个公平且具有挑战性的分布。

4.1 方法比较

虽然我们的重点是基于软提示的迁移方法，但在像 SIB-200 这样的大规模多语言基准测试中，这种情况下相对较少有成熟的基线。

我们的方法主要是交叉提示编码器（XPE），这是一种面向多语言转移的参数高效的软提示编码方法。为了隔离提示编码器的贡献，我们通过从 XPE 中移除编码器来对其进行消融，得到标准软提示调优（SPT），这对应于以往工作中广泛使用的经典软提示调优方法。这个变体同时作为基线和我们方法的直接消融。

我们还评估了一个混合设置，称为双软提示，它在固定的提示预算内结合了 SPT 和 XPE 组件。此设置在保留总体软提示嵌入数量的同时，融合了两种提示类型。我们尝试了两种配置：$\text{DUAL}^{\text{XPE-70}}$ 和 $\text{DUAL}^{\text{XPE-30}}$，其中 70 和 30 表示分配给 XPE 的提示嵌入百分比。

为了使我们的方法表现更具上下文性，我们将其与多个基线进行比较，包括使用大型语言模型的零样本提示、使用单一源语言的基于提示的迁移方法，以及在 SIB-200 基准上的完整模型微调。所有模型——除了零样本提示 LLM 基线——都基于 XLM-R Large 架构，就像我们的模型一样。

对于零样本跨语言迁移（ZS-XLT），我们包括了来自多个基于提示的方法的结果：Phi-3.2-mini、GPT-3.5 和 GPT-4，分别在纯零样本设置下评估，未进行任何特定任务的调整。我们还与 RoSPrompt 进行比较，这是一种最近的方法，结合了软提示和硬提示，仅使用英语作为唯一的源语言。虽然它也是在一个主题分类任务上训练的，但 RoSPrompt 使用一个辅助数据集（DBPedia14），它的标签空间与 SIB-200 不同，因此这种设置不完全具有可比性。最后，我们包括了 SIB-200 ZS-XLT 的基线，其中涉及在单一源语言（英语、阿拉伯语和中文）上进行完整模型微调，然后在目标语言上进行零样本评估。

我们比较了我们基于参数高效微调的全监督多源 XLT 方法与原始 SIB-200 基准中报告的单语全模型微调基线。两种设置都涉及为每种目标语言训练一个独立的模型；然而，我们的方法首先进行多源训练，然后再适应每种目标语言，从而在更新不到 0.3 % (1.6M) 的模型参数的同时实现跨语言知识转移。相比之下，SIB-200 基线在目标语言监督下训练所有模型参数，而没有结合任何跨语言信号。

值得注意的是，我们分别报告了零样本和完全监督场景分别在 10 次和 6 次随机种子运行中的平均结果。

www.xueshuxiangzi.com

Target	Source	# Source	SPT	DUAL$^{\text{XPE-30}}$	DUAL$^{\text{XPE-70}}$	XPE
LowPerf.	EnArZho	3	35.2			35.3
	Joshi5	7				
	Seen	92				
Unseen	EnArZho	3	54.8			53.7
	Joshi5	7				
	Seen	92				
Seen /wo J5	EnArZho	3				82.6
	Joshi5	7				
All /wo J5	EnArZho	3				66.2
	Joshi5	7				

Table 1: ZS-XLT 在不同目标组上的表现（准确率）。每种方法都在不同的源语言组上进行了训练。较深的黄色表示较好的表现（针对每个目标组）。请注意，所有模型都在固定的最大优化步骤下进行训练，而不考虑数据集的大小。J5 指的是 Joshi5 语言。

	ZS Prompting			ZS-XLT						
	Phi-3.5	GPT-3.5	GPT-4		SIB-200		RoS	SPT	DUAL$^{\text{XPE-70}}$	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: 目标语言组的平均准确率。第一行标题表示总体设置类别，而接下来的两行则指定了各个方法及其对应的源语言。基线（Phi-3.5-mini、GPT-3.5、GPT-4、RoSPrompt 和 SIB-200）来源于之前的研究，而 SPT、XPE 和 DUAL 变体是我们训练的模型。“J5”指的是 Joshi5 语言组。

4.2 实现细节

软提示长度固定为 20 个虚拟嵌入。优化使用恒定的学习率和带重启的余弦调度（2 个周期）的 Adafactor 进行。对于 XPE，我们对提示编码器和分类头都使用学习率 5e−5 和权重衰减 0.1 。在 SPT 中，仅训练软提示，使用更高的学习率 5e−3 且无权重衰减，而分类头仍保持与 XPE 相同的设置。DUAL 配置反映了其相应组件应用的相同设置。训练以批量大小 32 进行。对于所有方法和源语言配置，我们使用固定的预算：源训练 24,000 次优化步骤，目标训练 6,000 次。在第一个学习率周期之后应用提前停止，对源训练有 20 个 epoch 的耐心，对目标有 30 个 epoch。所有实验均在单个 NVIDIA A100 GPU 上运行，每次训练大约需要 30 分钟。我们使用 HuggingFace 生态系统 (Wolf et al., 2020) 访问所需的资料，符合允许的科学用途。

5 结果

我们在 SIB-200 数据集上进行零样本实验，广泛评估我们提出的 XPE 和 DUAL 方法 (Adelani et al., 2024) 。此外，我们还在同一数据集上进行完整微调场景的评估。

5.1 零样本实验

在 Table 1 中，我们呈现了在零样本场景中不同软提示方法和目标与源语言的不同组合的结果。我们将 SPT 与 XPE 以及两种 DUAL 变体 DUAL$^{\text{XPE-70}}$ 和 DUAL$^{\text{XPE-30}}$ 进行比较。对于表现较差语言的挑战性集合，XPE 在训练所有 92 种已见语言时实现了 41.9 的最佳准确性。在这种情况下，减少 XPE 嵌入的比例而增加 SPT 会降低性能。当仅在 7 个源语言（Joshi5）上进行训练时，结果模式仍然相同，但当将训练语言减少到 3 个（英语、阿拉伯语、普通话）时，使用 SPT 和 XPE 的混合更加有利。类似的一般模式可以观察到，当所有未见语言被用作目标语言时也是如此。当在所有 92 种已见语言上进行训练时，纯 XPE 达到了最佳性能（60.8 的准确性）。随着训练语言数量的减少，XPE 与 SPT 的混合变得有利。当考虑已见语言作为目标的较不具挑战性的迁移场景时（不包括 Joshi5），SPT 和 XPE 的混合优势变得非常明显。在这种情况下，DUAL$^{\text{XPE-70}}$ 以 86.6 的准确率达到最高性能。当考虑所有语言（除 Joshi5 外）时，DUAL$^{\text{XPE-70}}$ 也达到了最高性能。

SOTA 比较。 最重要的是，我们提出的在 Joshi5 上训练的 DUAL$^{\text{XPE-70}}$ 方法在所有目

www.xueshuxiangzi.com

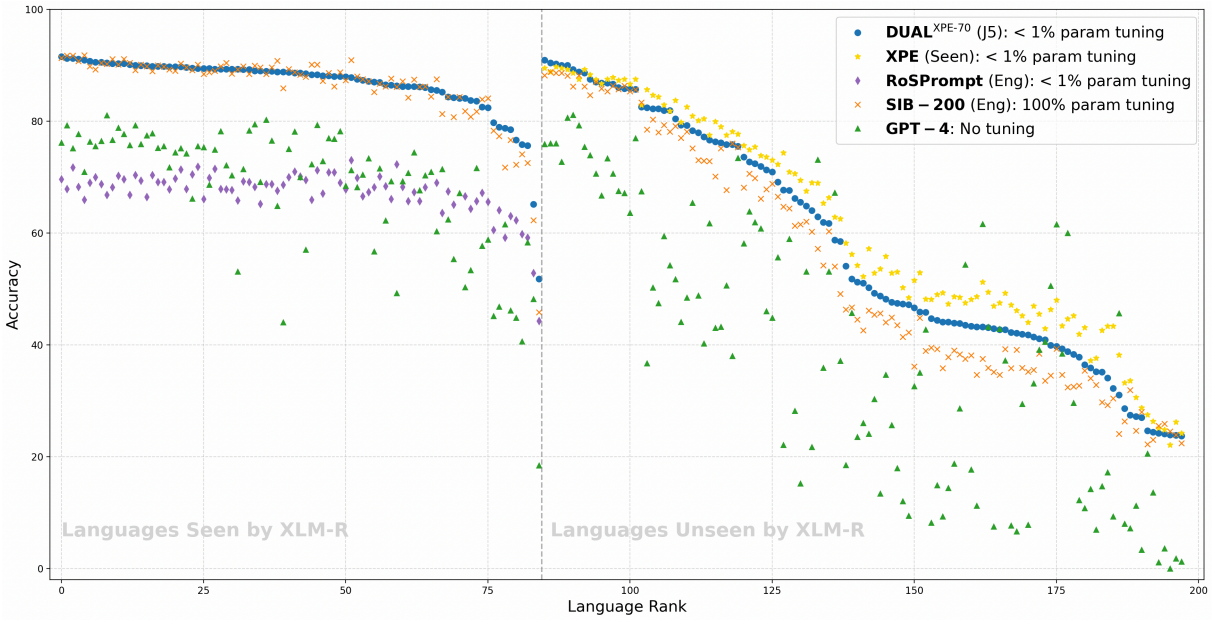

Figure 2: 在 SIB-200 数据集上比较不同方法。我们将语言根据它们是否在 XLM-R 的预训练语料库中出现进行分组。方法旁边括号中提供了源语言组，其中 J5 指的是 Joshi5 组。每组中的语言按 DUAL$^{\text{XPE-70}}$ (J5) 性能排序。所有方法都是 ZS-XLT，除了 GPT-4，它是 ZS 提示。需要注意的是，RoSPrompts 使用英文 DBPedia14 作为源数据集，虽然它是一个主题检测任务，但在标签空间上有所不同。

标语言配置中表现优于所有之前的工作。考虑除 Joshi5 以外的所有语言作为目标语言，DUAL$^{\text{XPE-70}}$ 达到了 70.0 的准确率，紧随其后的是 SPT（在 Joshi5 上训练，68.7 的准确率），以及 SIB-200（在阿拉伯语上训练，68.3 的准确率）。对于未见过的语言，包括表现较差语言的子集，我们在使用对所有已见语言进行训练的纯 XPE 时，达到了更高的性能。此处，XPE 实现了 60.8 的准确率，之后是 DUAL$^{\text{XPE-70}}$ 的 57.6 的准确率。之前工作的最佳结果是 SIB-200 在阿拉伯语上训练的 54.7 的准确率。这些结果表明，我们的方法能够在具有挑战性的跨语言任务转移场景中有效整合来自多个源语言的训练信号。在 Figure 2 中，我们展示了不同方法的单语言准确率图表。我们提出的方法所带来的改进是一致性的。只有极少数语言，我们的方法被通常表现不佳的 GPT-4 超过。

5.2 全监督实验

除了零样本设定外，我们还在目标语言中有监督数据可用的情况下评估了我们提出的方法。为了计算上的可行性，我们在第 **??** 节中描述的 23 种已见语言和 23 种未见语言的代表性子集上进行了评估。结果显示在 Table 3 中。总体而言，我们的 DUAL$^{\text{XPE-70}}$ 方法比 SIB-200 略有改善。当比较未见与已见目标语言的增益时，我们发现 DUAL$^{\text{XPE-70}}$ 尤其在未见语言上表现突出，而在已见语言上略处于劣势。

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

Table 3: 全监督方法的比较。直接的全模型微调基线来源于 SIB-200 论文。我们的顺序 XLT 方法使用 Joshi5 作为源语言组，并使用 DUAL$^{\text{XPE-70}}$ 作为方法。结果是针对目标语言组进行报告的。

6 讨论

6.1 泛化与专业化

我们的结果揭示了 XPE 和 SPT 在不同语言类型之间的一致性能差异。XPE 在表现较低、类型多样的目标语言上表现优于 SPT，这表明它更适合用于泛化。这可能源于其基于编码器的结构，鼓励抽象并捕捉在各语言中泛化的任务相关模式。相比之下，SPT 在已见过的语言上获得更高的分数，与骨干预训练数据的对齐更为紧密。这表明它倾向于专门化，SPT 嵌入记忆可直接利用的语言特定模式，当训练数据和目标数据之间存在足够的重叠时，这种模式尤为明显。SPT 和 XPE 的行为有两个核心区别。首先，在 SPT 中，每个提示嵌入在特定位置是独立更新的，直接基于其在序列中的角色。相比之下，XPE 嵌入通过一个共同的编码器，使其表示相互依赖且共同成形。其次，SPT 通过下游任务损失直接受到监督，接收未经中介的

www.xueshuxiangzi.com

梯度信号。而 XPE 另一方面引入了一个额外的转换步骤，更新流经编码器，这使得梯度平滑并泛化表示。

我们的结果表明，将 SPT 和 XPE 结合成 DUAL 配置在各种源语言和目标语言场景中都能够表现良好是有益的。

6.2 语言多样性改变了向泛化倾斜的平衡

我们的实验覆盖了多个源语言配置，表明源语言的多样性在塑造跨语言性能时起到了关键作用。随着源语言数量的增加——从 3 到 7 再到 92——泛化的好处变得更加明显。具体来说，我们观察到具有较强泛化能力的模型（例如 XPE-70 和完整的 XPE）随着多样性的增加而持续改进，往往超过更为专注的方法，如 SPT。这种模式在看过的和未看过的目标组中都成立，表明语言多样性放大了语言无关任务表示的价值。重要的是，这些收益在所有源配置和目标分组中没有例外地出现。这种一致性凸显了源语言多样性的普遍益处，并支持泛化在多语言迁移中变得愈发重要的说法。

6.3 另一种解释：容量匹配

虽然优化计划保持不变，但不同源配置中的独特训练样本数量会有所不同。因此，有人可能会将我们的发现归因于容量匹配效应：较小的模型（例如 SPT）在数据较少时表现更佳，而较大的模型（例如 XPE）则从更大的多样性中受益。然而，证据却指向结构偏差。无论源的多样性和大小如何，SPT 在已知的、对齐良好的语言上始终表现最佳。相反，XPE 在所有配置中均在低表现目标上优于 SPT。这种持续的差异表明，归纳偏差——专业化与泛化——比模型大小或训练量发挥了更为决定性的作用。另外，DUAL 设置在多样化环境中的持续优势表明，结合结构偏差比单纯的模型容量更为重要。

7 结论

我们引入了交叉提示编码器（XPE），这是一种多源参数高效转移学习（PELT）方法，在更新不到 0.3 % 的模型参数的同时，在最具挑战性的环境中——对表现较低的语言进行零样本转移——实现了显著的提升。为了进一步增强适应性，我们提出了一种双软提示机制，将 XPE 与标准软提示结合起来，利用抽象的、可转移的模式和特定语言的记忆。这个混合设计使得跨多个目标语言的强大多语言传递成为可能，每种语言在不同程度上受益于这两个组件互补优势的结合。

8

局限性

这项工作集中在一个单一的主干模型（XLM-R）上，这限制了关于 XPE 和 DUAL 对其他架构如编码器-解码器或仅解码器模型的普遍适用性的结论。我们在 SIB200 数据集上评估了我们的方法。虽然该数据集包含多种语言，但它集中于一个任务：多语言主题分类。需要进一步的研究来评估我们的方法在不同任务类型（包括推理和语言生成任务）上的泛化能力。最后，虽然我们探索了多语言迁移，但跨任务——更广泛地说，跨任务和跨语言的通用化——正如 PolyPrompt 论文中所探讨的，依然是未来研究的一个开放方向。

9

致谢 本工作部分由欧盟地平线欧洲项目“GAIN”（GA # 101078950）资助，以及德国联邦研究、技术和空间部（BMFTR）作为项目 TRAILS (01IW24005) 的一部分。

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 Vuli, Marie-Francine Moens, and Jochen De Weerdt. 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

www.xueshuxiangzi.com

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 promptraining. 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 Mohta, Tenghao Huang, Mohit Bansal, and Colin A 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

www.xueshuxiangzi.com

prompt translator. In 2024 IEEE International Conference on Multimedia and Expo (ICME) , pages 1–6.

Sona Elza Simon and Preethi Jyothi. 2025. Deftx: Denoised 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, Rogerio 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.

www.xueshuxiangzi.com