通过跨语言的上下文内预训练增强大模型的语言适应性

Linjuan Wu[1*], **Haoran Wei**[2†], **Huan Lin**[2], **Tianhao Li**[2], **Baosong Yang**[2], **Weiming Lu**[1,‡]
[1]Zhejiang University
[2]Tongyi Lab, Alibaba Group
[1] { wulinjuan525,luwm } @zju.edu.cn
[2] { funan.whr, lilai.lh, chongsheng.lth, yangbaosong.ybs } @alibaba-inc.com

Abstract

大型语言模型（LLMs）表现出显著的多语言能力，尽管预训练以英语为主，这归因于预训练期间的跨语言机制。现有的增强跨语言迁移的方法仍然受到平行资源的限制，局限于有限的语言和领域覆盖。我们提出了跨语言上下文预训练（CrossIC-PT），这是一种简单且可扩展的方法，通过简单的下一词预测利用语义相关的双语文本来增强跨语言迁移。我们通过将语义相关的双语维基百科文档交织到一个单一的上下文窗口中来构建 CrossIC-PT 样本。为了解窗口大小限制，我们实施了系统的分段策略，将长的双语文档对分割为块，同时调整滑动窗口机制以保持上下文连贯性。我们进一步通过语义检索框架扩展数据可用性，从网络抓取语料库中构建 CrossIC-PT 样本。实验结果表明，CrossIC-PT 在六种目标语言上的三个模型（Llama-3.1-8B、Qwen2.5-7B 和 Qwen2.5-1.5B）提升了多语言性能，分别带来了 3.79 %、3.99 % 和 1.95 % 的性能提升，并在数据增强后获得了额外改进。

1 介绍

最近最先进的大型语言模型（LLMs）（Achiam et al., 2023; Anthropic; Reid et al., 2024）展示了显著的多语言能力。这些模型通常是在海量网络抓取的数据集上进行预训练的，其中英语文本在数量上占据绝对主导地位（Brown et al., 2020; Dubey et al., 2024）。然而，目前的 LLMs 在非英语语言上的表现异常强大，而这不能完全用预训练期间英语文本相对数据比例来解释。研究人员将这一现象归因于 LLM 训练中的跨语言迁移，即从高资源语言（尤其是英语）习得的语言模式和知识在提升其他语言上的表现时有效地进行迁移（Artetxe et al., 2020; Scao et al., 2022; Wang et al., 2024）。

一系列研究已经探索了解释和增强语言模型预训练期间跨语言迁移的方法。Blevins and Zettlemoyer (2022) 揭示即使在以英语为主的预训练数据中，也能识别出数百万个非英语的标记，它们对于多语言能力至关重要。一些研究试图从共享词汇和表示相似性的角度分析跨语言迁移能力（Patil et al., 2022; Lin et al., 2023），尽管它们的结论主要适用于特定的语言群体。主要的研究范式已经集中于通过利用监督信号（如平行语料库（Zhang et al., 2024b; Ming et al., 2024; Ji et al., 2024; Gosal et al., 2024; Gilabert et al., 2024）、代码转换数据集（Singh et al., 2024; Yoo et al., 2024）或细粒度信号如跨语言实体链接（Yamada and Ri, 2024））显式增强跨语言迁移。然而，这些方法仍受到现有双语资源（例如词典和平行句对）数量有限、领域覆盖范围有限以及形态多样性有限的限制。

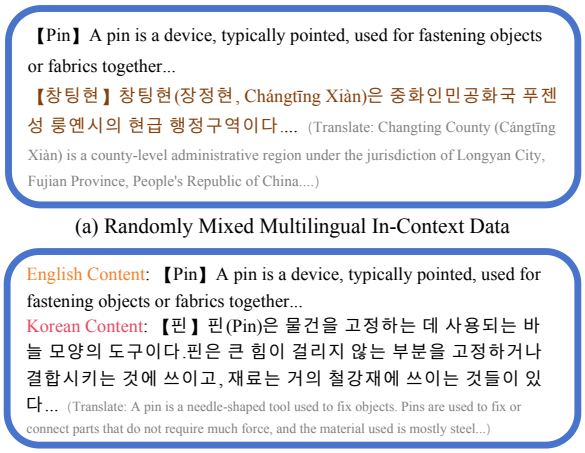

Figure 1: 现有的工作在输入窗口中随机混合多语言文本 (a)。我们的方法将语义相关的文本 (b) 分组，以增强跨语言的迁移。

我们的方法基于 LLM 预训练的基本原则：通过固定长度文本窗口内的下一个词预测（NWP）损失优化进行上下文建模。由于 LLM 可以通过这种机制有效地学习单语语义，我们假设在语义相关的跨语言内容上扩展 NWP 优化——使用源语言上下文来预测目标语言序列——可以增强跨语言转移能力。如图 1 (b) 所示，我们的方法通过交错语义相关的双语文

[*] Work done during internship at Tongyi Lab.
[†] Contributed equally.
[‡] Corresponding authors.

本对构建跨语言上下文样本。随后，我们通过在这些复合样本上进行标准 NWP 损失计算来优化 LLM。所提出的跨语言上下文预训练（CrossIC-PT）消除了对平行语料库的依赖，可以应用于不同类型的文本，为跨语言迁移学习提供了一种简单且可扩展的模式。

为了验证我们的方法，我们通过在现有的大型语言模型（LLMs）（Dubey et al., 2024; Yang et al., 2024）上进行持续预训练（CPT）来实现提出的 CrossIC-PT 方法。与从头开始训练相比，该策略收敛得更快，为多语言实验提供了一种具成本效益的解决方案（Zheng et al., 2024）。利用现成的多语言维基百科数据，我们通过将两个关于同一实体的双语维基百科文章连接成跨语言的上下文语料库，如图 2 所示。为了减轻上下文窗口长度限制，我们将文章对分割成双语子对，使用专门的 [SPLIT] 标记作为分隔符（图 2 (b)）。我们进一步优化了滑动窗口机制，确保下一个窗口从当前窗口的最后一个 [SPLIT] 之后的标记开始，从而保持上下文的一致性，并增强跨语言对齐学习。为了进一步评估我们方法的普遍性，我们开发了一个跨语言语义检索框架，该框架建立在扩展超出维基百科数据的基础上，通过结合网络爬取的文本。如图 3 所示，该框架使用目标语言维基百科文章的标题和部分内容关键词作为查询，从 English Fineweb_edu（Lozhkov et al., 2024）数据集中检索语义相关的段落。

我们在三种大型语言模型（Llama-3.1-8B、Qwen2.5-7B、Qwen2.5-1.5B）的基础上，针对六种语言进行了实验，并在七个任务上对其进行了测试。基于维基百科构建的 CrossIC-PT 模型，相较于基础模型，平均性能分别提高了 3.79 %、3.99 % 和 1.95 %。数据的扩充进一步为 Llama-3.1-8B 提升了 0.73 % 的性能。

我们的贡献可以总结如下：

- 我们提出了 CrossIC-PT，这是一种新颖的方法，通过利用语义相关的上下文数据来增强 LLMs 的跨语言迁移。
- 为了解决输入窗口长度的限制，我们设计了一种窗口分割策略，使用 [SPLIT] 标记和优化的滑动窗口机制来保持跨语言的上下文连贯性。
- 我们还设计了一个跨语言语义检索框架来增强训练数据，这进一步提高了模型性能，证明了我们方法的稳健性和可扩展性。

2 相关工作

许多现有的研究工作集中在收集多语言数据，以增强 LLM 的跨语言能力（Yang et al., 2024; Dubey et al., 2024; Ming et al., 2024; Ji et al., 2024）。来自不同语言的样本以随机方式打包成固定窗口大小（例如，4096），在自注意力中没有交叉污染。即便如此，这些模型已经展示出多语言能力。在此基础上，我们假设连接语义相关的英语和目标语言数据（图 1（b））可能通过利用隐式的监督信号来增强跨语言迁移。

跨语言监督信号已被证明能有效增强大语言模型（LLM）的跨语言迁移能力（Singh et al., 2024; Yamada and Ri, 2024）。大多数方法依赖于双语语料库作为明确的监督信号（Zhang et al., 2024b; Ming et al., 2024; Ji et al., 2024; Gosal et al., 2024; Gilabert et al., 2024）。一些工作，例如（Zhang et al., 2024b）通过反向翻译从 LLM 中提取翻译对来创建监督信号。其他方法，如（Singh et al., 2024; Yamada and Ri, 2024），应用代码转换技术来替换或增加英文翻译的词汇。（Yoo et al., 2024）也探讨了使用课程学习在不同层次上的代码转换。然而，平行语料库在类型、领域（大多数双语语料库是短句级别的平行文本，通常从新闻网站提取）和数量上都有限制。然而，通过反向翻译构建的合成平行文档在文本质量上有限。相比之下，我们的方法从互联网上的真实数据构建语义相关的文档对，这种方法更具可扩展性和较少的问题。

3 方法

多语言 LLM 的预训练通常会随机将来自不同语言的文档打包到固定大小的上下文窗口中。我们假设将语义相关的英文和目标语言文档连接起来，可以增强跨语言迁移。这种方法允许模型在连接的序列中使用单语和跨语言上下文来预测下一个词元。我们称这种连接的样本为跨语言上下文数据，其中英语作为学习目标语言的指导上下文。基于此，我们提出了一种利用跨语言上下文数据的预训练方法 CrossIC-PT。

由于 LLMs 使用固定的标记窗口大小（例如 4096 个标记）进行预训练，跨语言上下文数据通常比普通的单语文档长两倍，可能会超出大小限制。按长度简化打包可能会破坏跨语言关系。为解决此问题，我们精心设计了一种双语感知的窗口拆分策略，以构建跨语言上下文数据。此外，为避免传统滑动窗口机制拆分连接的上下文，我们进一步优化了滑动窗口机制以确保上下文连贯性。

我们利用维基百科的数据来实现我们的方法，如图 2 所示，包含三个关键步骤：（1）数据准备，我们从维基百科中提取并对齐双语文

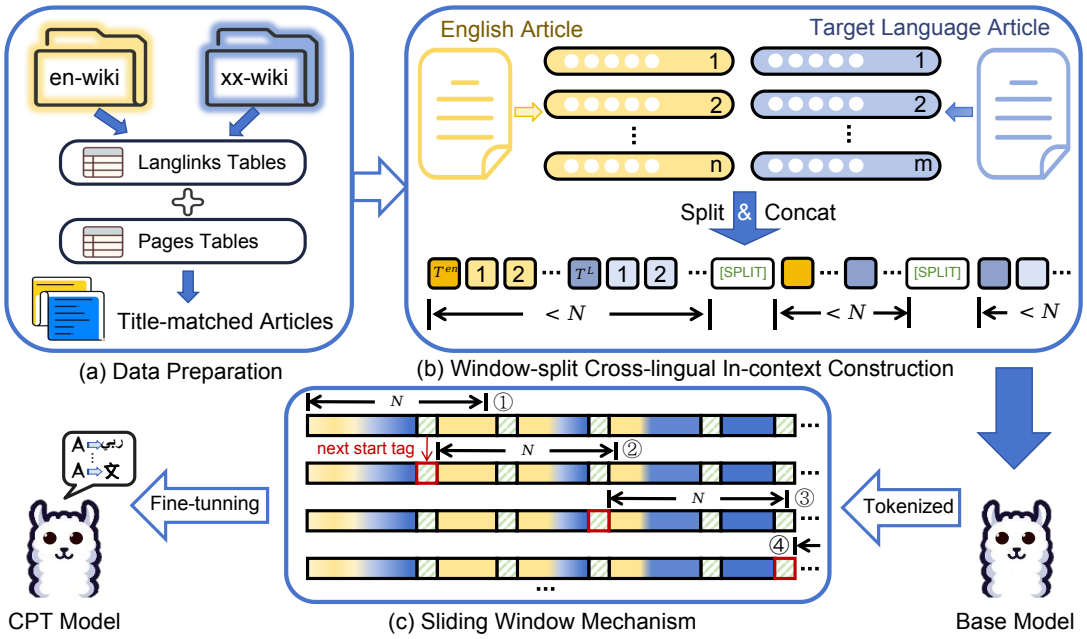

Figure 2: 我们的 CrossIC-PT 方法的实现过程，基于维基百科数据构建跨语言的上下文，并对现有的多语言模型进行持续预训练（CPT）。其中，N 表示模型的输入窗口长度。T 表示文章的标题，L 表示目标语言。

章对（第 **??** 节）；（2）窗口分割跨语言上下文构建，我们分割多语言上下文以匹配输入窗口的长度（第 3.1 节）；以及（3）通过优化的滑动窗口机制进行训练，以增强跨语言表示学习（第 **??** 节）。为了测试我们方法的泛化能力，我们提出一个跨语言语义检索框架来扩充训练数据（第 **??** 节）。

为了获得英语和目标语言（表示为 L ）的对齐文章对，我们利用来自 Wikimedia 的三个关键表，通过三个步骤：

1. 语言 L 的 Langlinks 表：包含语言 L 与其他语言在标题匹配的情况下的文章 ID 映射，以及相应的标题名称 T 。该表有助于识别与语言 L 匹配的英文文章 ID 和标题名称，映射为 $(ID^L, (ID^{en}, T^{en}))$ 。

2. 英文页面表：英文的 ‘pages‘ 表提供了文章 ID 及其对应的标题。我们利用它从步骤 (1) 的初始映射中移除标题为空或无效的英文文章，从而得到最终的 ID 对 (ID^L, ID^{en}) 。

3. 英语和目标语言的文章表格 L ：这两个语言的 “文章” 表格包含文章 ID 和网页上的完整信息，其中包括文章内容。使用双语文章 ID 对 (ID^L, ID^{en}) ，我们提取具有匹配标题的相应文章对。为了确保完整性，我们还执行反向映射 (ID^{en}, ID^L) ，并将结果与前向映射结合以获得一个全面的双语文章对集合。这个过程确保我们捕获了所有可能在英语和目标语言之间标题匹配的文章。

3.1 窗口分割跨语言上下文构建

为了适应上下文大小 N ，我们制定了一种处理长文章对的策略，通过将它们分段成段落并依次对齐。具体来说，对于每一个双语文章对 (A_{en}, A_L) ，我们提取标题 T ，并通过信号"\n\n "将文章分成段落：

$$A_{en} = [p_1^{en}, p_2^{en}, ..., p_n^{en}], \quad A_L = [p_1^L, p_2^L, ..., p_m^L].$$

我们迭代地选择段落对 (p_i^{en}, p_i^L) ，直到添加第 k 对段落会超过长度 N ，然后如下连接段落：

$$(T^{en}, p_1^{en}; p_2^{en}; ...; p_{k-1}^{en}; T^L, p_1^L; p_2^L; ...; p_{k-1}^L),$$

，其中所有英文段落在目标语言 L 段落之前，并以" \n\n " 作为分隔符。每个连接的序列以一个特殊的 [SPLIT] 标记结束以标记上下文窗口的结束。如果一种语言的段落在另一种语言之前耗尽，我们继续连接剩余语言的段落，直到达到长度限制 N 或所有段落都被使用。此过程将每个双语文章对转换为一个或多个窗口分割的多语言上下文，每个都在长度限制 N 内。

在标准的预训练中，滑动窗口机制将所有训练数据连接起来，并以固定的窗口大小进行滑动。然而，这可能随机打破我们跨语言的上下文，破坏连贯性。为了解决这个问题，我们通过引入的标签 "[SPLIT]" 优化了滑动窗口。具体来说，如图 2 所示，所有窗口在最后一个 "[SPLIT]" 标记后设置起始边界。留在结束

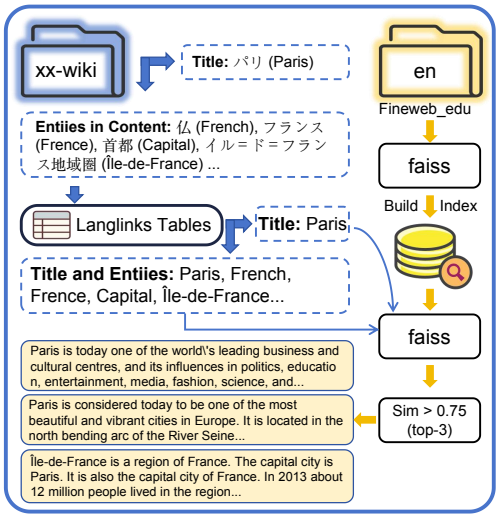

Figure 3: 基于 FAISS 相似搜索工具的跨语言语义检索框架。

边界和最新 "[SPLIT]" 标记之间的标记将被丢弃。通过这种方式，我们尽力在窗口内保持跨语言的连贯性。

3.1.1 训练策略

如前所述，持续预训练（CPT）在跨语言迁移中是具有成本效益的。因此，我们在所有实验中都采用了它。最近的研究，如 (Whitehouse et al., 2024)，表明低秩适配（LoRA）在与完整微调的竞争中表现非常优秀，特别是在低数据和跨语言迁移场景中。在我们的实验中，我们也在持续预训练期间采用了 LoRA，结果显示 LoRA 始终提供更好和更稳定的性能。

为了验证我们的方法，我们使用维基百科语料库，该语料库包含近 200 种语言的数据，并通过匹配的标题进行链接。虽然各语言间的内容并不严格平行，但它涵盖了相同的主题，使其适合我们的需求。为了增强我们方法的泛化性，我们引入了一个基于 FAISS 相似性搜索工具的跨语言语义检索框架 (Johnson et al., 2019)，如图 3 所示。这个框架通过整合来自 Fineweb_edu (Lozhkov et al., 2024) 数据集的相关英文文章来增强训练数据，这些文章是根据标题和内容关键词（每篇文章最多 10 个）从维基百科数据中提取并检索得到的。

首先，从目标语言的维基百科页面中提取关键词，并通过 `langlinks` 表映射到英语。Fineweb_edu 通过 FAISS 索引以进行相似度计算。我们采用使用 FAISS 的两步检索过程：(1) 基于标题关键词的检索，以及 (2) 基于标题和内容关键词的检索。最终的相似度分数是这两步的平均值，平衡了标题（可能模糊）和内容关键词的重要性。根据经验观察，我们设定了 0.75 的相似度阈值，并为每个目标语言文章检索出最多三个相关样本，以构建窗口拆分的跨语言上下文数据。这些样本与原始的维基百科数据结合，形成增强数据集。

我们的训练数据主要来源于维基百科（记为 W），其英语和每种目标语言的标记数量列在表 **??** 中。我们选择了六种目标语言 L：阿拉伯语（ar）、西班牙语（es）、日语（ja）、韩语（ko）、葡萄牙语（pt）和泰语（th）。为了进一步扩充数据集，我们从 Fineweb_edu 的一个子集（记为 F）中提取了相关的英文数据，该子集的文件大小为 17.44GB。扩展数据的标记数量也在表 **??** 中提供。

我们在三个基础模型上进行了实验：Llama-3.1-8B、Qwen2.5-7B 和 Qwen2.5-1.5B。对于 LoRA，我们将秩设置为 64，alpha 设置为 128，dropout 设置为 0.05。输入窗口长度设为 4096，批大小为 128。所有模型都训练一个 epoch，使用热身比例 0.05、余弦学习率调度器和 AdamW 优化器。我们随机选择 0.1% 的数据作为验证集，种子数为 32。对于 Llama-3.1-8B 和 Qwen2.5-7B，一个 epoch 训练后的模型被用作最终模型。对于 Qwen2.5-1.5B，我们每 100 步验证模型，并保存验证损失最低的检查点作为最终模型。训练是在 8 个 A100 GPU 上进行的。

3.2 基准

我们在来自最新多语言和多任务基准 P-MMEVAL (Zhang et al., 2024a) 的几个任务上评估了我们的模型，其中包括：生成（FLORES-200 (Costa-jussà et al., 2022)）、理解（XNLI (Conneau et al., 2018)、MHELLASWAG [*]）、知识（MMMLU [†]）、逻辑推理（MLOGIQA）和数学推理（MGSM (Shi et al., 2023)）。为了进一步评估模型的段落理解能力，我们加入了一个阅读理解任务（MRC）。MRC 测试数据包括阿拉伯语（ar）和韩语（ko）的 TydiQA-GoldP (Clark et al., 2020)，西班牙语（es）、葡萄牙语（pt）和泰语（th）的 XQuAD (Artetxe et al., 2020)，以及日语（ja）的 JaQuAD (So et al., 2022) 中的 1,200 个样本。评估设置的详细信息可以在附录中找到 A。

除了基础模型（Llama-3.1-8B, Qwen2.5-7B 和 Qwen2.5-1.5B）外，我们还包含以下基线：

- EMMA-500 (Ji et al., 2024)：一个在 Llama-2-7B (Touvron et al., 2023) 上的模型 CPT，包含 1360 亿个 token，覆盖超过 500 种语言。
- LEIA (Yamada and Ri, 2024)：一种随机将实体的英文翻译添加到目标语言维基百科数据中的方法，以用于预训练，利用跨语言实体

[*] https://huggingface.co/datasets/alexandrainst/m_hellaswag
[†] https://huggingface.co/datasets/openai/MMMLU

Model		Languages					
		ar	es	ja	ko	pt	th
Llama-2-7B	base	24.77	37.10	37.76	35.05	40.90	23.27
	EMMA-500	30.14	31.18	32.77	32.49	28.06	33.31
Llama-3.1-8B	base	37.96	42.11	43.02	43.82	44.36	38.79
	LEIA	37.04±0.49	44.03±0.24	44.86±0.89	44.11±0.48	44.48±0.58	42.90±0.82
	Mix-PT	38.09	43.46	44.81	44.75	46.45	42.38
	CrossIC-PT	40.57	45.49	47.27	46.87	49.09	43.51
Qwen2.5-7B	base	50.91	54.71	56.95	55.52	56.49	53.81
	Mix-PT	54.48	58.71	57.69	57.39	60.30	56.19
	CrossIC-PT	55.97	59.44	59.00	59.03	61.59	57.33
Qwen2.5-1.5B	base	37.83	43.90	42.26	39.75	44.35	41.40
	Mix-PT	38.14	44.37	41.85	39.48	45.63	40.92
	CrossIC-PT	40.21	45.09	43.96	41.47	48.25	42.23

Table 1: 我们基于三个基础大语言模型（Llama-3.1-8B、Qwen2.5-7B 和 Qwen2.5-1.5B）的 CrossIC-PT 模型的平均结果，与对应的基准在六种目标语言中进行了对比。CrossIC-PT 使用的跨语言上下文数据集来源于维基百科。

监督。我们使用提供的代码复现此方法，以构建数据并在 Llama-3.1-8B 上执行 CPT，确保目标语言的 token 数与我们的相匹配。我们用三个随机种子 (32, 111, 222) 进行了实验，并报告了结果的平均值和方差。

- Mix-PT：一种方法，使用我们从图 2 (a) 中得到的标题匹配文章对进行预训练。

3.3 结果

3.3.1 基础结果

基于六种语言的维基百科数据，基线方法和我们方法的平均结果如表 1 所示。每项任务的详细结果可在附录 **??** 中找到。CrossIC-PT 始终提升了基础大型语言模型的性能，并优于其他基线方法，证明了使用语义相关的跨语言上下文语料库进行预训练的有效性。

与基础的 LLMs 相比，我们的 CrossIC-PT 方法在 Llama-3.1-8B, Qwen2.5-7B 和 Qwen2.5-1.5B 上分别提高了 3.79 %、3.99 % 和 1.95 % 的性能，涵盖了六种语言。值得注意的是，在葡萄牙语（pt）中，CrossIC-PT 在 Llama-3.1-8B 上提高了 4.73 % 的性能，超过了最强的基线方法 2.64 %。对于 Qwen2.5 模型来说，随着模型规模的增加，性能提升更为显著，这可能是因为 CPT 的性能受限于模型最初的能力。

我们的方法在所有语言中都能一致提高性能。在泰语上的提升在 Qwen2.5-1.5B 上不太明显，这可能是由于数据集较小的原因。LEIA 方法在某些语言（西班牙语、日语和泰语）上显示了显著的增长，但其性能不稳定且依赖于数据。例如，日语和泰语的标准差超过 0.8。这表明，与 LEIA 使用的实体对齐信号相比，我们跨语言上下文数据中的隐式监督信号在各语言中更稳健和适应性更强。

Data	Model	Languages					
		ar	es	ja	ko	pt	th
	Llama-3.1-8B	37.96	42.11	43.02	44.14	44.36	38.79
W	Mix-PT	38.09	43.46	44.81	44.75	46.45	42.38
	CrossIC-PT	40.57	45.49	47.27	46.87	49.09	43.51
W+F	Mix-PT	40.19	44.58	44.75	44.48	46.62	42.05
	CrossIC-PT	41.18	46.93	48.10	47.32	49.97	43.72

Table 2: 我们使用基于维基百科的数据以及在维基百科和 Fineweb_edu 数据上构建的扩充数据对 CrossIC-PT 模型和 Mix-PT 基线模型的结果。

Mix-PT 模型是一个强有力的基线，经过训练的非合并维基百科标题匹配文章对，在所有六种语言上相比于三个基础 LLM 提高了性能。然而，我们的方法在 Llama-3.1-8B 上将平均性能提高了 2.15 %。我们的方法通过合并跨语言数据并设计优化的滑动窗口机制进一步增强了 Mix-PT。

为了探索我们方法的泛化性，我们提出了一个跨语言语义检索框架（如图 3 所示）来增强训练数据，其结果在表 2 中报告。检索后，数据量增加了 0.06B–0.23B。尽管这是一个相对较小的增长，但它将我们方法的平均性能提高了 0.73 %。这表明即使英语数据与目标语言没有完美对齐，语义相关性仍然有助于跨语言迁移。语义相似性检索过程的简单性可以轻松扩展到各种数据来源。

此外，我们保存了几个中间检查点来评估数据量对性能的影响。如图 4 所示，在较早的检查点中，我们的方法在所有六种语言中都优于基线 LLM，并在四种语言中超过了强基线 Mix-PT。这表明 CrossIC-PT 可以快速从跨语言上下文数据中获取有用的跨语言迁移能力。

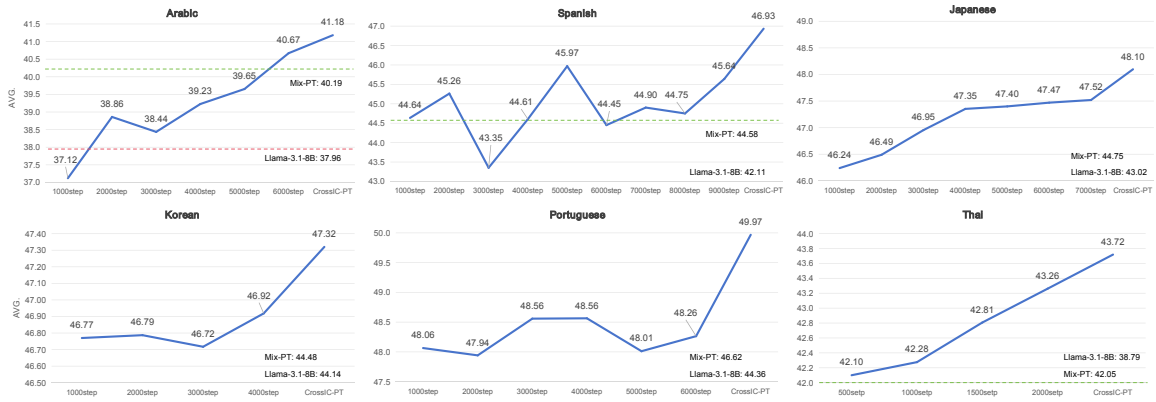

Figure 4: 基于 Llama-3.1-8B 的 CrossIC-PT 在中间检查点的性能进展。我们的方法在早期阶段就超越了基线 LLM，表明其快速获得跨语言迁移能力，并且随着数据量的增加保持缓慢上升趋势。

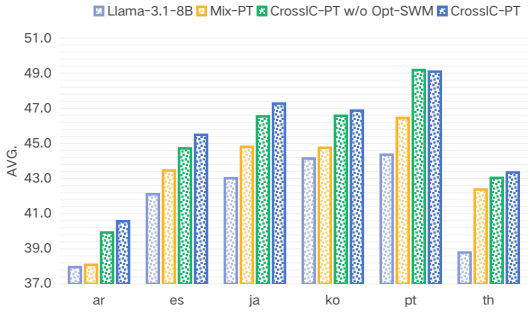

Figure 5: CrossIC-PT 在没有优化滑动窗口机制（Opt-SWM）下的消融结果。

尽管随着数据量的增加，性能提升的速度变得较慢，但仍然观察到了一致的上升趋势。

我们进行了一项消融研究，以评估我们优化的滑动窗口机制（Opt-SWM）的影响，该机制引入了 [SPLIT] 符号，并确保每个窗口在最后一个 [SPLIT] 符号之后开始。具体来说，我们比较了使用和不使用 Opt-SWM 的 CrossIC-PT 的性能，记作 CrossIC-PT w/o Opt-SWM。图 5 中显示的结果表明，即使不使用 Opt-SWM，仅使用窗口拆分的跨语言上下文数据，CrossIC-PT 在所有语言中的性能持续提升。添加优化的滑动窗口机制进一步提高了性能，突显了其在维护跨语言上下文连贯性和改进语言迁移中的作用。这证明了我们设计中所有步骤的有效性。

4 分析

我们认为，在跨语言上下文数据中拼接语义相关的英语和目标语言文本可以帮助模型更好地理解由英语上下文引导的目标语言。因此，我们将顺序设置为英语在前，目标语言在后。为了验证这种顺序是否更有益，我们分析了拼接顺序，并测试了模型在英语中的性能，以确保不会出现灾难性遗忘。

4.1 连接方向分析

为了评估串联方向对性能的影响，我们比较了原始方向（英语在前，目标语言在后）和反向方向（目标语言在前，英语在后），以及两者 1:1 随机混合的情形。此前，我们只报告了翻译任务中 en-xx 方向的结果。在本次实验中，我们还提供了 FLORES-200 上 xx-en 方向的结果。

六种语言在各项任务上的平均结果列于表格 3 中。数据连接顺序对翻译任务的影响最为显著，这符合直觉。当连接方向与翻译方向一致时，翻译性能最佳。在结合两种方向时，CrossIC-PT 在翻译任务中持续优于 Mix-PT 方法，表明即使是非平行的双语数据也能改进翻译。总体来看，先英语后目标语言的连接方式给出了最佳结果，这与我们用英语作为上下文来引导目标语言学习的初衷是一致的。

4.2 英语任务的表现

为了防止灾难性遗忘，确保英语性能的维持是很重要的。为了验证这一点，我们在英语任务上测试了六个目标语言模型的性能，使用的是之前相同的任务。结果如图 **??** 所示。

图 **??** 的上半部分显示了每个目标语言模型在英语任务上的平均表现，x 轴按照 Llama-3.1-8B 在目标语言与英语之间的表现差距进行排序。趋势表明，更大的表现差距在训练后对英语表现有更大的影响。例如，泰语（th）和阿拉伯语（ar）的英语表现较低。然而，这主要是因为在一个任务上的显著下降。为了进一步研究，图 **??** 的下半部分展示了目标语言模型与基础模型 Llama-3.1-8B 在七个任务上的表现差异的统计显著性（“p”）。结果显示，除了使用数据增强训练的泰语模型（在英语表现上表现出显著下降）外，其他目标语言模型没有显著差异。这表明 CrossIC-PT 在提高目标语言性能的同时能有效保留英语能力。我们认为这可能是因为在跨语言上下文语料库中至少包含

Model	XLOGIQA	XHELLASWAG	MMMLU	XNLI	MRC	FLORESE-200		MGSM	AVG.
						en-xx	xx-en		
Llama-3.1-8B	33.25	35.33	40.10	56.17	57.49	38.56	29.41	38.00	41.04
Mix-PT	34.75	36.68	43.05	59.17	60.36	39.63	32.72	36.96	42.92
CrossIC-PT	36.00	39.71	43.15	62.17	63.02	41.39	30.44	39.68	44.44
CrossIC-PT $_{mix}$	34.75	32.33	43.55	58.33	62.20	40.75	33.53	36.96	42.80
CrossIC-PT $_{reverse}$	35.50	33.69	43.00	57.67	62.41	39.51	34.12	36.40	42.79

Table 3: CrossIC-PT 的平均任务结果，包含跨语言上下文数据的双向混合（CrossIC-PT $_{mix}$ ）和反向（CrossIC-PT $_{reverse}$ ）。

50 % 的英语词汇，有助于缓解严重的遗忘。这一结果进一步验证了 CrossIC-PT 在跨语言转移上的稳健性和实用性。

我们的研究通过关注语义相关的多语言上下文，探索了一种增强大型语言模型跨语言迁移能力的特殊角度。我们假设，将语义相关的英语和目标语言语料库连接为跨语言上下文数据是容易获取的，并提供一种隐含的跨语言监督信号。基于这一假设，我们提出了 CrossIC-PT，一种基于跨语言上下文数据的预训练方法。我们使用维基百科数据实施此方法，并对现有的大型语言模型进行持续预训练。为了应对模型训练期间输入窗口长度的限制，我们设计了一种窗口拆分策略，结合优化的窗口滑动机制。实验结果表明，CrossIC-PT 在三个模型—Llama-3.1-8B、Qwen2.5-7B 和 Qwen2.5-1.5B—和六种目标语言中提高了多语言性能，与基线模型相比，分别实现了 3.79 %、3.99 % 和 1.95 % 的性能提升。使用语义检索框架进行数据增强后，进一步观察到了改进。我们的方法易于扩展至多语言大型语言模型预训练，并提供了一种有效扩展数据量的方法。

据我们所知，这项工作存在以下局限性：

- 由于资源限制，我们的实验限制在上下文窗口长度为 4096 个标记。更长的窗口可以更好地保留文章的完整性，并允许来自多于两种语言的相似多语言数据的连接，可能进一步增强跨语言迁移能力。
- 我们的实验集中在验证跨语言上下文数据的级联效力，因此我们进行的是单语言数据的持续预训练，而不是混合多语言数据。虽然这种选择符合我们的研究目标，但我们的方法也为开发多语言大型语言模型的开发者提供了宝贵的见解。
- 我们的数据扩展方法基于检索，目前演示了如何使用目标语言的维基百科数据从外部来源检索额外的英语数据。然而，这种方法可以轻松扩展以检索更多样化的数据。维基百科的广泛领域覆盖使其成为从其他来源检索目标语言和英语数据的理想枢纽。通过使用适当的相似性阈值控制检索过程，检索到的双语数据可以用于构建高质量的跨语言上下文数据。

5

致谢 本工作得到了阿里巴巴研究实习项目的支持。

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, et al. 2023. Gpt-4 technical report.

Anthropic. The claude 3 model family: Opus, sonnet, haiku.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020 , pages 4623–4637.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of english pretrained models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022 , pages 3563–3574. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. ArXiv , abs/2005.14165.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins,

and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. Trans. Assoc. Comput. Linguistics , 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018 , pages 2475–2485. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. CoRR , abs/2207.04672.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. CoRR , abs/2407.21783.

Javier García Gilabert, Carlos Escolano, Aleix Sant Savall, Francesca de Luca Fornaciari, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Investigating the translation capabilities of large language models trained on parallel data only. CoRR , abs/2406.09140.

Gurpreet Gosal, Yishi Xu, Gokul Ramakrishnan, Rituraj Joshi, Avraham Sheinin, Zhiming Chen, Biswajit Mishra, Natalia Vassilieva, Joel Hestness, Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Onkar Pandit, Satheesh Katipomu, Samta Kamboj, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Soundar Doraiswamy, Mohamed El Karim Chami, and Preslav Nakov. 2024. Bilingual adaptation of monolingual foundation models. CoRR , abs/2407.12869.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, et al. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. CoRR , abs/2409.17892.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data , 7(3):535–547.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, André F. T. Martins, and Hinrich Schütze. 2023. mplm-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In Findings .

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong Xu, Yangyang Liu, Xiaohu Zhao, Hao Wang, Heng Liu, Hao Zhou, Huifeng Yin, Zifu Shang, Haijun Li, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement. CoRR , abs/2412.04003.

Vaidehi Patil, Partha Pratim Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. ArXiv , abs/2203.01976.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. ArXiv , abs/2403.05530.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang,

Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. CoRR , abs/2211.05100.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 . OpenReview.net.

Vaibhav Singh, Amrith Krishna, Karthika NJ, and Ganesh Ramakrishnan. 2024. A three-pronged approach to cross-lingual adaptation with multilingual llms. CoRR , abs/2406.17377.

ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. 2022. Jaquad: Japanese question answering dataset for machine reading comprehension. CoRR , abs/2202.01764.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. CoRR , abs/2307.09288.

Hetong Wang, Pasquale Minervini, and E. Ponti. 2024. Probing the emergence of cross-lingual alignment during llm training. ArXiv , abs/2406.13229.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. Low-rank adaptation for multilingual summarization: An empirical study. In Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024 , pages 1202–1228. Association for Computational Linguistics.

Ikuya Yamada and Ryokan Ri. 2024. LEIA: facilitating cross-lingual knowledge transfer in language models with entity-based data augmentation. In Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024 , pages 7029–7039. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. CoRR , abs/2412.15115.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2024. Code-switching curriculum learning for multilingual transfer in llms. CoRR , abs/2411.02460.

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, and Jingren Zhou. 2024a. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. CoRR , abs/2411.09116.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024b. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024 , pages 11189–11204. Association for Computational Linguistics.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. 2024. Breaking language barriers: Cross-lingual continual pre-training at scale. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024 , pages 7725–7738. Association for Computational Linguistics.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, et al. 2023. Gpt-4 technical report.

Anthropic. The claude 3 model family: Opus, sonnet, haiku.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 4623–4637.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of english pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3563–3574. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguistics*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Javier García Gilabert, Carlos Escolano, Aleix Sant Savall, Francesca de Luca Fornaciari, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Investigating the translation capabilities of large language models trained on parallel data only. *CoRR*, abs/2406.09140.

Gurpreet Gosal, Yishi Xu, Gokul Ramakrishnan, Rituraj Joshi, Avraham Sheinin, Zhiming Chen, Biswajit Mishra, Natalia Vassilieva, Joel Hestness, Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Onkar Pandit, Satheesh Katipomu, Samta Kamboj, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Soundar Doraiswamy, Mohamed El Karim Chami, and Preslav Nakov. 2024. Bilingual adaptation of monolingual foundation models. *CoRR*, abs/2407.12869.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, et al. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. *CoRR*, abs/2409.17892.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, André F. T. Martins, and Hinrich Schütze. 2023. mplm-sim:

Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Findings*.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong Xu, Yangyang Liu, Xiaohu Zhao, Hao Wang, Heng Liu, Hao Zhou, Huifeng Yin, Zifu Shang, Haijun Li, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement. *CoRR*, abs/2412.04003.

Vaidehi Patil, Partha Pratim Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. *ArXiv*, abs/2203.01976.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Vaibhav Singh, Amrith Krishna, Karthika NJ, and Ganesh Ramakrishnan. 2024. A three-pronged approach to cross-lingual adaptation with multilingual llms. *CoRR*, abs/2406.17377.

ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. 2022. Jaquad: Japanese question answering dataset for machine reading comprehension. *CoRR*, abs/2202.01764.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Hetong Wang, Pasquale Minervini, and E. Ponti. 2024. Probing the emergence of cross-lingual alignment during llm training. *ArXiv*, abs/2406.13229.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. Low-rank adaptation for multilingual summarization: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1202–1228. Association for Computational Linguistics.

Ikuya Yamada and Ryokan Ri. 2024. LEIA: facilitating cross-lingual knowledge transfer in language models with entity-based data augmentation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 7029–7039. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2024. Code-switching curriculum

learning for multilingual transfer in llms. *CoRR*, abs/2411.02460.

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, and Jingren Zhou. 2024a. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *CoRR*, abs/2411.09116.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024b. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11189–11204. Association for Computational Linguistics.

Wenzhen Zheng, Wenbo Pan, Xu Xu, Libo Qin, Li Yue, and Ming Zhou. 2024. Breaking language barriers: Cross-lingual continual pre-training at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7725–7738. Association for Computational Linguistics.

A 评估环境

我们使用的每个任务的提示在表 4 中显示。由于我们的方法旨在将英语能力转移到目标语言，提示主要是用英语设计的，示范也从英语数据中选择。对于数学推理任务 (MGSM)，我们进行了一个 8-shot 测试；对于阅读理解任务 (MRC)，我们采用零-shot 设置来评估模型对目标语言的理解；对于其他任务，我们进行了一个 5-shot 测试。对于选择题任务（例如，XNLI、MMLU、XHELLASWAG、XLOGIQA），我们通过预测下一个 logits 来直接获得答案。对于其他任务，我们使用贪婪搜索生成答案，并通过正则表达式匹配提取最终答案。

Task	Prompt
XLOGIQA	`Passage: { context } \nQuestion: { question } \nChoices:\nA. { option_a } \nB. { option_b } \nC. { option_c } \nD. { option_d } \nAnswer:`
XHELLASWAG	`{ premise } \nOptions: \nA. { option_1 } \nB. { option_2 } \nC. { option_3 } \nD. { option_4 } \nQuestion: Which is the correct ending for the sentence from A, B, C, and D? \nAnswer:`
MMMLU	`The following is a multiple-choice question.\n\n { question } \nA. { option_a } \nB. { option_b } \nC. { option_c } \nD. { option_d } \n\nAnswer:`
XNLI	`Take the following as truth: { premise } \nThen the following statement: " { hypothesis } " is\nOptions:\nA. true\nB. inconclusive\nC. false\nAnswer:`
MRC	`Refer to the passage below and answer the following question:\nPassage: { context } \nQuestion: { question } \nAnswer: Based on the passage, the answer to the question is "`
FLORES-200	`Translate from [source] to [target].\n[source]: </X>\n[target]:`
MGSM	`Solve this math problem. Give the reasoning steps before giving the final answer on the last line by itself in the format of "[The answer is ]". Do not add anything other than the integer answer after "The answer is".\n\n { question }`

Table 4: 任务提示。“[]”表示可选内容。对于 FLORESE 任务，“[source]”表示源语言，“[target]”表示翻译的目标语言。对于 MGSM，“[答案是]”是根据测试语言对“答案是”的翻译。

我们的方法和基线在每个任务中的六种语言的平均结果如表 **??** 所示。